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, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 265 Crittenden Blvd., Box 630, Rochester, NY 14642, USA, andrea_baran@URMC.Rochester.edu

Arun H. Patil, Lieber Institute for Brain Development, Johns Hopkins University, 855 North Wolfe St. Suite 300, Baltimore, MD 21205, USA, arun26feb@gmail.com

Ernesto Aparicio-Puerta, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 265 Crittenden Blvd., Box 630, Rochester, NY 14642, USA, Ernesto_AparicioPuerta@URMC.Rochester.edu

Marc K. Halushka, Department of Pathology, Cleveland Clinic, 9500 Euclid Ave., Cleveland, OH 44195, USA, HALUSHM@ccf.org

Matthew N. McCall, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 265 Crittenden Blvd., Box 630, Rochester, NY 14642, USA, mccallm@gmail.com (corresponding author)

Keywords: microRNA, isomiR, differential expression, mixed model, aggregation, miRNA-seq

Abstract:

MicroRNA-seq data is produced by aligning small RNA sequencing reads of different miRNA transcript isoforms, called isomiRs, to known microRNAs. Aggregation to microRNA-level counts discards information and violates core assumptions of differential expression (DE) methods developed for mRNA-seq data. We establish miRglmm, a DE method for microRNA-seq data, that uses a generalized linear mixed model of isomiR-level counts, facilitating detection of miRNA with differential expression or differential isomiR usage. We demonstrate that miRglmm outperforms current DE methods in estimating DE for miRNA, whether or not there is significant isomiR variability, and simultaneously provides estimates of isomiR-level DE.

42 **Background**

43 MicroRNA (miRNA) are a class of small, noncoding RNAs that perform a role in transcriptional regulation.
44 They are typically 18-24 nucleotides long single-stranded RNA molecules that bind to mRNA causing
45 translational suppression or mRNA degradation (1, 2). Through this mechanism, miRNA can regulate
46 entire pathways and drive disease pathogenesis (3, 4). The specificity of each miRNA:mRNA interaction
47 leads to discrete downstream consequences (5). Small RNA sequencing (sRNA-seq) can be used to
48 measure miRNA expression. This process involves enriching samples for small RNA species prior to
49 sequencing, followed by aligning the sequence reads to known miRNAs, tRNA fragments, or other RNA
50 species (6). miRNA count data, which we refer to as miRNA-seq data throughout, is produced when
51 sequence reads are aligned to known miRNAs, exclusively.

52 The specific mature sequence listed in miRNA databases is considered the canonical miRNA sequence
53 (7). Different isoforms of individual miRNAs, called isomiRs, can arise from both variation in the
54 nucleotide sequence or from variation in the transcript length (8). Sequence length variants have more
55 or fewer nucleotides at the 5' and/or the 3' end of the canonical sequence, whereas polymorphic
56 (internal) isomiRs include different nucleotides within the mature sequence (7). A third form of isomiR is
57 the addition of an adenosine or uracil tail by a terminal uridylyl transferase (TUT) or similar enzyme (9).
58 We will use the general term isomiR to describe the various sequence isoforms mapped to a miRNA,
59 without distinction between biologically or technically derived isomiRs, with the assumption that in
60 biological samples most sequences observed at sufficient levels to model are biologically relevant
61 isomiRs.

62 The collection of isomiRs that align to a given miRNA are referred to as the miRgroup (10). As an
63 example of isomiR diversity within miRgroup, Supplemental Tables 1 and 2 display isomiR-level miRNA-
64 seq read counts for two miRgroups: hsa-let-7a-5p and hsa-miR-26a-5p. Typically, read counts of each

65 isomiR in a miRgroup are aggregated by summation and summarized as a single read count for each
66 miRNA (as shown in the last row of Supplemental Tables 1 and 2). There are numerous combinations of
67 isomiR counts that produce the same miRNA counts; therefore, this aggregation leads to the loss of any
68 information contained in individual isomiR expression.

69 IsomiRs are thought to possess unique biological roles (11). Like canonical sequences, isomiRs are
70 conserved throughout evolution, and biogenesis of isomiRs is tightly regulated. The region of
71 heterogeneity between isomiRs may have unique implications for miRNA-mediated translation
72 regulation (8). 5' isomiRs have modifications at the 5' end of the miRNA resulting from differential
73 processing of paralogous pre-miRNA and can result in regulation of distinct target mRNA (12). 3' isomiRs
74 are more common and result from post-transcriptional trimming or tailing sequence modifications (13).
75 These modifications can alter miRNA function by impacting target recognition or extent of target
76 repression. Additionally, these modifications can impact the stability of the molecule as adenylation
77 protects miRNAs from degradation, while uridylation is associated with increased degradation (7).
78 Naturally occurring isomiRs have been shown to play distinct roles in a variety of biological processes
79 including cytokine expression, virus proliferation, apoptosis, and tumor progression (11). Differential
80 expression (DE) at the isomiR-level has identified isomiRs with cancer-specific expression (14). The
81 biological importance of isomiRs highlights the need for a miRNA-seq analysis method that can account
82 for distinct isomiR expression patterns in estimating miRNA-level differential expression, while also
83 producing more granular estimates of isomiR-level differences.

84 Similar to differential expression analyses typically performed using bulk mRNA-seq data, the goal of
85 most miRNA-seq studies is to study if, and which, miRNA differ in their expression between groups of
86 samples. Common DE tools developed for mRNA-seq data include DESeq2 (15), edgeR (16) and limma-
87 voom (17). All of these methods have been frequently used to analyze miRNA-seq data; however, there
88 are some key differences between miRNA-seq data and mRNA-seq data that may make the assumptions

89 of these methods invalid when applied to miRNA-seq data. Due to the compositional nature of RNA
90 sequencing data, the reads can be viewed as a random sample of a fixed size from the pool of all RNA in
91 the library, which can be modeled by a multinomial distribution. In bulk mRNA-seq data, the number of
92 unique mRNA expressed is large and the reads are distributed relatively evenly across the mRNA. In this
93 setting, the true underlying multinomial distribution can be approximated by marginal binomial
94 distributions. This approximation is a key assumption of the tools mentioned above but is violated in
95 miRNA-seq data in two important ways. First, there are generally fewer than 500 miRNA compared to
96 over 10,000 mRNA expressed in a sample, which makes it more likely that random fluctuation in the
97 expression of one miRNA substantially affects the expression of other miRNAs. Second, the distribution
98 of reads in miRNA-seq data is often skewed toward a small number of highly expressed miRNA
99 compared to the more uniform distribution of reads seen in mRNA-seq data (18). This can induce
100 negative correlation between highly expressed miRNA (due to competition of being counted) regardless
101 of their underlying biological correlation. This also results in sparse data with a small number of non-
102 zero features making traditional normalization approaches, such as size factor normalization or counts-
103 per-million (CPM) normalization, unstable or highly skewed.

104 Differences between miRNA and mRNA make the analysis of miRNA counts at the isomiR level feasible.
105 First, even using Illumina short read sequencing, miRNAs and all isomiRs are fully sequenced (18-24 nt),
106 whereas typically only fragments of mRNAs (100s of nt out of ~5,000-50,000 nt species) are sequenced.
107 Therefore, quantifying the expression of miRNAs based on the entire sequence is possible (6). Second,
108 as previously stated, there are far fewer miRNA expressed compared to mRNAs. Maintaining and
109 analyzing sequence level data at the scale needed for mRNA analysis would be computationally
110 burdensome. Recent work has highlighted the importance of analyzing miRNA-seq data at the isomiR
111 level (10). Through the analysis of 28 public miRNA-seq datasets and a newly generated human
112 endothelial cell hypoxia data set, the authors show substantial differences between isomiR expression

113 and their corresponding canonical miRNAs when applying DESeq2 to miRNA counts versus isomiR
114 counts. As noted above, aggregate miRNA level data violates the assumption of independence between
115 features due to the small number of unique miRNAs expressed in a sample and the highly skewed
116 distribution of miRNA expression. While analyzing isomiR level data greatly increases the feature space
117 and reduces the overall skew of the expression distribution, the isomiR level data introduce a new
118 source of dependence due to high correlation between isomiRs from the same miRNA, which also
119 violates a core assumption of the DESeq2 model. Despite this limitation, we agree with the authors'
120 conclusion that "a comprehensive re-evaluation of the miRNA-seq analysis practices" is needed.

121 Negative binomial mixed models (NBMM) can be used to model overdispersed count data when there is
122 a correlation structure among the counts (19). This is the case when working with miRNA-seq data at the
123 isomiR level, as counts that are observed for the same isomiR or for the same sample are correlated.
124 Modelling the raw (non-normalized) counts directly has advantages over the typical counts per million
125 (CPM) normalization. CPM is a dependent normalization strategy, so a change in any one miRNA read
126 count will lead to changes in all other miRNA values even in the absence of a change in absolute
127 expression (20). To account for the technical artifact of difference in the number of total reads between
128 samples, one can incorporate an offset term into the NBMM to adjust for total overall read count within
129 each sample. NBMMs have been proposed for other next-generation sequencing analyses, including
130 RNA-seq, metagenomic sequencing, and single cell RNA-seq analyses (21-24). However, the NBMMs
131 used in these methods all model aggregate feature level data rather than isoform level data and none
132 have been applied to miRNA-seq data.

133 In this manuscript, we propose miRglmm, a method to model isomiR-level counts using a generalized
134 linear mixed model to estimate miRNA-level DE while also obtaining estimates of isomiR DE and
135 variability. We compare miRglmm to several commonly used DE tools on simulated data, an
136 experimental benchmark data set, and real biological data sets. In both simulations and experimental

benchmark data, miRglmm has a lower Mean Squared Error (MSE) than other DE tools and better confidence interval coverage. Additionally, we find significant isomiR-level variability exists within most miRgroups in real biological data sets, further motivating the use of miRglmm to analyze miRNA-seq data.

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 (25), with the goal of minimizing possible technical variation, and therefore capturing 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 1). 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 isomiRs, 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 always 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 (26) in which randomly measuring more of one feature decreases the chance of measuring other features. Consider the extreme example of only two features;

160 with a fixed number of total reads, it is clear that a higher count for feature 1 implies a lower count for
161 feature 2, resulting in perfect negative correlation between the two features regardless of their
162 underlying biological correlation. Frequently, this induces a negative correlation between the two most
163 highly expressed miRNAs; however, this induced correlation is not always observed for the most highly
164 expressed isomiRs. To illustrate this, we selected a study of immune cell types (27), with the goal of
165 analyzing samples without cellular heterogeneity that can occur in tissue-level samples. We compared
166 the correlation between the two highest expressed miRNAs with the correlation between the two
167 highest expressed isomiRs for each cell type. While we observed a negative correlation between the
168 highest expressed miRNAs for all cell types, the same was not true at the isomiR level where we
169 observed a mix of positive and negative correlations depending on the cell type (Supplemental Figure 1).
170 This indicates that utilizing the isomiR-level count data could overcome technical biases seen with
171 aggregated miRNA-level count data.

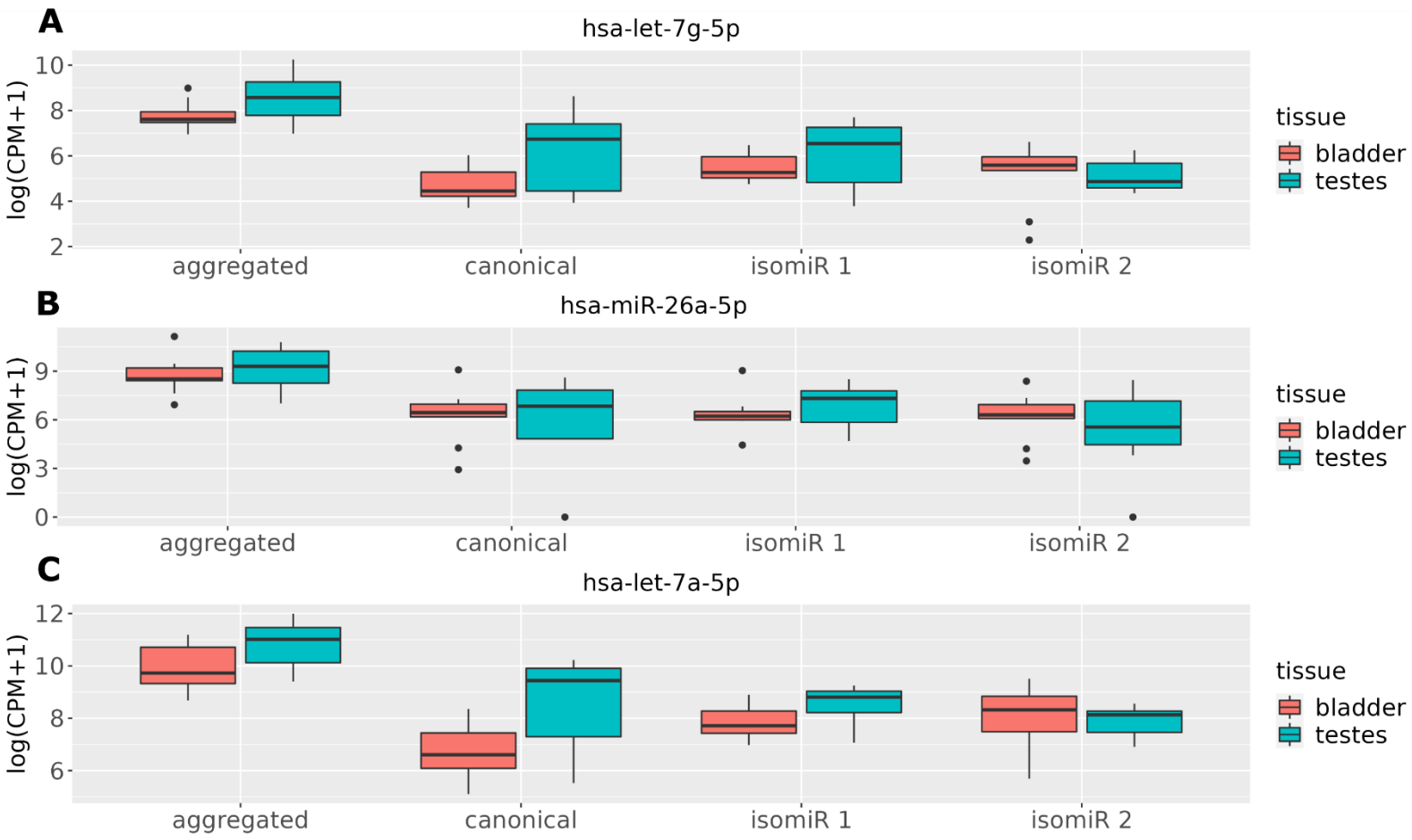


Figure 1: Tissue-specific counts are summarized by boxplots for three miRNAs (panel A: hsa-let-7g-5p, panel B: hsa-miR-26a-5p, panel C: hsa-let-7a-5p). Aggregated miRNA-level counts from summing counts across isomiRs within a miRNA are compared to the canonical/representative sequence and the two highest expressed non-canonical isomiRs (isomiR 1 and isomiR 2). The canonical sequence counts, and isomiR 1 are all

176 higher in testes than bladder, but isomiR 2 has the opposite trend, which is masked when counts are aggregated to miRNA-level. CPM: counts

177 per million.

178

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

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

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

190 To assess the performance of miRglmm in comparison to existing methods (DESeq2 (15), edgeR (16),
191 limma-voom (17) and Negative Binomial Generalized Linear Models (NB GLM)(28)), we used a collection
192 of monocyte samples to simulate differential expression at both the miRNA- and isomiR-level (see
193 Methods). Over 100 simulations, miRglmm provided the lowest mean MSE (mean = 0.015) and median
194 MSE (median = 0.014) among all the methods (Table 1, Supplemental Table 3). miRglmm was also the
195 method with the lowest standard deviation across simulations (standard deviation (SD) = 0.005) and the
196 lowest maximum MSE (maximum= 0.033). miRglmm minimized MSE in 97% of simulations (97/100)
197 when comparing methods within a given simulation. DESeq2 minimized the MSE in the remaining 3
198 simulations. On average, for these 3 simulations, the MSE for DESeq2 was only smaller than the MSE for
199 miRglmm by 0.001. miRglmm provided the highest mean coverage proportion (mean = 0.91) and highest
200 median coverage proportion (median = 0.93) among the 4 methods that produce standard error (SE)
201 estimates which can be used to calculate confidence interval coverage (Table 1, Supplemental Table 4).

202 Table 1: MSE and 95% confidence interval coverage proportion summary statistics across 100

203 simulations

method	mean MSE (10^{-3})	minimum MSE (10^{-3})	maximum MSE (10^{-3})	# of times method minimizes MSE	mean coverage proportion	minimum coverage proportion	maximum coverage proportion
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

204 MSE: Mean Squared Error, NA: not applicable

205 We can also assess performance within each "truth" group of miRNA: induced positive effect (N=20),

206 induced negative effect (N=20), or no effect induced (N=82). When we calculate MSE within each group

207 and summarize across simulations, we see that miRglmm provides markedly lower MSE in the groups

208 with the induced effect compared to all aggregation methods (Figure 2A). In the group of miRNA with no

209 change induced, DESeq2 tends to provide a lower MSE in most simulations. Similarly, the coverage

210 proportion in the two groups of miRNA with induced effects tends to be much closer to the 95% nominal

211 level than for the aggregation methods (Figure 2B). The aggregation methods have very poor confidence

212 interval coverage proportions (0.3-0.4). For the miRNA with no induced effect, DESeq2 tends to have the

213 highest coverage proportion, though miRglmm still provides coverage near the 95% nominal level.

214 An additional advantage of miRglmm is that it provides an estimate of the variability in the group effect

215 between isomiRs, facilitating the detection of miRNA with differential isomiR usage between groups.

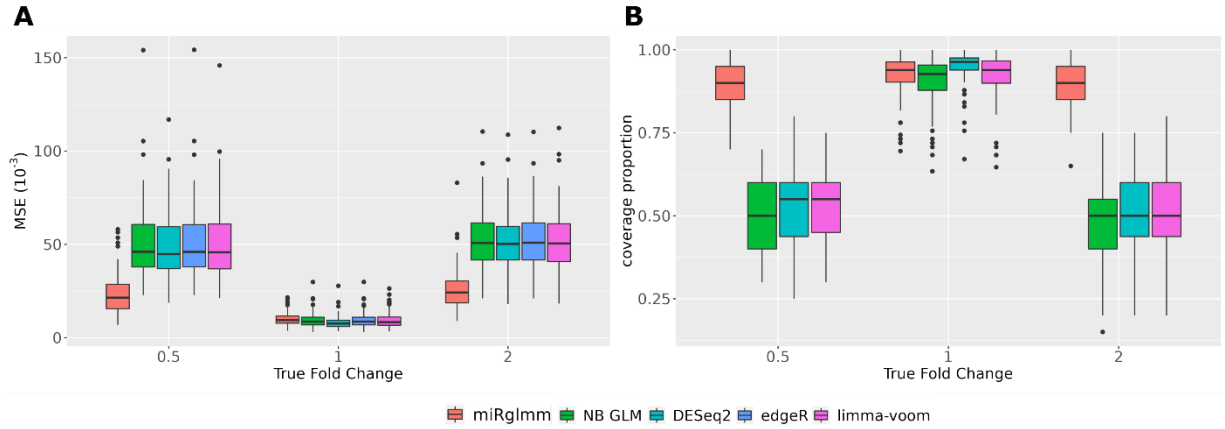
216 The other methods cannot estimate or account for this variability, and thus do not have the ability to

217 detect miRNAs with differential isomiR usage. We observe that the proportion of miRNAs with

218 significant variability between isomiRs is very high in the groups with the induced effect (Supplemental

219 Figure 2), indicating that our simulation procedure correctly implemented random isomiR DE variability

220 in the group effect. Taken together, these results support the conclusion that aggregation of isomiR
221 counts to miRNA-level counts and the resulting loss of information is detrimental to performance in
222 cases where significant isomiR DE variability exists. miRglmm, which can account for this variability,
223 provides consistently high performance whether or not there is significant isomiR DE variability.



224

225 Figure 2: miRglmm outperforms the aggregation methods in terms of Mean Squared Error (MSE) for
226 miRNA in which an effect is induced (panel A). DESeq2 provides the lowest MSE when there is no effect
227 induced. The 95% confidence interval coverage proportion of miRglmm is much higher than the
228 coverage proportion of the aggregation methods when an effect is induced (panel B). Results are based
229 on 100 independent simulated data sets.

230

231 **miRglmm provides accurate estimates of differential expression for isomiRs**

232 Along with providing miRNA-level estimates of differential expression, miRglmm also provides isomiR-
233 level estimates of differential expression. While rarely done, DESeq2 can also provide isomiR-level
234 estimates if run on the isomiR-level count matrix. Summarizing over 100 simulations, miRglmm
235 (MSE=0.022) provides better estimates of isomiR-level differential expression than DESeq2 (MSE=0.041).
236 When we assess performance within each “true effect” group, we see that miRglmm outperforms

237 DESeq2 when there is a negative effect (MSE=0.035 vs 0.044), when there is a positive effect
238 (MSE=0.035 vs 0.044), and when there is no differential expression (MSE=0.015 vs 0.040). This contrasts
239 with what we observed at the miRNA-level, where miRglmm slightly underperformed DESeq2 with
240 respect to MSE when there was no differential expression, but performed much better than DESeq2
241 when there was either positive or negative differential expression.

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

244 While variable isomiR usage is common in real biological data, we wanted to assess how miRglmm
245 would perform in the absence of isomiR variability. To assess the performance of miRglmm in this
246 context, we used a multi-protocol, multi-institution experimental benchmark data originally designed to
247 compare the performance of four different sRNA-seq library preparation methods (29). This experiment
248 used ratiometric pools of synthesized small RNAs with known variable amounts of differential
249 expression (see Methods), providing a known true fold change value for each miRNA that can be used to
250 evaluate the accuracy of DE methods. Because the data came from mixtures of synthesized miRNA
251 pools, there should be no biological variability in isomiR usage.

252 We used MSE and confidence interval coverage to assess the performance of the DE methods (Table 2)
253 as well as expression-based isomiR filtering thresholds (Supplemental Table 5). miRglmm accurately
254 estimated the known DE magnitudes between synthetic miRNA pools while maintaining greater than
255 nominal confidence interval coverage. In terms of MSE, miRglmm at any isomiR expression filtering
256 threshold provided smaller MSE than all aggregation methods. However, miRglmm with no filtering of
257 lowly expressed isomiRs had the highest MSE, and we observed that the estimates were systemically
258 biased towards the null due to the inclusion of very lowly expressed isomiRs (Supplemental Figure 3).
259 The four aggregation methods provide similar estimates but underestimate the true effect at all levels.

260 We chose log(median CPM) > -1 as the default filter for miRglmm because this retained the most

261 isomiRs while achieving similar MSE as more stringent filtering thresholds (Supplemental Figure 4). As

262 filtering gets more restrictive, we lose the ability to model some miRNA if fewer than two isomiRs are

263 retained.

264 Table 2: Comparing performance of miRglmm to aggregation methods

method	MSE (x 10^{-3})	coverage proportion
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

265 MSE: Mean Squared Error, NA: not applicable

266 To determine whether other methods would improve if we filtered lowly expressed isomiRs, we
267 compared MSEs based on estimates from running the other methods on aggregated counts from all
268 isomiRs versus filtering prior to aggregation of the isomiR counts. The gain in performance for miRglmm
269 due to filtering was not seen for the aggregation methods (Supplemental Figure 3). Since aggregating is
270 done via summation, the low counts do not influence the sum to the same degree that they influence
271 the miRglmm estimate.

272 Due to the synthetic nature of the data, we do not expect any biological isomiRs and presume that the
273 sequence isoforms we are modeling arise solely from technical variation. Hence, the random slope
274 parameter was not found to be significant in the synthetic data (Supplemental Figure 3). If we remove
275 this extraneous random slope parameter for isomiR from the model, the performance is similar in terms
276 of the MSE, confidence interval coverage and bias (Supplemental Table 6 and Supplemental Figure 5).
277 This indicates that retaining the random slope parameter, even if not needed, does not diminish the
278 ability of miRglmm to estimate the effect of interest.

279

280 **Differential miRNA expression and differential isomiR usage between immune cell types**

281 We performed differential expression analyses comparing miRNA expression between five immune cell
282 types: monocytes (N=39), Natural Killer Cells (N=38), CD4+ T lymphocytes (N=35), CD8+ T lymphocytes
283 (N=32) and CD19+ B lymphocytes (N=26). The samples were obtained from one experiment in the
284 miRNAome data (27), limiting the possibility of technical variation due to library preparation or
285 sequencer differences. Based on isomiR filtering strategies assessed using data from the Extracellular
286 RNA Communication Consortium (ERCC), we chose to implement miRglmm using a threshold of
287 log(median CPM) < -1 to remove isomiRs with low expression. After isomiR filtering, we retained 7,389

288 isomiRs from 104 miRNA, and the number of isomiRs mapping to any one miRNA ranged from 6 to 633
289 with a median of 35 isomiRs.

290 We calculated log fold-change (logFC) estimates for each miRNA using miRglmm and existing DE
291 methods (see Methods). B lymphocytes were used as the reference group in each logFC estimate. We
292 used intra-class correlation coefficients (ICC) to assess agreement in logFC estimates between methods
293 (Figure 3A). While all methods provided similar logFC estimates (ICC ≥ 0.95), we observed less
294 agreement regarding statistical significance (Bejamini Hochberg False Discovery Rate (FDR) of 0.05)
295 between methods (Cohen's Kappa 0.36-0.53). There does not appear to be substantial differences in ICC
296 between contrasts, i.e. specific cell type comparisons (Supplemental Figures 6-9), but some contrasts
297 have very poor agreement in terms of statistical significance (minimum Kappa=0.11 when comparing
298 edgeR and miRglmm for CD4+ T lymphocyte vs B lymphocyte and CD8+ T lymphocyte vs B lymphocyte
299 contrasts). An upset plot of statistically significantly DE miRNAs across all contrasts provides a visual
300 comparison of all 5 methods simultaneously (Figure 3B). There are 4 miRNA-contrast pairs significant for
301 miRglmm only but no aggregation method (Supplemental Table 7), and 26 miRNA-contrast pairs
302 significant for all aggregation methods but not miRglmm (Supplemental Table 8). We used likelihood
303 ratios tests (LRTs) to assess if there is variability in the cell type effect between isomiRs within a given
304 miRNA. All miRNAs had significant LRT p-values for the random slope parameter for isomiRs (Figure 3C).
305 The other methods we assessed cannot estimate or account for this variability.

306

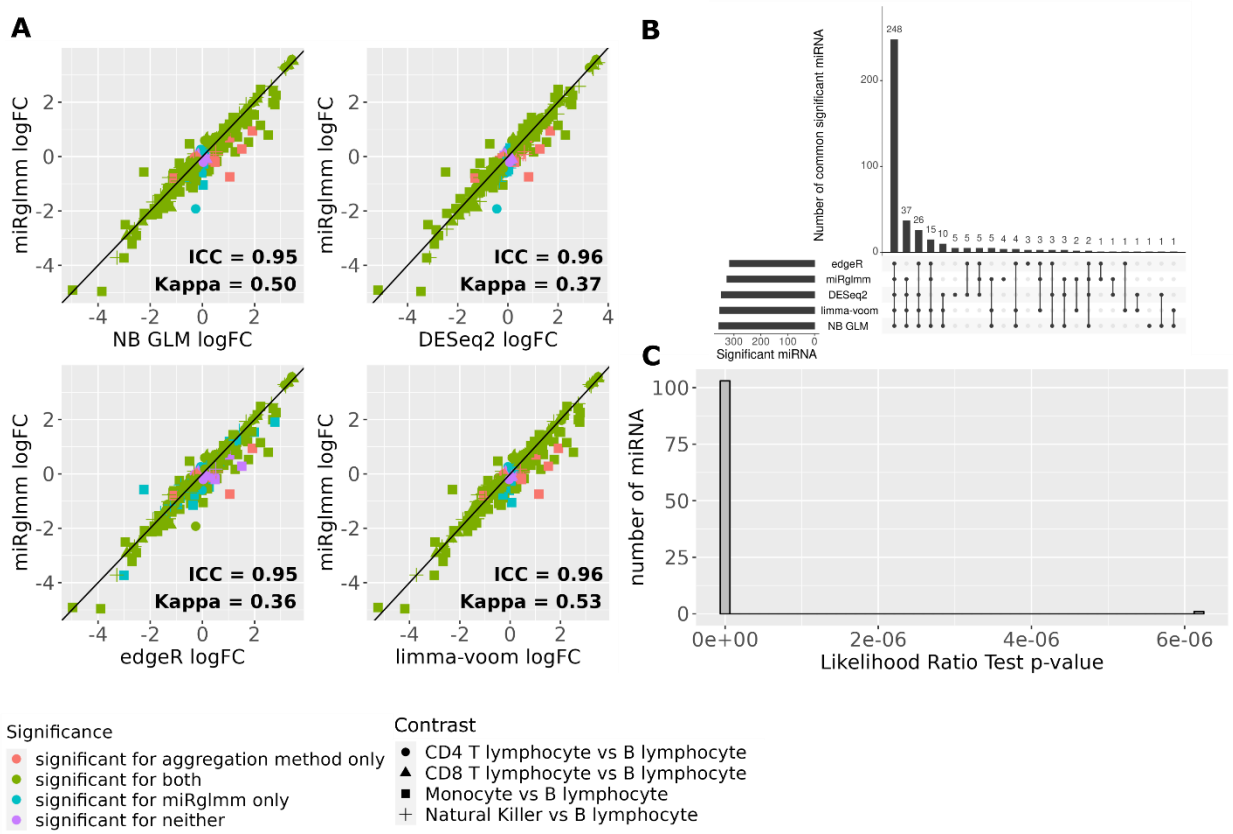
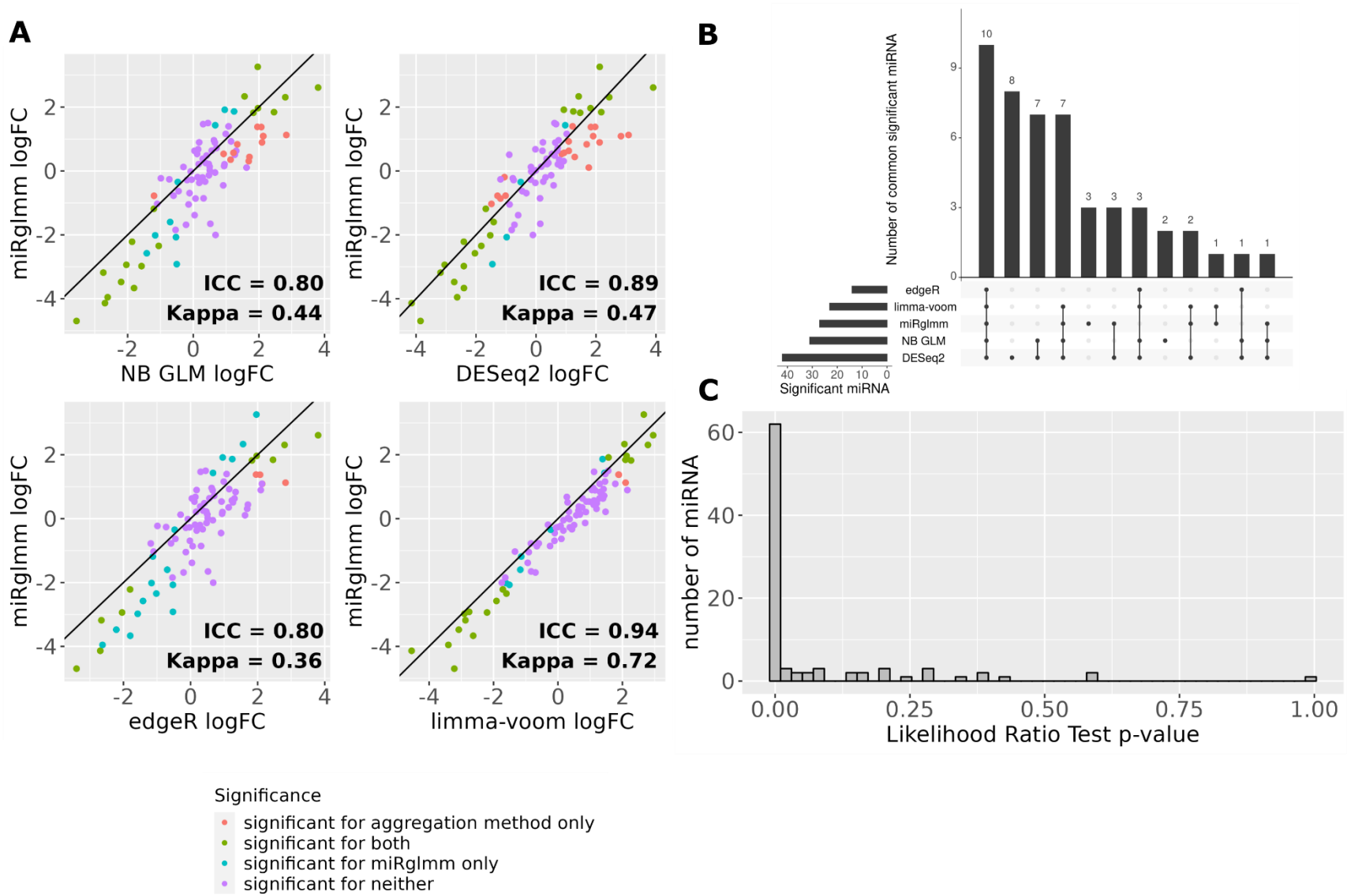


Figure 3: Log fold change (logFC) estimates of miRNA-level expression in cell types compared to B lymphocytes from miRglmm are plotted versus logFC estimates from each aggregation method (panel A). Each point is a unique contrast for a unique miRNA. Significant DE vs B lymphocytes is indicated by the color (Benjamini Hochberg FDR of 0.05). 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).

318 **Differential miRNA expression and differential isomiR usage between tissues**

319 We performed differential expression analyses comparing miRNA expression between bladder (N=9)
320 and testes (N=7). These tissue types represented the largest set of two distinct tissues in one single
321 experiment in the miRNAome data. Since the data came from one experiment, there was no known
322 technical source of variation to adjust for, and tissue-type was the only fixed effect in the model. We
323 performed miRNA-level and isomiR-level filtering (see Methods), and retained 3,036 isomiRs from 90
324 miRNA. The number of isomiRs mapping to any one miRNA ranged from 5 to 436 with a median of 18
325 isomiRs.

326 We estimated log fold-change (logFC) between tissue types for each miRNA using miRglmm and
327 compared the estimates to existing DE methods (see Methods). We used intra-class correlation
328 coefficients to assess agreement in logFC estimates between methods (Figure 4A). All methods provided
329 reasonably similar logFC estimates (ICC ≥ 0.8). As with the cell type analysis, agreement with respect to
330 statistically significant DE, measured via Cohen's Kappa, was much weaker. We similarly compared all
331 aggregation methods in a pairwise fashion (Supplemental Figure 10) and observed high ICC (≥ 0.88) and
332 weak-to-moderate kappas (≥ 0.35-0.68) between methods. An upset plot of miRNA deemed significantly
333 different between bladder and testes across the methods provides a visual comparison of all 5 methods
334 simultaneously (Figure 4B). 67 of 90 miRNA have significant LRT p-values for the random tissue
335 parameter for isomiRs (Figure 4C).


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 = 0.36
limma-voom logFC
ICC = 0.94
Kappa = 0.72
B
Number of common significant miRNA
Significant miRNA
edgeR
limma-voom
miRglmm
NB GLM
DESeq2
C
number of miRNA
Likelihood Ratio Test p-value
Significance
significant for aggregation method only
significant for both
significant for miRglmm only
significant for neither

336

337 Figure 4: Log fold change (logFC) estimates of miRNA-level expression in testes compared to bladder from miRglmm are plotted versus logFC
338 estimates from each aggregation method(panel A). Each point is a unique miRNA. Significant DE is indicated by the color (Benjamini Hochberg
339 FDR of 0.05). Intra-class correlation coefficients (ICC) are used to assess agreement in the values of the estimated logFC between methods.
340 Cohen's kappa is used to assess agreement in the significance of the estimate between methods. An upset plot shows the common number of
341 miRNAs significantly different between bladder and testes across the 5 methods (panel B). There are 3 miRNA found to be significant for all
342 aggregation methods but not miRglmm. There are 3 miRNA found to be significant for miRglmm but none of the aggregation methods. A
343 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
344 variability in the bladder-testes effect across isomiRs (panel C).

345 miRglmm can also provide isomiR-level estimates of differential expression. We visualized isomiR-level
346 expression and variability within several miRNA by plotting the estimated effects for each isomiR
347 mapping to the miRNA (Figure 5). We overlaid estimated fixed effects from miRglmm to compare the
348 miRNA-level differential expression estimates (fixed effect estimates from miRglmm) with underlying
349 individual isomiR effects (random isomiR effect estimates from miRglmm). We also overlaid the NB GLM
350 estimated effect as an example aggregation method (other aggregation method results are similarly
351 shown in Supplemental Figure 11). Figure 5A-5C show the 3 miRNA that only miRglmm identifies as
352 differentially expressed between bladder and testes (Figure 4B). Figure 5D-5F show the 3 miRNA that
353 are found to be significantly differentially expressed by all aggregation methods but not by miRglmm
354 (Figure 4B). miRglmm estimates of expression tend to follow patterns observed for the majority of
355 isomiRs, whereas NB GLM estimates of expression tend to follow patterns observed in only the highest
356 expressing isomiRs. In the case of hsa-miR-143-5p (Figure 5B) there is little variability in isomiRs,
357 increasing the precision of the estimate which results in a statistically significant difference when using
358 miRglmm but not for NB GLM where the estimated effect is similar but less precise. All miRNA
359 represented in Figure 5 with the exception of hsa-miR-143-5p (Figure 5B) were found to have significant
360 random slope variability.

361

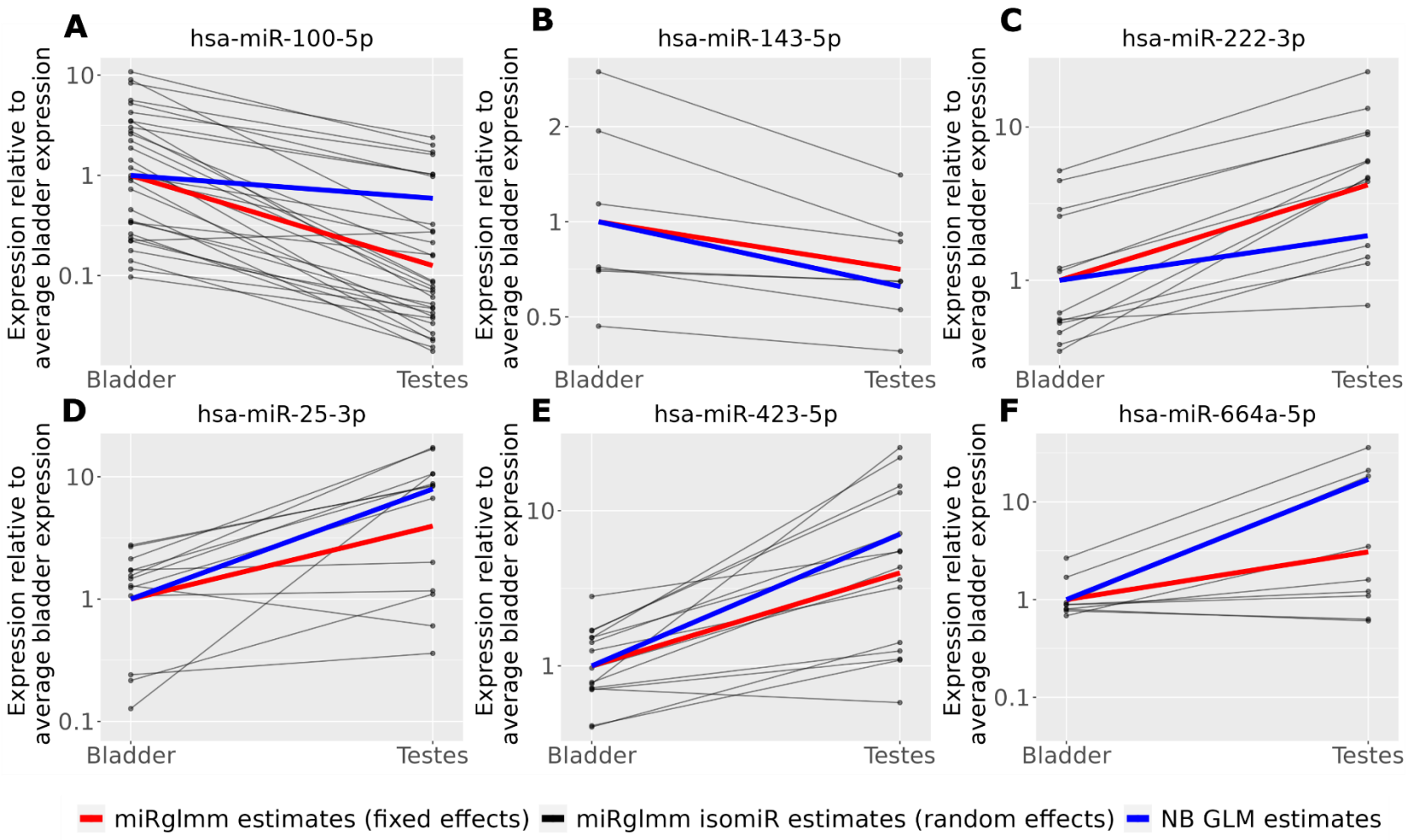
362

363

364

365

366



367

368 Figure 5: miRglmm estimates of relative expression are compared to estimates of relative expression from aggregated-count NB GLM models.

369 We obtain estimates of isomiR variability within miRNA from the random effects components of miRglmm. Panels A-C represent miRNA found to

370 be differentially expressed using miRglmm but not found by any of the aggregation methods. Panel D-F represent miRNA found to be

371 differentially expressed using all aggregation methods but not found by miRglmm. With the exception of hsa-miR-143-5p (panel B), all miRNA

372 presented have significant isomiR variability in the bladder vs testes effect.

373 **Discussion**

374 With miRglmm, we address the need for a DE tool specifically suited for miRNA-seq data that can also
375 utilize information from isomiRs in estimating miRNA-level DE estimates (10). Commonly used DE tools,
376 such as DESeq2 (15), edgeR (16) and limma-voom (17), were built for analysis of mRNA-seq data, and
377 some key features of miRNA-seq violate the independence between features assumption of these
378 methods. Additionally, these tools run on aggregated miRNA-level counts, resulting in the loss of
379 information contained in individual isomiRs. There is evidence of the biological importance of isomiRs
380 (11), so miRglmm allows for the discovery of heterogeneous DE among isomiRs of the same miRNA via
381 estimation of the random group effect for isomiR (and associated likelihood ratio test).

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

389 A limitation of miRglmm is that it does not aim to distinguish a true biological isomiR from a technically
390 arising isomiR and treats all isomiRs within a miRgroup equally. It has been shown that differences in
391 library preparation can induce bias in quantification of isomiRs (30, 31). By filtering lowly expressed
392 isomiRs, we eliminate some technical variability due to background count alignment. Our study designs
393 in the DE analysis of immune cell types as well as bladder vs testes tissue were implemented to minimize
394 technical variation by selecting samples processed by a single laboratory using consistent library
395 preparation techniques and sequencing equipment. Importantly, we observed the majority of miRNA

396 did have significant isomiR variability between cell types / tissues in both data sets where we assume we
397 are capturing primarily biological variability. If we were to analyze larger datasets that encompass
398 multiple studies, we would expect even higher rates of miRNA with significant isomiR variability, though
399 some of this would be purely technical. In this case, miRglmm is flexible and can adjust for technical
400 design variables, such as laboratory, library preparation method, or sequencer, as we did in analyzing
401 the ERCC benchmark data set (29).

402 **Conclusions**

403 We provide a new method and analysis tool, miRglmm, that uses isomiR variability to improve
404 differential expression analysis of miRNAs from miRNA-seq datasets. miRglmm provides superior
405 performance to alternative DE tools, whether or not significant isomiR variability exists, and estimates
406 both miRNA-level and isomiR-level differential expression.

407 **Methods**

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

409 The statistical model implemented in miRglmm is as follows. Let C_{ij} denote the count for sample $i =$
410 $1, \ldots, n$ 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
411 offset term in the model to adjust for variable sequencing depth across samples. For each sample i, let
412 X_i be a p-vector of covariates. For each miRNA, miRglmm fits the following negative binomial mixed
413 model (NBMM):

414 $$C_{ij} \sim \mathrm{NB}\left(\mu_{ij}, \Theta\right)$$

415 $$log\left(\mu_{ij}\right) = log(T_i) + \left(\beta + \tau_{1j}\right)X_i + \tau_i + \tau_{0j}$$

416 $$\begin{bmatrix}\tau_{0j} \\ \tau_{1j}\end{bmatrix} \sim N(0, \Omega_i), \Omega_i = \begin{bmatrix}\sigma_{0j}^2 & \sigma_{01j} \\ \sigma_{01j} & \sigma_{1j}^2\end{bmatrix}$$

417 $$\tau_i \sim N(0, \sigma_1^2)$$

418 β is the fixed effect of primary interest. τ_i is the random intercept term for sample i, τ_{0j} is the random
419 intercept for isomiR j, and τ_{1j} is the random slope term for isomiR j. The random effects for sample and
420 isomiR are independent, while the intercept and slope random effects for isomiR have a dependence
421 structure specified by Ω. The variance of the random slope σ_{1j}^2 can be used to assess variable DE
422 between isomiRs. We built miRglmm using the function glmer.nb in the lme4 package, which fits an
423 NBMM via a Laplace approximation to the maximum likelihood with a variety of optimizer choices (32).
424 The analyses presented in this paper are reproducible using code in the GitHub repository found at
425 https://github.com/mccall-group/miRglmm-paper.

426 **Software, data structures, and outputs**

427 miRglmm is an R library that consists of one core function that can be easily integrated into DE analysis
428 pipelines by replacing DE methods designed for mRNA-seq (such as DESeq2, edgeR, limma/voom).
429 IsomiR-level count matrices in the form of core Bioconductor structures, SummarizedExperiment
430 objects, are taken as input, and a list of model fit summaries for each miRgroup analyzed is returned.
431 The function can be run in parallel across miRgroups and can handle flexible design matrices. Additional
432 functions are included in the package to extract and summarize values of interest from the model fit
433 summaries, including miRNA-level estimates of DE with confidence intervals, isomiR-level estimates of
434 expression within a miRgroup, estimates of variability in DE across isomiRs within a miRgroup, and a
435 likelihood ratio test for significant isomiR variability within a miRgroup. The miRglmm package vignette
436 includes examples of how to import isomiR-level count data from either miRge (33) or sRNAbench (34).

437 **The microRNAome data resource**

438 The miRNAome dataset was assembled to more fully understand miRNA expression patterns across
439 primary cell types (35, 36). The dataset was built upon 2,077 samples from 175 public datasets across

440 196 primary cell types. miRNA annotation and quantification was performed using the miRge3.0 pipeline
441 (33). Briefly, miRge3.0 is a multi-step miRNA alignment program. From a FASTQ file, miRge3.0 collapses
442 identical sequences and processes them through repeated Bowtie alignment steps to identify canonical
443 miRNAs, isomiRs, and other RNA species. In addition to a final read count per miRNA, it outputs an
444 alignment file containing counts of all aligned isomiRs, which was used here.

445 **Filtering of lowly expressed miRNAs and isomiRs**

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

453 **Inducing known effects in real biological data to simulate differential expression**

454 We created ground truth data by inducing a known artificial effect into real biological data, allowing us
455 to assess the performance of miRglmm under conditions seen in real data. We searched the miRNAome
456 data (35) for one cell type with large sample counts coming from the same study to have a relatively
457 homogenous starting dataset. We used 39 monocyte samples from one study (27) and retained 122
458 miRNA with sufficient expression to model using a log(median CPM) cutoff of 5, as described above.

459 To induce an artificial “group” effect, we randomly split the samples into 2 groups, with 19 samples
460 labelled as Group A and 20 samples labelled as Group B (Supplemental Figure 12). Of 122 total miRNA,
461 we used stratified sampling to select 20 miRNA to be overexpressed in Group A, and another 20 miRNA
462 to be underexpressed in Group A. The sampling was stratified by total miRNA expression to manipulate

463 miRNA across the full range of expression values. We used the default log(median CPM) cutoff of -1 to
464 retain isomiRs for analysis. For the 20 miRNA overexpressed in Group A, we multiplied the counts of all
465 retained isomiRs in the miRgroup by a random truncated normal variable with mean value of 2, variance
466 of 1, lower bound of 1 and upper bound of 3 in Group A samples only. These 20 miRNAs now have a
467 known miRNA-level fold change (B vs A) = 0.5, with random variability between isomiRs. For the 20
468 miRNA underexpressed in Group A, we multiplied the counts of all retained isomiRs in the miRgroup by
469 a random truncated normal variable with the same parameters as above but this time in Group B
470 samples only. These 20 miRNAs now have a known miRNA-level fold change (B vs A) = 2, with random
471 variability between isomiRs. Importantly, we recalculated the total counts used in the offset term after
472 the artificial signal is induced. The stratification used in sampling miRNA ensures the effect of the signal
473 being added is consistent across groups, even though the total counts increase for all samples. The
474 entire procedure, including randomly splitting samples into two groups, was repeated 100 times.

475 For each miRNA, we estimated a miRNA-level differential group effect, measured via log fold-change
476 (logFC) estimates, using miRglmm. We compared these estimates to differential group effects produced
477 by other commonly used differential expression tools used on aggregated data: DESeq2 (15), edgeR (16),
478 and limma-voom (17). We also included a Negative Binomial Generalized Linear Models (NB GLM), fit
479 using glm.nb from the MASS R package (28), which is similar to the miRglmm model but without random
480 effects that can model the isomiR variability. Data was aggregated to the miRNA-level prior to running
481 NB GLM, DESeq2, edgeR and limma-voom, and hence we collectively call these methods “aggregation
482 methods”. Aggregated count values were produced by summing counts from all isomiRs that align to a
483 given miRNA.

484 We used Mean Squared Error (MSE) and 95% confidence interval coverage proportions to assess the
485 performance. The MSE compared the estimated logFC to the induced effect for each miRNA, and the
486 coverage proportion assessed the proportion that the true logFC fell in the 95% confidence interval

487 estimated by the model. We cannot estimate a confidence interval coverage proportion for edgeR as
488 that algorithm does not provide standard error (SE) estimates. We also separately looked at each
489 method's MSE and coverage when the effect was induced up or down, and those where no effect was
490 induced.

491 To identify miRNA with significant variability in the group effect between isomiRs, we tested whether
492 the random slope parameter τ_{1j} in the miRglmm model contributes significant information with a 1-
493 degree of freedom (1-df) Likelihood Ratio Test (LRT) comparing likelihoods from the model specified
494 above and a model removing only the τ_{1j} parameter. With this test, we can identify miRNA with
495 significant random slope effects and summarize by induced effect groups.

496 **Analysis of a benchmark experiment using synthetic miRNA pools**

497 We used experimental benchmark data with known expression differences to assess the performance of
498 miRglmm and existing methods to estimate DE. The Extracellular RNA Communication Consortium
499 (ERCC) was established to facilitate expansion of the field of extracellular RNA (exRNA) biology and
500 consisted of collaborative projects to develop robust methods for isolation and analysis of exRNA data
501 (37). One of these projects was a multi-protocol, multi-institution assessment of the bias of four sRNA-
502 seq library preparation methods, using ratiometric pools of synthesized small RNAs (29). Chemically
503 synthesized RNA oligonucleotides were added in varying ratios to pool A and pool B, from 10 to 1 and 1
504 to 10, leading to 15 levels of DE (from logFC = -2.3 to logFC = 2.3), providing a known fold change value
505 for each miRNA that can be used to evaluate the accuracy of miRglmm and other methods
506 (Supplemental Figure 13).

507 The ERCC sequence runs with 4N method (4 random nucleotides on both ends of the reads) were
508 chosen (n=104 runs). The reference sequence database consisting of 286 human miRNAs and 48 other
509 spikein miRNAs were indexed using bowtie-index. The reads were processed for illumina adapters

510 'TGGAATTCTCGGGTGCCAAGGA' using cutadapt (38) followed by alignment using bowtie aligner (39).
511 The parameters for bowtie include "No mismatch (-n 0)", "Trim 4 nucleotides on both ends (-5 4 -3 4)",
512 "Avoid alignment against the reverse-complement reference strand (--norc)", and output in SAM format
513 (-S). The SAM files were processed for read counts across each mapped miRNA and to account for PCR
514 duplicates (4N-based) using custom Python scripts. The miRNA counts were used for the downstream
515 analysis in miRglmm.

516 The ERCC data was processed by 6 different laboratories (Lab), and we used dimensionality reduction
517 using non-metric multidimensional scaling (NMDS) to assess if there was sufficient laboratory variability
518 to require adjustment for Lab in the analysis. The samples separated along the first dimension by Lab,
519 and the second dimension appeared to capture the Pool effect (Supplemental Figure 14A). We included
520 a fixed effect for Lab in the models for all methods to adjust for the technical effect of the differences in
521 lab-specific sample handling, processing, and sequencer on the counts. The ERCC dataset included
522 synthetic RNAs that were much longer at 50-90 nucleotides than the typical 18-24 nucleotide miRNAs.
523 We excluded any RNA specified to have a length > 45, resulting in a set of 303 small RNAs.

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

534 distributions separated around a log(median CPM) value of -1 (Supplemental Figure 14B). We ran
535 miRglmm without any sequence filtering, and also compared the performance of miRglmm under a
536 range of reasonable filters (log(median CPM)>-1 (default) to 2). We also assessed the effect of filtering
537 prior to aggregation when using the aggregation methods.

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

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

544 Our goal in selecting samples for these analyses was to minimize possible technical variation, and
545 therefore capture true biologically relevant isomiR-level differences. We chose bladder and testes
546 tissues to compare because they represented a large set of tissues available from a single study (25),
547 where sample processing and sequencer would be consistent across samples. Additionally, we chose a
548 set of immune cell types that were also derived from a single study as our cell type differential
549 expression analysis set (27). We aimed to produce miRNA-level differential expression (DE) estimates
550 using miRglmm and compare these results to other commonly used differential expression tools that
551 run on aggregated miRNA-level count data.

552 For each miRNA, we estimated a miRNA-level differential group (tissue or cell type) effect, measured via
553 log fold-change (logFC) estimates, using miRglmm. We compared these estimates to differential group
554 effects produced by existing aggregation methods using an intraclass correlation coefficient (ICC).
555 Significance at a Benjamini Hochberg False Discovery Rate (FDR) of 0.05 was compared using Cohen's
556 Kappa. We also used an upset plot to simultaneously compare miRNA considered differentially

557 expressed between all 5 methods. We can produce isomiR-level estimates of expression within group by
558 summing fixed and random effects for each isomiR. We utilized LRT as previously described to identify
559 miRNA within which there is significant variability in the group effect between isomiRs.

560

561 **Declarations**

562 **Ethics approval and consent to participate**

563 Not applicable

564 **Consent for publication**

565 Not applicable

566 **Availability of data and materials**

567 Biological sample metadata used in this manuscript were accessed through the microRNAome database
568 (35, 36). The original raw count data for the specific studies analyzed can be accessed via the sequence
569 read archive (SRA) under the following accession numbers: SRP110505 (27) and SRP007946 (25).
570 Synthetic data used in this manuscript were accessed via SRA accession number SRP098949 (29).
571 Processed data files used in this manuscript can be found on Zenodo (10.5281/zenodo.12571079).

572 **Competing interests**

573 The authors declare that they have no competing interests.

574 **Funding**

575 This work was supported by NIH grant R01GM139928.

576 **Authors' contributions**

577 MNM and MKH conceived the project. AMB and MNM analyzed data. MKH interpreted results. AMB,
578 AP, EAP and MNM developed the software. AMB and MNM wrote the manuscript. All authors read,
579 edited and approved the final manuscript.

580 **Acknowledgements**

581 We acknowledge the helpful feedback and discussion from other members of the McCall research
582 group.

583 1. Ambros V. The functions of animal microRNAs. Nature. 2004;431(7006):350-5.
584 2. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell. 2004;116(2):281-
585 97.
586 3. Zhang B, Farwell MA. microRNAs: a new emerging class of players for disease diagnostics and
587 gene therapy. Journal of Cellular and Molecular Medicine. 2008;12(1):3-21.
588 4. Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. Cell. 2012;148(6):1172-
589 87.
590 5. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation
591 by miRNAs and siRNAs. Genes Dev. 2006;20(5):515-24.
592 6. Baras AS, Mitchell CJ, Myers JR, Gupta S, Weng LC, Ashton JM, et al. miRge - A Multiplexed
593 Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy. PLoS One.
594 2015;10(11):e0143066.
595 7. Tomasello L, Distefano R, Nigita G, Croce CM. The MicroRNA Family Gets Wider: The IsomiRs
596 Classification and Role. Frontiers in Cell and Developmental Biology. 2021;9.
597 8. Ameres SL, Zamore PD. Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol.
598 2013;14(8):475-88.
599 9. Yang A, Bofill-De Ros X, Stanton R, Shao T-J, Villanueva P, Gu S. TENT2, TUT4, and TUT7
600 selectively regulate miRNA sequence and abundance. Nature Communications. 2022;13(1):5260.
601 10. Schmauch E, Attia Y, Laitinen P, Turunen TA, Bartos P, Vaananen M-A, et al. The critical role of
602 isomiRs in accurate differential expression analysis of miRNA-seq data. bioRxiv.
603 2024:2024.03.28.587190.
604 11. Bofill-De Ros X, Yang A, Gu S. IsomiRs: Expanding the miRNA repression toolbox beyond the
605 seed. Biochim Biophys Acta Gene Regul Mech. 2020;1863(4):194373.
606 12. Tan GC, Chan E, Molnar A, Sarkar R, Alexieva D, Isa IM, et al. 5' isomiR variation is of functional
607 and evolutionary importance. Nucleic Acids Res. 2014;42(14):9424-35.
608 13. Yang A, Bofill-De Ros X, Shao TJ, Jiang M, Li K, Villanueva P, et al. 3' Uridylation Confers miRNAs
609 with Non-canonical Target Repertoires. Mol Cell. 2019;75(3):511-22.e4.
610 14. Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I. Knowledge about the
611 presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer
612 types. Nucleic Acids Res. 2017;45(6):2973-85.
613 15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq
614 data with DESeq2. Genome Biol. 2014;15(12):550.
615 16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
616 expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

617 17. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools
618 for RNA-seq read counts. Genome Biology. 2014;15(2):R29.
619 18. Jun S-H, McCall M. Statistical Modeling for MicroRNA Sequencing Data. bioRxiv.
620 2024:2024.05.07.592964.
621 19. Booth JG, Casella G, Friedl H, Hobert JP. Negative binomial loglinear mixed models. Statistical
622 Modelling. 2003;3(3):179-91.
623 20. Witwer KW, Halushka MK. Toward the promise of microRNAs - Enhancing reproducibility and
624 rigor in microRNA research. RNA Biol. 2016;13(11):1103-16.
625 21. Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK, Yi N. Negative binomial mixed models for
626 analyzing microbiome count data. BMC Bioinformatics. 2017;18(1):4.
627 22. Zhang X, Pei Y-F, Zhang L, Guo B, Pendegraft AH, Zhuang W, Yi N. Negative Binomial Mixed
628 Models for Analyzing Longitudinal Microbiome Data. Frontiers in Microbiology. 2018;9.
629 23. Tsonaka R, Spitali P. Negative Binomial mixed models estimated with the maximum likelihood
630 method can be used for longitudinal RNAseq data. Brief Bioinform. 2021;22(4).
631 24. He L, Davila-Velderrain J, Sumida TS, Hafler DA, Kellis M, Kulminski AM. NEBULA is a fast
632 negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject
633 single-cell data. Communications Biology. 2021;4(1):629.
634 25. Li X, Chen J, Hu X, Huang Y, Li Z, Zhou L, et al. Comparative mRNA and microRNA expression
635 profiling of three genitourinary cancers reveals common hallmarks and cancer-specific molecular events.
636 PLoS One. 2011;6(7):e22570.
637 26. Silverman JD, Durand HK, Bloom RJ, Mukherjee S, David LA. Dynamic linear models guide design
638 and analysis of microbiota studies within artificial human guts. Microbiome. 2018;6(1):202.
639 27. Juzenas S, Venkatesh G, Hübenthal M, Hoeppner MP, Du ZG, Paulsen M, et al. A comprehensive,
640 cell specific microRNA catalogue of human peripheral blood. Nucleic Acids Res. 2017;45(16):9290-301.
641 28. Ripley WNVaBD. Modern Applied Statistics with S. Fourth ed: Springer; 2002.
642 29. Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, et al.
643 Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling.
644 Nature Biotechnology. 2018;36(8):746-57.
645 30. Wright C, Rajpurohit A, Burke EE, Williams C, Collado-Torres L, Kimos M, et al. Comprehensive
646 assessment of multiple biases in small RNA sequencing reveals significant differences in the
647 performance of widely used methods. BMC Genomics. 2019;20(1):513.
648 31. Gómez-Martín C, Aparicio-Puerta E, van Eijndhoven MAJ, Medina JM, Hackenberg M, Pegtel DM.
649 Reassessment of miRNA variant (isomiRs) composition by small RNA sequencing. Cell Rep Methods.
650 2023;3(5):100480.
651 32. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal
652 of Statistical Software. 2015;67(1):1 - 48.
653 33. Patil AH, Halushka MK. miRge3.0: a comprehensive microRNA and tRF sequencing analysis
654 pipeline. NAR Genom Bioinform. 2021;3(3):lqab068.
655 34. Aparicio-Puerta E, Gómez-Martín C, Giannoukakos S, Medina JM, Scheepbouwer C, García-
656 Moreno A, et al. sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for
657 model and non-model organisms. Nucleic Acids Res. 2022;50(W1):W710-w7.
658 35. McCall MN, Kim MS, Adil M, Patil AH, Lu Y, Mitchell CJ, et al. Toward the human cellular
659 microRNAome. Genome Res. 2017;27(10):1769-81.
660 36. Patil AH, Baran A, Brehm ZP, McCall MN, Halushka MK. A curated human cellular microRNAome
661 based on 196 primary cell types. Gigascience. 2022;11.
662 37. Das S, Ansel KM, Bitzer M, Breakefield XO, Charest A, Galas DJ, et al. The Extracellular RNA
663 Communication Consortium: Establishing Foundational Knowledge and Technologies for Extracellular
664 RNA Research. Cell. 2019;177(2):231-42.

665 38. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011.
666 2011;17(1):3.
667 39. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short
668 DNA sequences to the human genome. Genome Biology. 2009;10(3):R25.

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686 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**

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

688 isomiRs

689

690

691

692

693 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**

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

695 isomiRs

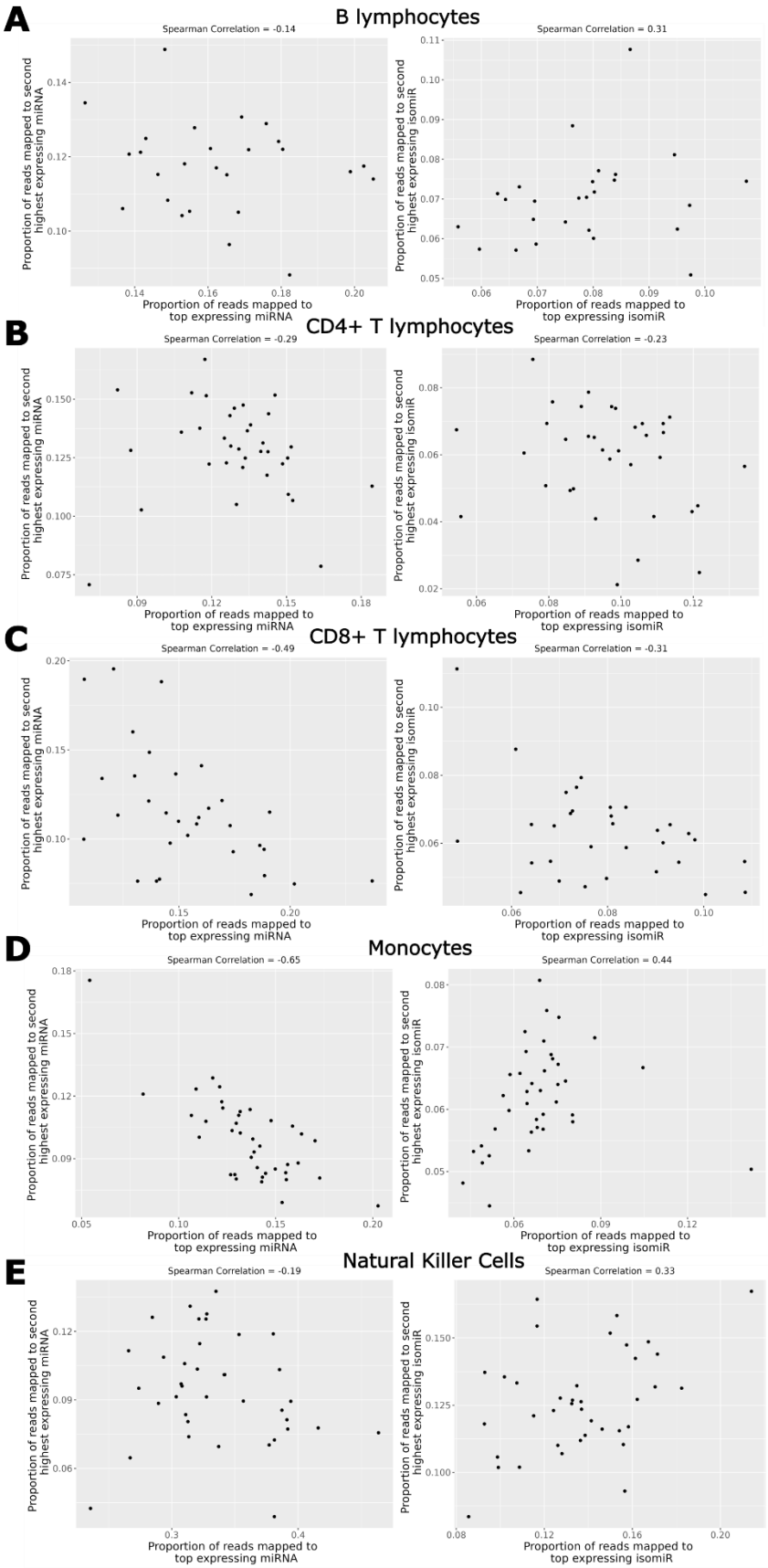
696

697

698

699

700


A
B lymphocytes
Spearman Correlation = -0.14
Spearman Correlation = 0.31
Proportion of reads mapped to second highest expressing miRNA
Proportion of reads mapped to second highest expressing isomiR
Proportion of reads mapped to top expressing miRNA
Proportion of reads mapped to top expressing isomiR
B
CD4+ T lymphocytes
Spearman Correlation = -0.29
Spearman Correlation = -0.23
C
CD8+ T lymphocytes
Spearman Correlation = -0.49
Spearman Correlation = -0.31
D
Monocytes
Spearman Correlation = -0.65
Spearman Correlation = 0.44
E
Natural Killer Cells
Spearman Correlation = -0.19
Spearman Correlation = 0.33

701

702 Supplemental Figure 1: Negative correlation bias is observed when considering the top two highest
703 expressing miRNA after aggregation to miRNA-level counts (left column, all panels), whereas a mix of
704 positive (right column, panels A, D and E) and negative correlations (right column, panels B and C) are
705 observed when considering the two highest expressing isomiRs.

706 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

707

708

709

710

711

712

713

714

715

716

717

718

719

720 Supplemental Table 4: Confidence interval coverage proportion summary statistics across 100

721 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

722

723

724

725

726

727

728

729

730

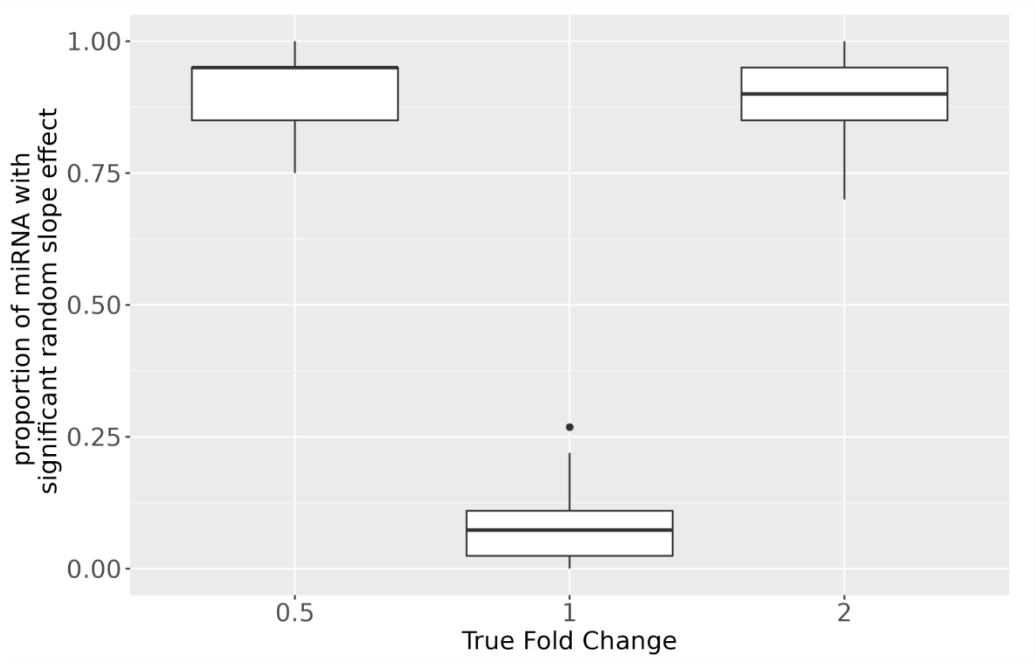
731

732

733

734

735



736

737 Supplemental Figure 2: The simulation procedure creates significant variability in the group effect across

738 isomiRs.

739

740

741

742

743

744

745

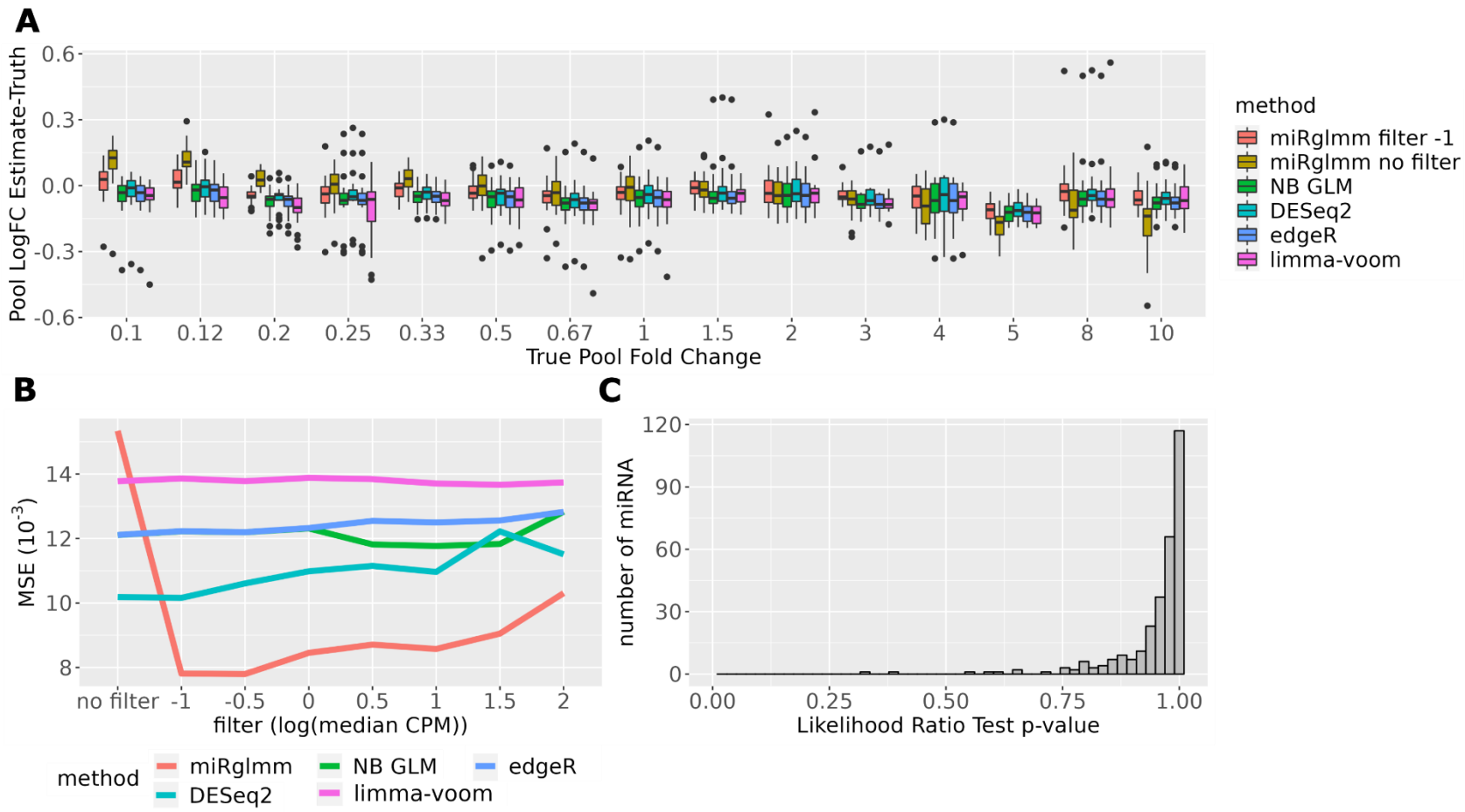
746

747

748 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 proportion
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

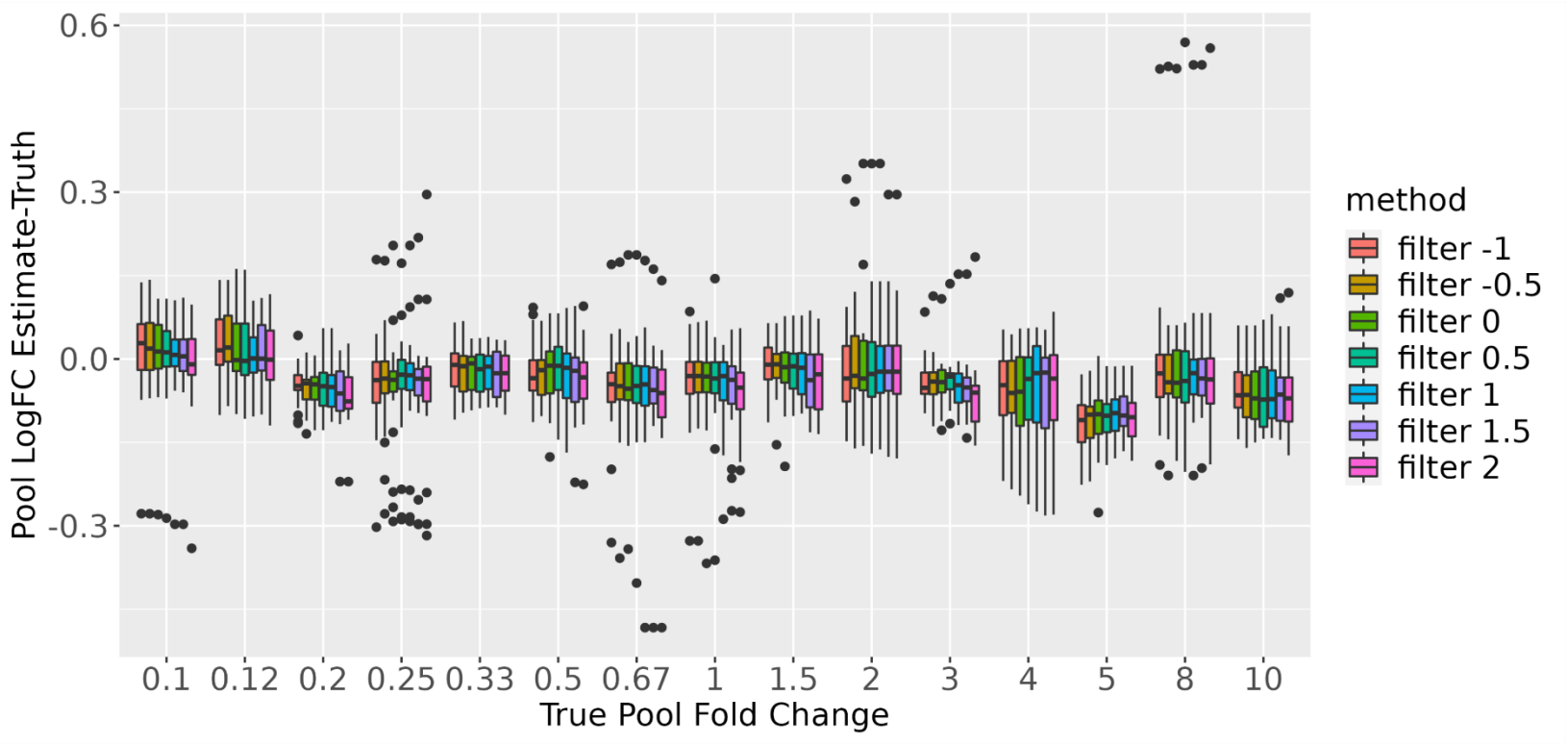
749 MSE: Mean Squared Error, NA: not applicable, *denotes default filter implemented in miRglmm



Supplemental Figure 3: 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 aggregation 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 aggregation

755 methods (panel B). There are no miRNA with significant variability in the Pool effect across isomiRs as measured via a Likelihood Ratio Test p-

756 value (panel C).



Supplemental Figure 4: 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.

762 Supplemental Table 6: Comparing performance by miRglmm filters and comparing to aggregation

763 methods when random slope effect removed

method	MSE (10^{-3})	number of miRNA estimated	minimum number of sequences	maximum number of sequences	coverage proportion
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

764

765

766

767

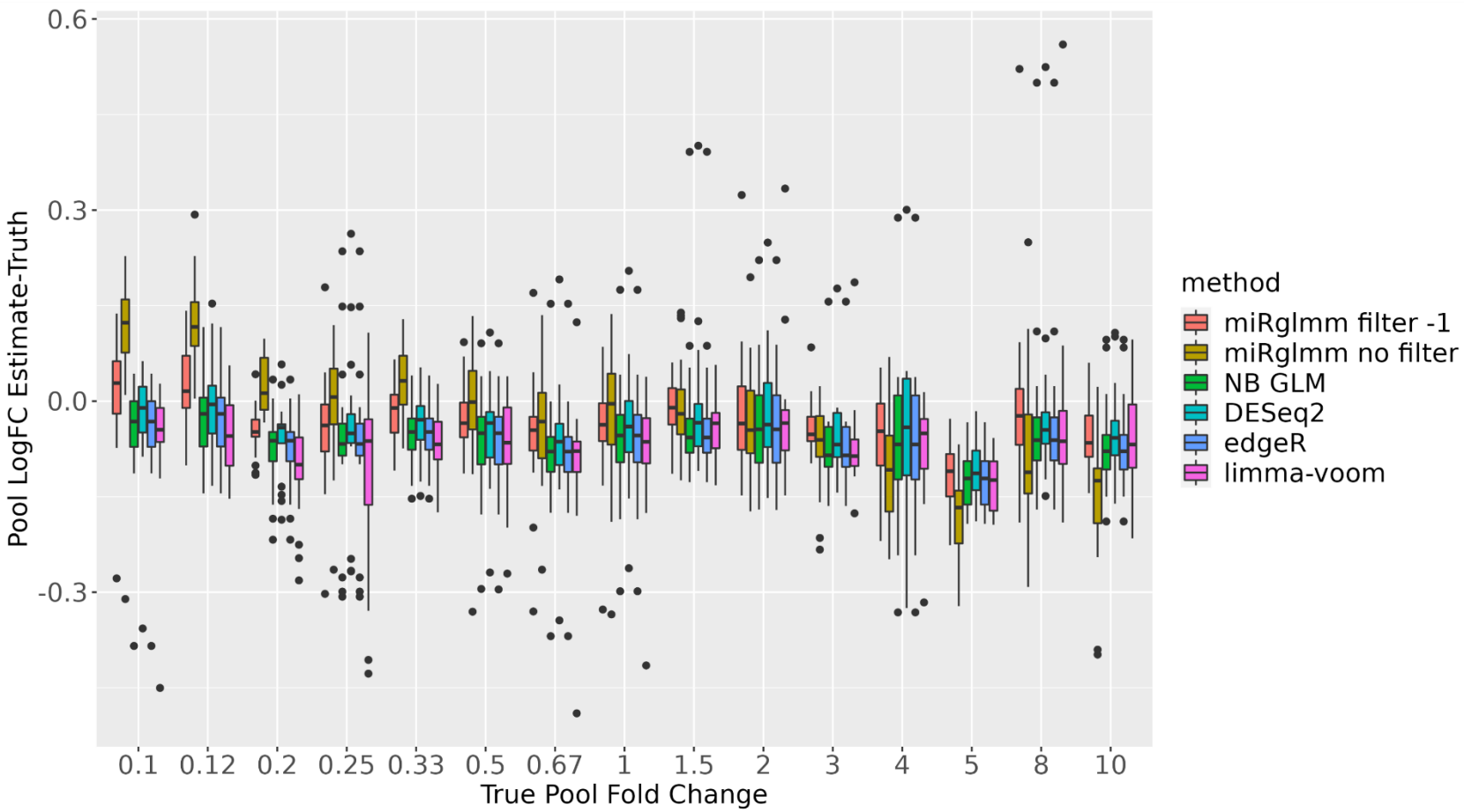
768

769

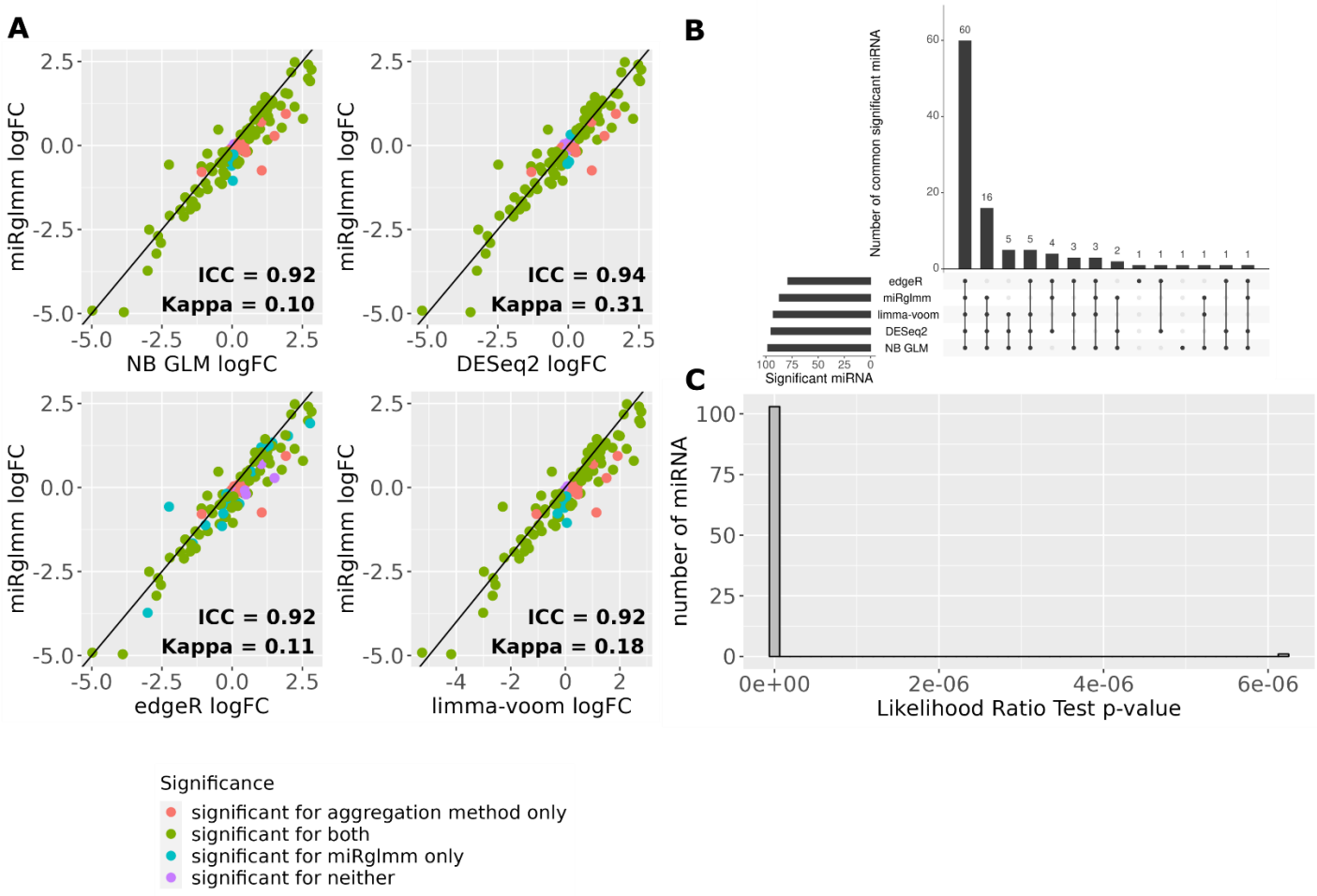
770

771

772



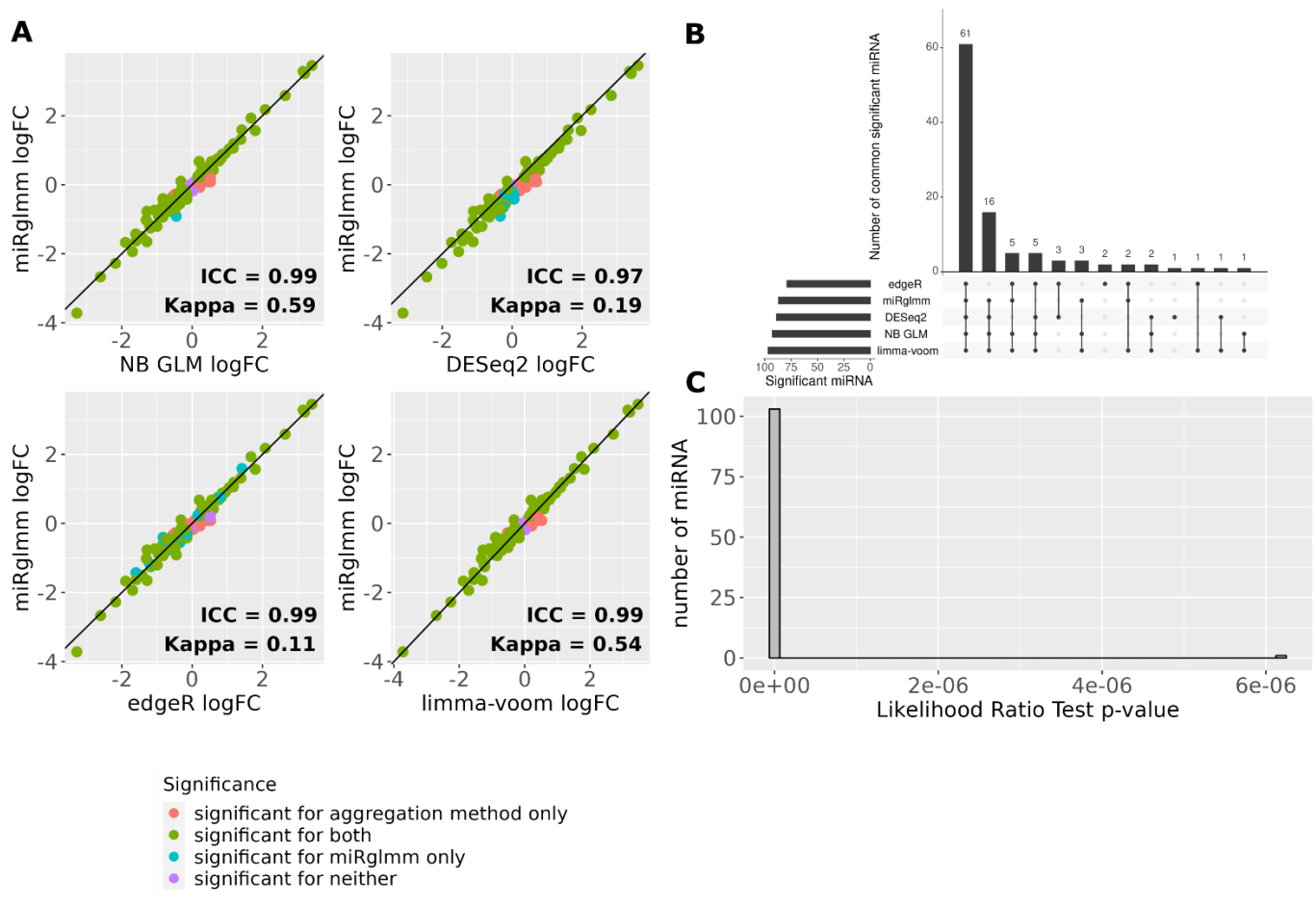
Supplemental Figure 5: The performance of miRglmm without the random Pool effect for isomiR is compared to the performance of aggregation methods. Log fold change (logFC) estimates of Pool B expression compared to Pool A expression are compared to known true values for each miRNA and summarized within each True Pool Fold Change group.



Supplemental Figure 6: Log fold change (logFC) estimates of miRNA-level expression in monocytes compared to B lymphocytes from miRglmm are plotted versus logFC estimates from each aggregation method (panel A). Each point represents the estimate for the monocyte vs B lymphocyte contrast for a unique miRNA. Significant DE vs B lymphocytes is indicated by the color (Benjamini Hochberg FDR of 0.05). Intra-class

781 correlation coefficients (ICC) are used to assess agreement in the values of the estimated logFC between methods. Cohen's kappa is used to
782 assess agreement in the significance of the estimate between methods. An upset plot shows the common number of miRNAs differentially
783 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
784 indicates that most of the miRNA have significant variability in the cell-type effect across isomiRs (panel C).

785



Supplemental Figure 7: Log fold change (logFC) estimates of miRNA-level expression in Natural Killer (NK) cells compared to B lymphocytes from miRglmm are plotted versus logFC estimates from each aggregation method (panel A). Each point represents the estimate for the NK vs B lymphocyte contrast for a unique miRNA. Significant DE vs B lymphocytes is indicated by the color (Benjamini Hochberg FDR of 0.05). 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

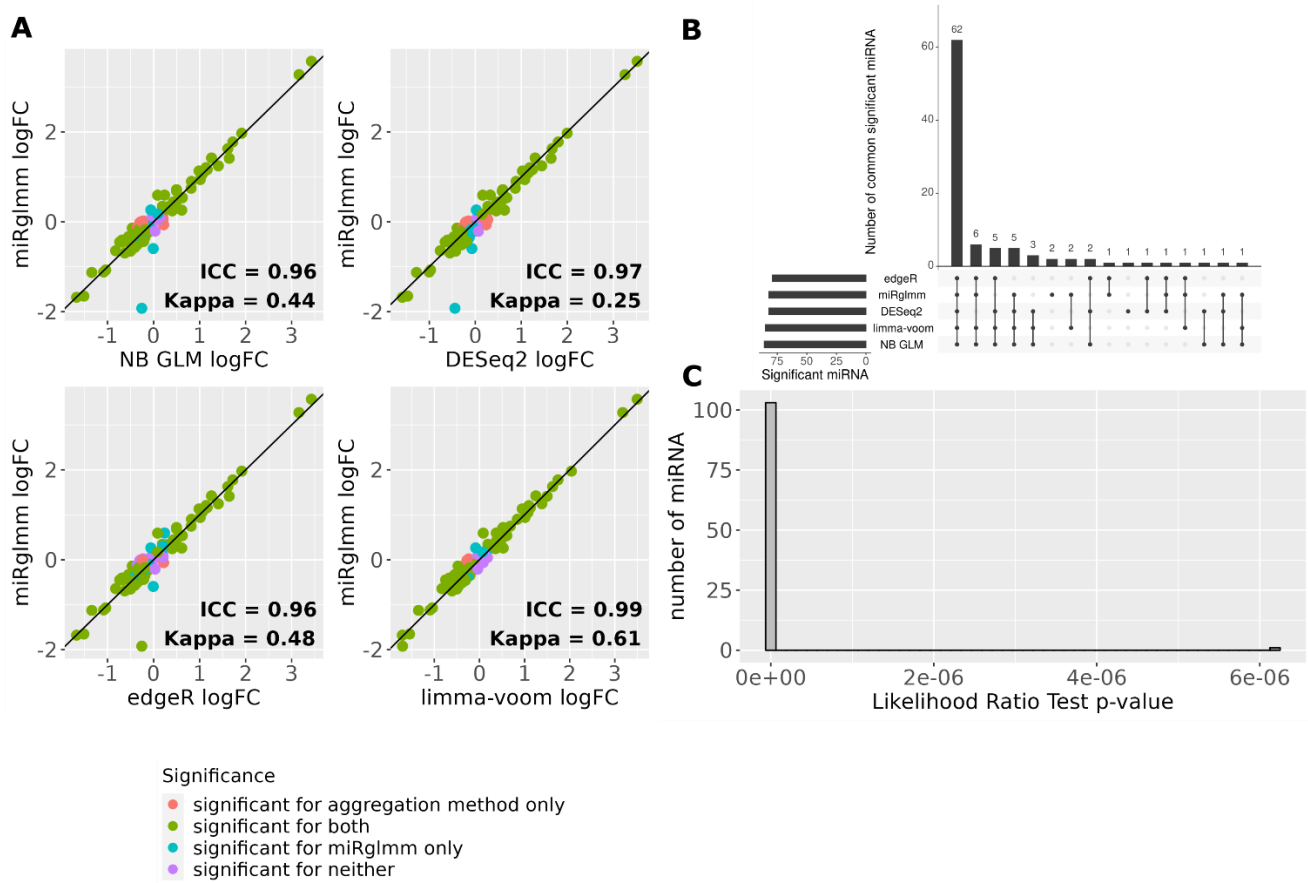
791 assess agreement in the significance of the estimate between methods. An upset plot shows the common number of miRNAs differentially

792 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

793 indicates that most of the miRNA have significant variability in the cell-type effect across isomiRs (panel C).

794

795



Supplemental Figure 8: Log fold change (logFC) estimates of miRNA-level expression in CD4+ T lymphocytes compared to B lymphocytes from miRglmm are plotted versus logFC estimates from each aggregation method (panel A). Each point represents the estimate for the CD4+ T vs B lymphocyte contrast for a unique miRNA. Significant DE vs B lymphocytes is indicated by the color (Benjamini Hochberg FDR of 0.05). 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

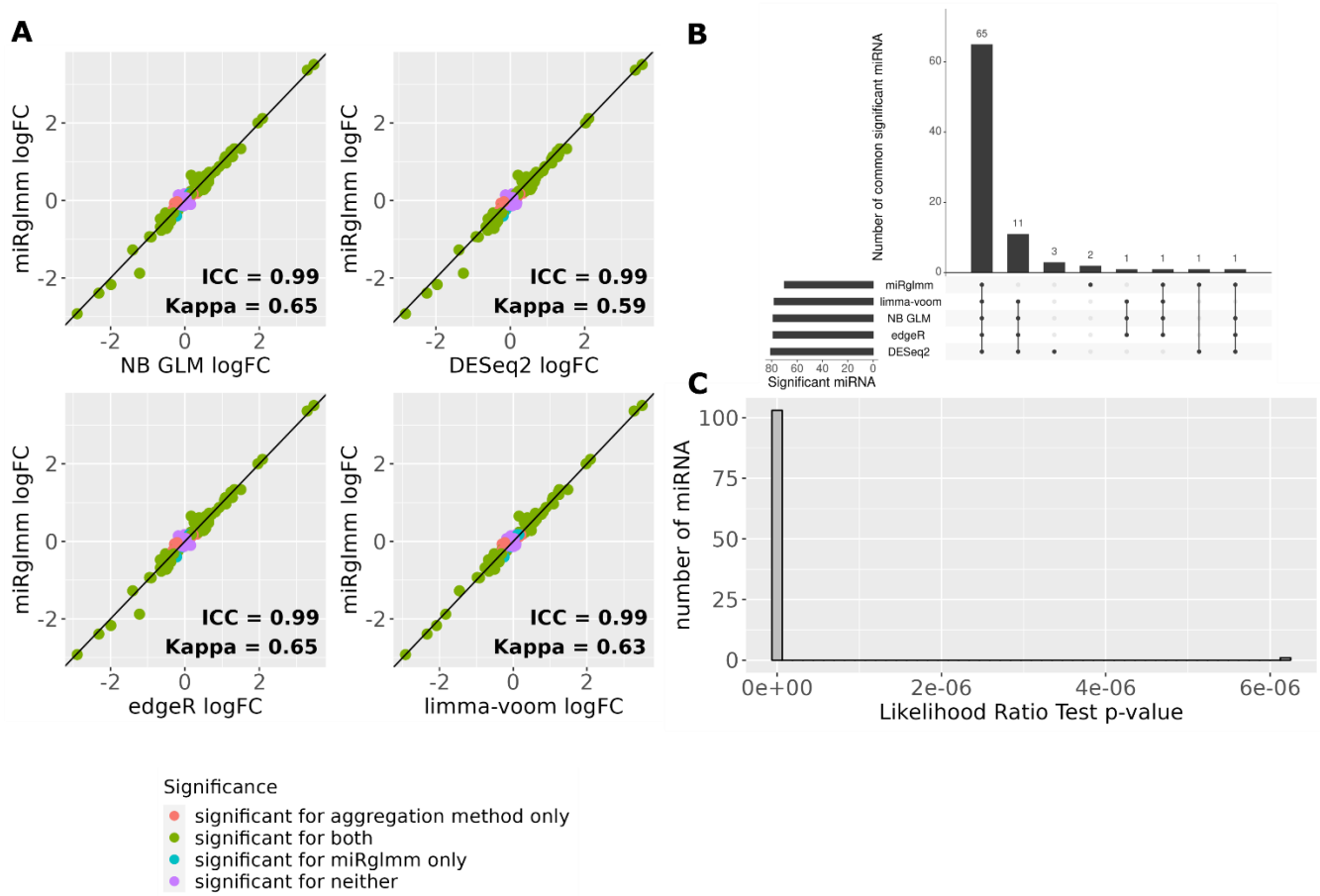
801 assess agreement in the significance of the estimate between methods. An upset plot shows the common number of miRNAs differentially

802 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

803 indicates that most of the miRNA have significant variability in the cell-type effect across isomiRs (panel C).

804

805



Supplemental Figure 9: Log fold change (logFC) estimates of miRNA-level expression in CD8+ T lymphocytes compared to B lymphocytes from miRglmm are plotted versus logFC estimates from each aggregation method (panel A). Each point represents the estimate for the CD8+ T vs B lymphocyte contrast for a unique miRNA. Significant DE vs B lymphocytes is indicated by the color (Benjamini Hochberg FDR of 0.05). 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

811 assess agreement in the significance of the estimate between methods. An upset plot shows the common number of miRNAs differentially

812 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

813 indicates that most of the miRNA have significant variability in the cell-type effect across isomiRs (panel C).

814 Supplemental Table 7: LogFC estimates for miRNA-contrast pairs that are significant for miRglmm but

815 not significant for any aggregation method

miRNA	Contrast (vs B lymphocyte)	miRglmm	NB GLM	DESeq2	edgeR	limma-voom
hsa-miR-32-5p	CD4 T lymphocyte	0.26	-0.06	0.02	-0.06	-0.08
hsa-miR-340-5p	CD8 T lymphocyte	-0.4	-0.22	-0.19	-0.22	-0.25
hsa-miR-423-3p	CD8 T lymphocyte	-0.15	-0.06	0	-0.06	-0.07
hsa-miR-423-5p	CD4 T lymphocyte	-0.38	-0.25	-0.15	-0.25	-0.26

816

817

818

819

820

821

822

823

824

825

826

827

828

829

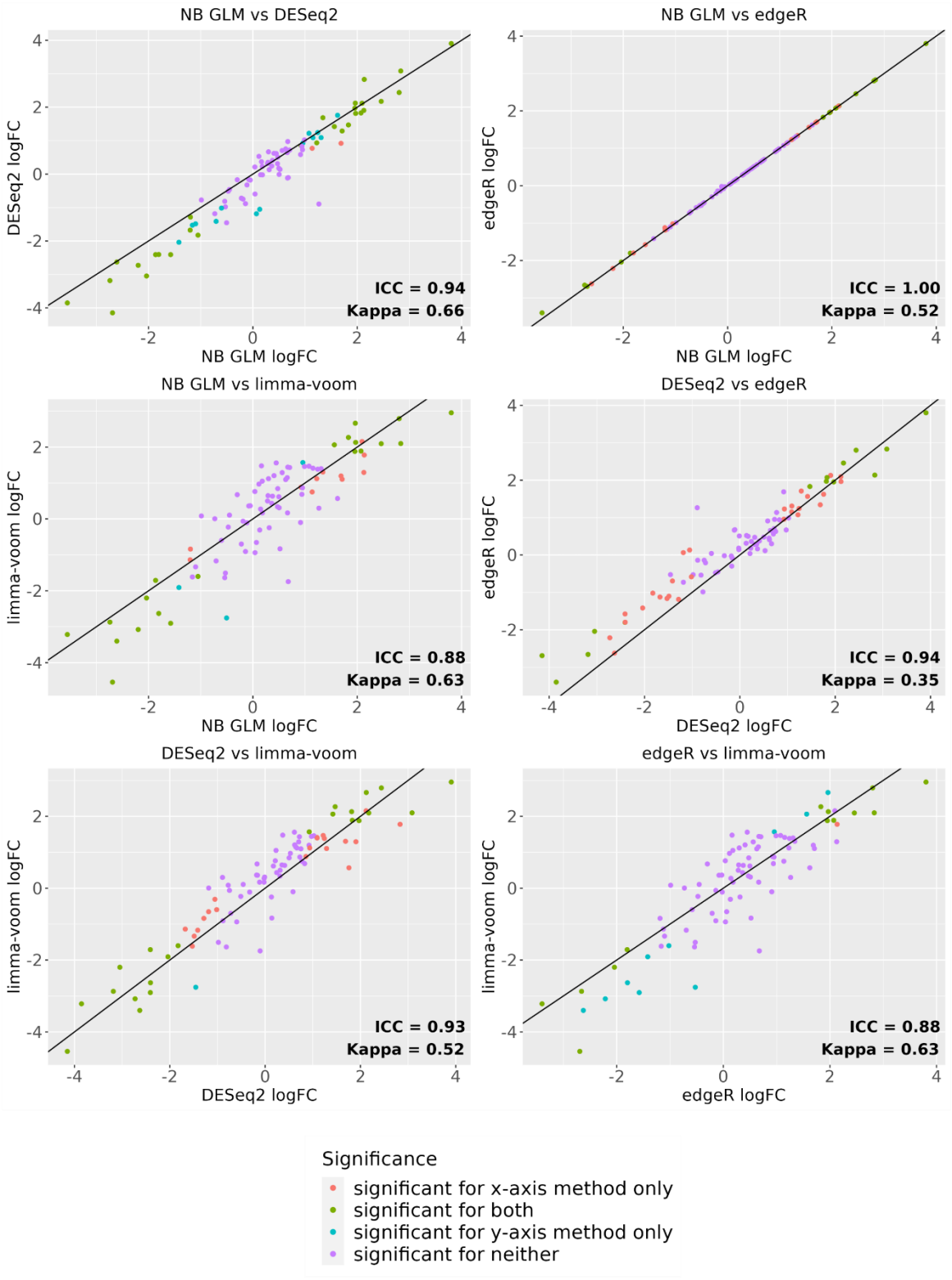
830 Supplemental Table 8: LogFC estimates for miRNA-contrast pairs that are not significant for miRglmm

831 but significant for all aggregation methods

miRNA	contrast	miRglmm	NB GLM	DESeq2	edgeR	limma-voom
hsa-let-7b-5p	Monocyte	0.09	0.45	0.22	0.45	0.47
hsa-miR-103a-3p	CD8 T lymphocyte	0.14	0.17	0.21	0.17	0.16
hsa-miR-106b-3p	CD8 T lymphocyte	-0.07	-0.28	-0.25	-0.28	-0.3
hsa-miR-107	Natural Killer	-0.07	0.22	0.39	0.22	0.22
hsa-miR-107	CD8 T lymphocyte	0.12	0.17	0.21	0.17	0.16
hsa-miR-140-3p	CD4 T lymphocyte	-0.1	-0.19	-0.12	-0.19	-0.18
hsa-miR-140-3p	CD8 T lymphocyte	-0.12	-0.22	-0.17	-0.22	-0.21
hsa-miR-148b-3p	Monocyte	0.94	1.91	1.69	1.91	1.92
hsa-miR-148b-3p	CD8 T lymphocyte	0.18	0.32	0.34	0.32	0.26
hsa-miR-181a-3p	Monocyte	-0.75	1.05	0.83	1.05	1.14
hsa-miR-181a-3p	CD4 T lymphocyte	-0.13	-0.28	-0.22	-0.28	-0.23
hsa-miR-181a-3p	CD8 T lymphocyte	-0.06	-0.27	-0.2	-0.27	-0.19
hsa-miR-181b-5p	CD4 T lymphocyte	0	-0.28	-0.2	-0.28	-0.23
hsa-miR-221-3p	Natural Killer	-0.27	-0.48	-0.31	-0.48	-0.53
hsa-miR-26b-3p	CD8 T lymphocyte	0.17	0.22	0.26	0.22	0.21
hsa-miR-28-5p	CD8 T lymphocyte	-0.05	-0.15	-0.11	-0.15	-0.17
hsa-miR-30d-5p	Monocyte	-0.04	0.33	0.1	0.33	0.33
hsa-miR-30d-5p	CD4 T lymphocyte	-0.07	-0.2	-0.13	-0.2	-0.19
hsa-miR-320a-3p	CD8 T lymphocyte	0.11	0.14	0.2	0.14	0.13
hsa-miR-339-3p	Monocyte	-0.79	-1.08	-1.31	-1.08	-1.06
hsa-miR-374a-3p	Natural Killer	0.09	0.43	0.6	0.43	0.4
hsa-miR-374a-5p	Natural Killer	0.05	0.19	0.36	0.19	0.18
hsa-miR-378a-3p	CD8 T lymphocyte	-0.23	-0.32	-0.27	-0.32	-0.3
hsa-miR-652-3p	CD4 T lymphocyte	-0.01	-0.23	-0.19	-0.23	-0.28
hsa-miR-652-3p	CD8 T lymphocyte	-0.03	-0.21	-0.17	-0.21	-0.23
hsa-miR-941	Natural Killer	0.08	0.52	0.7	0.52	0.52

832

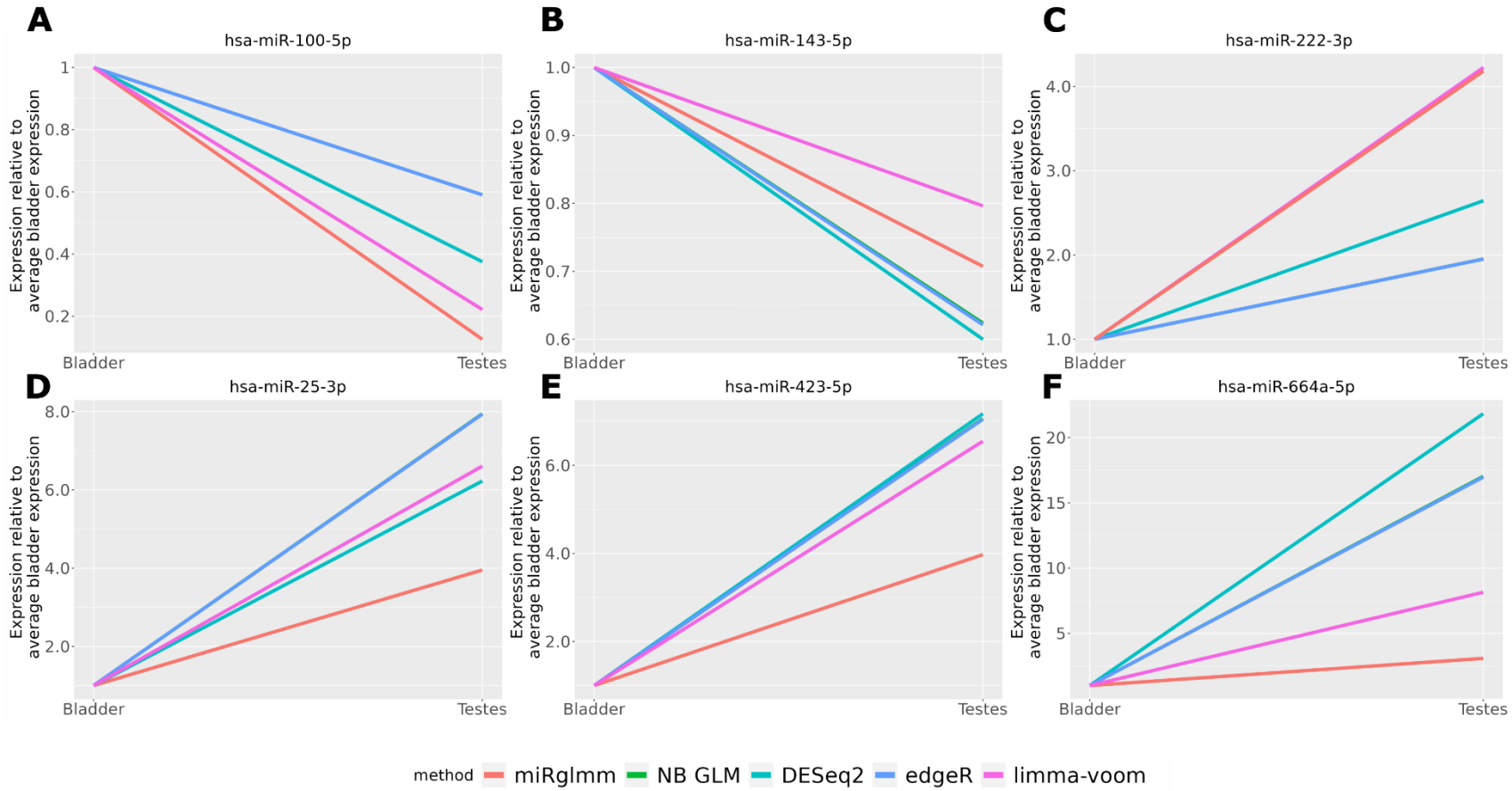
833


NB GLM vs DESeq2
DESeq2 logFC
NB GLM logFC
ICC = 0.94
Kappa = 0.66
NB GLM vs edgeR
edgeR logFC
NB GLM logFC
ICC = 1.00
Kappa = 0.52
NB GLM vs limma-voom
limma-voom logFC
NB GLM logFC
ICC = 0.88
Kappa = 0.63
DESeq2 vs edgeR
edgeR logFC
DESeq2 logFC
ICC = 0.94
Kappa = 0.35
DESeq2 vs limma-voom
limma-voom logFC
DESeq2 logFC
ICC = 0.93
Kappa = 0.52
edgeR vs limma-voom
limma-voom logFC
edgeR logFC
ICC = 0.88
Kappa = 0.63
Significance
significant for x-axis method only
significant for both
significant for y-axis method only
significant for neither

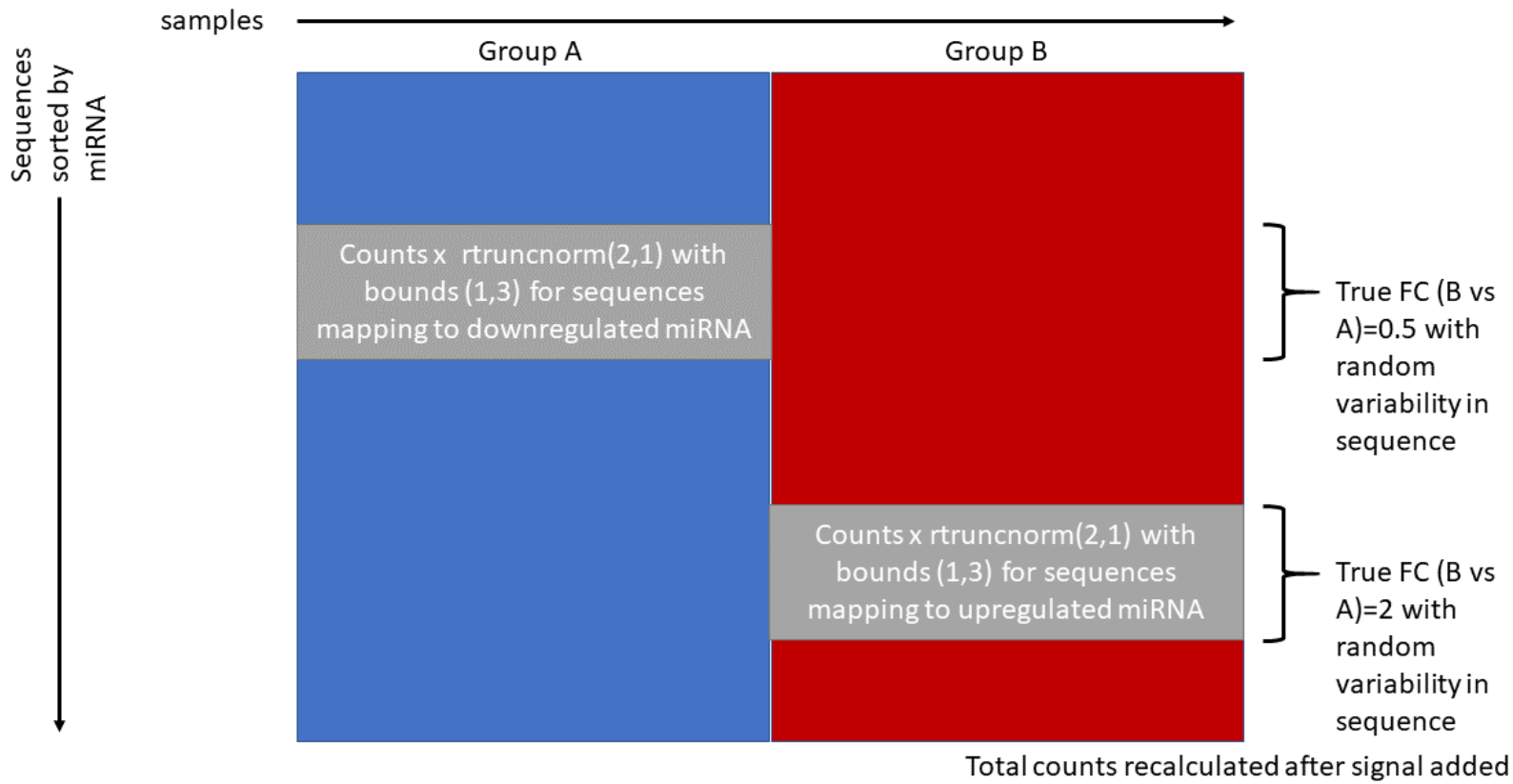
835 Supplemental Figure 10: Log fold change (logFC) estimates of expression in testes compared to bladder

836 are compared between aggregation methods with differences in significance (Benjamini Hochberg FDR

837 of 0.05) indicated by color.



Supplemental Figure 11: Estimates of testes expression relative to bladder are compared between miRglmm and all aggregation methods. miRglmm and limma-voom provide nearly identical estimates for hsa-miR-222-3p. NB GLM and edgeR provide nearly identical estimates for all miRNA.

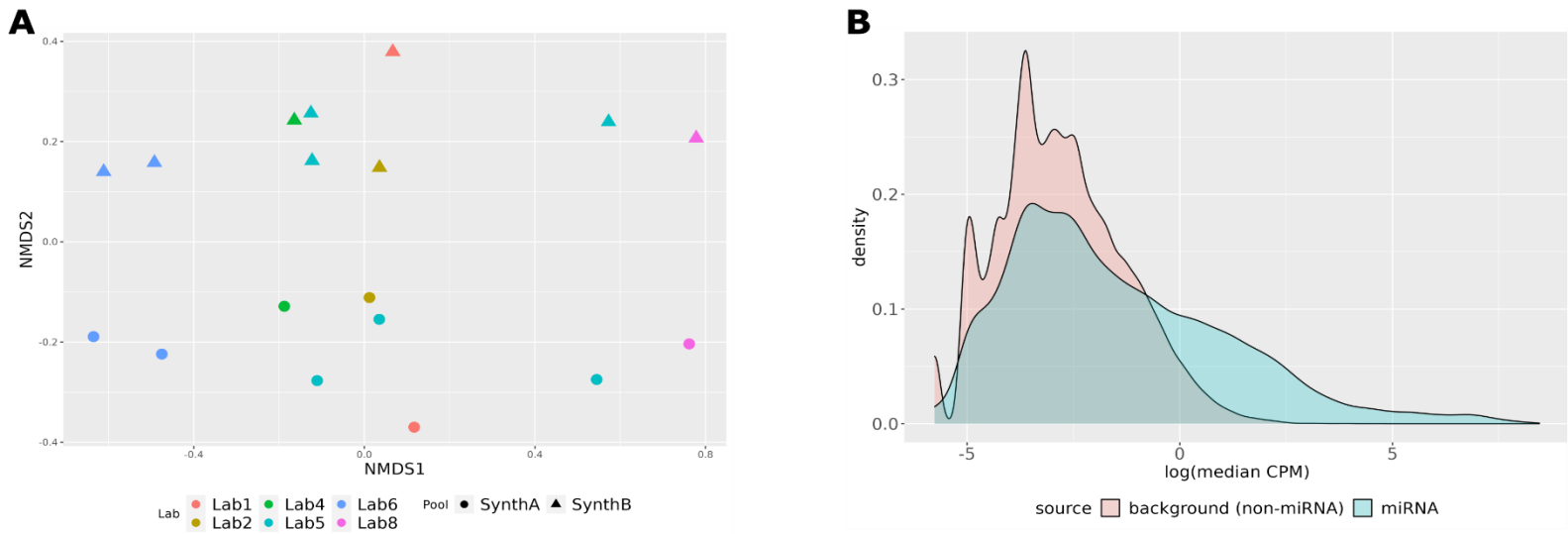


Supplemental Figure 12: Inducing an artificial group effect with random sequence variability into monocyte miRNA-seq data. The entire procedure from random splitting of samples into groups is repeated 100 times.

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 13: 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 14: 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).