The poor performance of TMM on microRNA-Seq

LANA X. GARMIRE^{1,3} and SHANKAR SUBRAMANIAM^{2,3}

¹Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, USA

²Department of Bioengineering, Department of Chemistry and Biochemistry, Department of Cellular and Molecular Medicine, and Department of Nano Engineering, University of California, San Diego, La Jolla, California 92093, USA

We published our third-party comparison of seven different normalization methods that were previously employed in microarray analysis, and TMM, a method developed by Robinson and Oshlack (2010) for RNA-Seq analysis, in the context of microRNA sequencing (miRNA-Seq). We used various evaluation metrics (MSE, KS statistics, ROC curves, linear regression, and differential expression test similarities) on two independent public miRNA-Seq profiling results that had qPCR results for validation. Based on the results that used relevant R packages *at the time of publication*, we found (Garmire and Subramaniam 2012) that quantile and lowess normalization worked the best on the two public data sets, whereas the normalization step documented in TMM, *at the time of manuscript preparation*, performed the worst among all methods compared.

This publication caught a tremendous amount of interest from Robinson and colleagues. They incorrectly stated many facts concerning both their software, which was made available publicly and subsequently revised, and our analysis, both in private communications as well as in their Divergent Views paper in *RNA*. We feel obligated to straighten the records and facts to the readers on several questions that Robinson and colleagues raised in their Divergent Views. For software that is actively being developed, such as edgeR, which uses the TMM method in their original paper, we look forward to additional improvements, such as the changes that are made in their code supplemented to their Divergent Views.

The authors first questioned the reproducibility of our analysis. To answer this question, we must point out that the authors changed the software and user guidelines since the preparation of our manuscript. The changes in MSE from TMM in their Supplemental Figure 1 come from the changes in their method. While preparing this publication, we actively sought the opinions of Robinson and colleagues given the results in Figure 3 of our recent publication (Garmire and Subramaniam 2012). We asked Robinson and colleagues *directly* if one should reverse the order of the numerator and denominator in calculating the normalization factor docu-

mented in their original *Genome Biology* paper (since we observed that doing so would improve MSE of TMM slightly). Unfortunately, Robinson and colleagues did not directly answer us, but we were pointed to their online example, which was consistent with their publication. We therefore used the normalization factor as described by Robinson and colleagues. This normalization factor was used *directly* in their earlier version of user guideline (dated October 14, 2010; a copy of this guideline is presented in Supplemental Material) to normalize the *count* data (shown by their code on pages 3 and 4 and Figure 1 on page 5 of this version of the guideline; screen shots of this code are shown below).

This record is contrary to their claim in their Divergent Views that “the normalization factors modify the *library* size, not the *count* data.” We were told by Robinson and colleagues after our publication, that by introducing an extra step to invert the normalization factor, as we had suspected in our prior e-mail inquiry to the authors, the MSE and KS metrics would improve their TMM method. For the record, we attach our reanalysis here on F-data (Fig. 2), which we had shared with Robinson and colleagues privately (Fig. 3 of our publication [left], and the result after adding an additional inverse step [right]). The new results of TMM still have poor performance based on this plot.

Robinson and colleagues questioned the suitability of using MSE and K-S metrics to evaluate miRNA-Seq data. First, we need to point out that we knew their limitation reasonably well (also discussed in our reference by Xiong et al.); we convinced the reviewers and the Editor with the following facts: First, MSE and K-S are not the sole criteria that we used; instead, we stressed the coherent results of several criteria, stated in the beginning of our Divergent Views. Second, we assume that the *majority* of the data in experimental conditions are not drastically different when there are no biological replicates, which is reasonable from the MA plots. Since the current version of TMM does not deliver the best performance in these data, we think it is not an optimal method when overall similarities are expected between samples, although TMM might be better when significant differences are expected between samples (e.g., comparisons of kidney and liver in their publication).

Robinson and colleagues questioned our omission of zero counts in the evaluation. Our publication centered on

³Corresponding authors
E-mail lgarmire@gmail.com
E-mail shankar@ucsd.edu

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.039271.113>.

```

> d <- calcNormFactors(d)
> d$samples
      group lib.size norm.factors
R1L1Kidney Kidney 1804977      1.209
R1L2Liver  Liver 1691734      0.821
R1L3Kidney Kidney 1855190      1.225
R1L4Liver  Liver 1696308      0.823

> eff.libsize <- d$samples$lib.size * d$samples$norm.factors
> maPlot(d$counts[, 1]/eff.libsize[1], d$counts[, 2]/eff.libsize[2],
+        normalize = FALSE, pch = 19, cex = 0.4, ylim = c(-8, 8))
> grid(col = "blue")

```

FIGURE 1. A snapshot of the code on pp. 3 and 4 in the edgeR user guideline (dated October 14, 2010).

evaluation with different metrics but not downstream analysis. Fold-change, MA plots, and miRNA-Seq versus qPCR comparison cannot be done with zero counts. Moreover, when no biological replicates exist (although this practical issue has been gradually changed with the popularity of the small RNA-Seq method and the price drop of NGS), uncertainty abounds for zero counts. Surely, one needs to examine carefully the samples that have zero count in one condition, but significant counts in other conditions. We did not propagate omission of zero counts in the downstream data analysis.

The authors also have reservations toward the rounding of transformed data back into integer counts, as well as using count-based statistical tests to normalize the rounded integer counts. The Fisher's test, binomial, Poisson, and χ^2 tests employed in our paper are widely used by the NGS community, when one has to deal with the reality that no biological replicates are done, such as in the data sets we used in the paper.

Regarding questioning of the quality of publicly available qPCR validation “truth” data sets we used for our publication, we will have to rely on the authors who generated the F-data and K-data. However, we want to point out that the “truth” data that Robinson and colleagues defined seem to have much smaller numbers than what we had used, and this can obscure the AUC interpretation based on limited “resolutions.”

In summary, our paper described a comprehensive effort to compare seven different normalization methods, of which the method of Robinson and colleagues is one, and our implementation and use of their method (as one part of our study) were based entirely on the then-available information on the authors' website. If they did, indeed, make an error then and subsequently corrected it, we are not at fault in our comparisons. Most significantly, even with their improvements, their method does not perform well.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

REFERENCES

- Garmire LX, Subramaniam S. 2012. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA* **18**: 1279–1288.
- Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25. doi: 10.1186/gb-2010-11-3-r25.

FIGURE 2. Reanalysis on the evaluation of normalization methods with empirical statistics. (Left) Figure 3 on F-data in our original publication; (right) results as in Figure 3 after adding an additional inverse step for the normalization factor.