miRNA-Seq normalization comparisons need improvement

XIAOBEI ZHOU,[1,2] ALICIA OSHLACK,[3] and MARK D. ROBINSON[1,2,4]

[1]Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland
[3]Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria 3052, Australia

BACKGROUND

Currently there is no method of best practice for the normalization of microRNA sequencing data (miRNA-Seq). Therefore, we read with interest a recent article in *RNA* by Garmire and Subramaniam that set out to compare various normalization strategies specifically for this application (Garmire and Subramaniam 2012). They compared methods currently in use for normalization of messenger RNA sequencing (mRNA-Seq) data, such as total-depth normalization ("raw") and Trimmed Mean of M-values ("TMM"). Additionally, they compared many methods not used previously with sequencing data, such as global scaling, and borrowed from strategies applied to microarray studies, such as quantile normalization (QN). The article attracted our attention for many reasons, but notably for the claimed poor performance and "abnormal results" of our TMM method (Robinson and Oshlack 2010). After investigating, we discovered that TMM's claimed poor performance was the result of an error that shifted log-ratios in the wrong direction. Furthermore, we felt that various practical issues were not satisfyingly discussed; we comment briefly on these here and provide reproducible reanalyses to support our claims (see Supplemental Material).

REPRODUCIBILITY

The authors were confused about how to introduce the TMM normalization factors (private e-mail to us November 6, 2010; code sent privately to us on August 3, 2012). While we did not answer this question directly in the original exchange, we pointed them to our online example code where the TMM normalization factors are introduced to the statistical test. Importantly, as mentioned in the TMM article (Robinson and Oshlack 2010), the normalization factors modify the *library size*, not the *count data*. Therefore, Garmire and Subramaniam's abnormal TMM results can be attributed to introducing these factors in the wrong direction (see Supplemental Note S1 for the correction). We make our R code publicly available, so others can reproduce our analyses and test new situations; documentation for applying TMM in a standard setting is readily available in the edgeR software package (Robinson et al. 2010). However, it is the user's responsibility to ensure correct usage in a nonstandard setting (e.g., operations on log-ratios instead of differential expression statistics).

We have reproduced some of the metrics presented in the Garmire and Subramaniam paper and conclude that the corrected TMM normalization is an average performer and represents an improvement over total-depth normalization (Supplemental Note S2). However, the integration of TMM normalization factors within an established statistical framework provides a clear path from raw data to interpretable statistical summaries (e.g., *P*-values), whereas other methods (e.g., QN) may not, at least in small samples where parametric models are used. Therefore, we question the validity of some of Garmire and Subramaniam's comparisons and also the overall conclusions of the paper, as discussed below.

MSE AND K-S METRICS ARE NOT APPROPRIATE IN THIS SETTING

The purpose of normalization is to remove technical bias while maintaining true biological signal. Garmire and Subramaniam employed mean-squared error (MSE) and the Kolmorogov-Smirnov (K-S) test metrics, among others, to assess normalization performance. A small MSE or K-S statistic, applied here to single samples from *different* biological conditions, was taken by Garmire and Subramaniam to be evidence of good performance. Unfortunately, this comparison gives no consideration to the presence of truly differentially expressed miRNAs, which directly affect these scores. Low MSE favors normalization that removes all evidence of differential expression, which is an undesirable property when true biological differences exist (e.g., here, evidence from corresponding miRNA qPCR data). Notably, the cited reference that uses MSE as a performance metric does so from known (simulated) fold-changes (Xiong et al. 2008). A more appropriate performance metric would be MSE or scale-free coefficient of variation between biological *replicates* of the same condition, as recently reported for comparing mRNA-Seq normalization strategies (Dillies et al. 2012).

[4]**Corresponding author**
E-mail mark.robinson@imls.uzh.ch
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.037895.112.

The K-S test measures the similarity of two cumulative distributions. We question the motivation for this, at two levels: (i) Samples with different "composition" exhibit different marginal distributions (e.g., comparisons of kidney and liver tissue; Supplemental Fig. S8 in Additional file 1 of Robinson and Oshlack 2010); and (ii) QN would always achieve a zero K-S statistic, were it not for the treatment of ties (Supplemental Note S3). Therefore, QN is always put in a favorable light by this comparison, regardless of any nonlinear effects introduced.

It is worth noting that Garmire and Subramaniam's performance comparisons disregard features that are unobserved in one of the two conditions (i.e., count of zero), since fold-changes cannot be computed. However, miRNAs present in one condition and absent in another may be biologically interesting and should not be ignored, which calls into question how to apply QN in practical situations and whether these performance comparisons are representative of the whole data set. Discarding data for the purposes of performance evaluation may be permissible, but removing such data in downstream analyses is clearly undesirable.

STATISTICAL METHODS FOR COUNT DATA NEED COUNTS

As mentioned, TMM preserves the count data by introducing normalization factors as *offsets* in the statistical model (Robinson and Oshlack 2010). In contrast, Garmire and Subramaniam proceeded to use count-based statistical tests (Fisher exact, Binomial, Poisson, and χ^2) to normalized non-count data. We have two reservations about this approach: (i) The tests employed do not have the capacity to address biological variability, which is essential to generalizable conclusions (Hansen et al. 2011); (ii) transforming count data into nonintegers can distort the mean-variance relationships implied by existing count models (Oshlack and Wakefield 2009). Regardless, clear recommendations of how to apply normalization in a practical setting are needed.

REFERENCE DATA SETS

In order to make decisive claims about method performance, "reference" data sets are critical. Such data sets include an independent truth (e.g., measurements from an independent platform) that can be used to evaluate the performance of an algorithm. Garmire and Subramaniam employed receiver operator characteristic (ROC) curves using miRNA qPCR as the independent truth to define truly differential (and non-differential) miRNAs. Our reanalyses of this data set suggest that ROC results are sensitive to decisions made in determining the "truth" (Supplemental Note S4). Altogether, we conclude that the ROC analysis performed by Garmire and Subramaniam is not conclusive, without a further sensitivity analysis of parameters affecting the selection of true positive and true negatives.

SUMMARY

As developers and users of informatics strategies, we are keenly interested in the relative merits of competing approaches. Crucially, there has been relatively little investigation into normalization strategies for miRNA-Seq data and the timely article from Garmire and Subramaniam promised to shed light on this issue. Unfortunately, errors in the implementation, poor choice of performance metrics (or poor choice of data set), few details about practical implementation (e.g., elimination of features containing zero count), and sensitivity to choices made regarding the reference truth data set have left many open questions about the best analysis methods for miRNA-Seq data. In this paper, we have discussed some of the subtle yet critical parameters that need to be carefully investigated.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

X.Z. is supported by SNSF project grant (143883). A.O. is supported by an NHMRC career development fellowship (ID: 1051481). M.D.R. acknowledges funding from the European Commission through the 7th Framework Collaborative Project RADIANT (Grant Agreement Number: 305626).

REFERENCES

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* doi: 10.1093/bib/bbs046.

Garmire LX, Subramaniam S. 2012. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA* **18:** 1279–1288.

Hansen KD, Wu Z, Irizarry RA, Leek JT. 2011. Sequencing technology does not eliminate biological variability. *Nat Biotechnol* **29:** 572–573.

Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4:** 14.

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.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.

Xiong H, Zhang D, Martyniuk CJ, Trudeau VL, Xia X. 2008. Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data. *BMC Bioinformatics* **9:** 25.