Published in final edited form as:

Int J Comput Biol Drug Des. 2014 ; 7(2-3): 102–112. doi:10.1504/IJCBDD.2014.061642.

A comparison of microRNA sequencing reproducibility and noise reduction using mirVana and TRIzol isolation methods

Yan Guo,

Department of Cancer Biology, Vanderbilt University, Nashville TN 37232, USA

Amma Bosompem,

Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville TN 37232, USA

Xue Zhong,

Department of Cancer Biology, Vanderbilt University, Nashville TN 37232, USA

Travis Clark,

Vanderbilt Technologies for Advanced Genomics, Vanderbilt University, Nashville TN 37232, USA

Yu Shyr, and

Department of Biostatistics, Vanderbilt University, Nashville TN 37232, USA

Annette S. Kim*

Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville TN 37232, USA

Abstract

MicroRNAseq (miRNAseq) is a form of RNAseq technology that has become an increasingly popular alternative to miRNA expression profiling. Unlike messenger RNA (mRNA), miRNA extraction can be difficult, and sequencing such small RNA can also be problematic. We designed a study to test the reproducibility of miRNAseq technology and the performance of the two popular miRNA isolation methods, mirVana and TRIzol, by sequencing replicated samples using microRNA isolated with each kit. Through careful analysis of our data, we found excellent repeatability of miRNAseq technology. The mirVana method performed better than TRIzol in terms of useful reads sequenced, number of miRNA identified, and reproducibility. Finally, we identified a baseline noise level for miRNAseq technology; this baseline noise level can be used as a filter in future miRNAseq studies.

Keywords

next generation sequencing; microRNA; miRNA; mirVana; TRIzol

*Corresponding author: annette.s.kim@vanderbilt.edu.

This paper is a revised and expanded version of a paper entitled A comparison of microRNA sequencing reproducibility and noise reduction using mirVana and TRIzol isolation methods presented at the *International Conference on Intelligent Biology and Medicine (ICIBM 2013)*, Nashville, TN, USA, 11–13 August, 2013.

1 Introduction

MicroRNA (miRNA) are a family of small (21–25 base pairs), non-coding RNAs that serve key roles in transcriptional and post-transcriptional repression of gene expression (Chen and Rajewsky, 2007). These miRNA silence genes by base pairing with complementary sequences primarily within the 3' untranslated regions of mRNA molecules, causing translational repression or target degradation (Bartel, 2009; Kusenda et al., 2006). Since multiple miRNA may regulate the expression of a single mRNA and since each miRNA may regulate multiple mRNA, miRNA act as master regulators of this complex network of translational regulation. According to the latest build of miRBase (v 19) (Kozomara and Griffiths-Jones, 2011) there are 1600 precursors and 2042 mature miRNAs that have been identified in humans. The importance of miRNA is underscored by the impressive list of diseases that have been found to be associated with abnormal miRNA expression- the database miR2Disease has documented 163 diseases with reported miRNA association (Jiang et al., 2009).

Microarrays have been used frequently for conducting high-throughput mRNA and miRNA expression profiling. Recently, the introduction of RNAseq technology has had a revolutionary impact on the field of mRNA expression research. RNAseq refers to the use of next-generation sequencing (NGS) technologies to sequence cDNA in order to get information about the RNA content of a given sample. Traditional RNAseq targets mRNA. Although both microarray and RNAseq are used to measure gene expression, there are fundamental differences between the two methods in terms of technology, data format, and analysis. Microarrays measure gene expression based on fluorescence intensity, resulting from hybridisation of the sample(s) to a spatially fixed set of complementary nucleotide probes. By contrast, RNAseq measures gene expression based on the number of cDNA sequences that map to a given region. Correspondingly, there are differences in their data format and analysis. The most popular normalisation method for microarray analysis is robust multi-array average (RMA) (Irizarry et al., 2003), a form of quantile normalisation, while the most popular normalisation method for RNAseq is reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008).

RNAseq data has been shown to estimate expression level with high reproducibility (Marioni et al., 2008). The majority of these previously published studies showed moderate to good concordance rates of gene expression between microarray and RNAseq results. The majority of existing studies have focused on non-human samples such as *Candida parapsilosis* (Bloom et al., 2009), *Candida albicans* (Marioni et al., 2008), fission yeast *Schizosaccharomyces pombe* (Wilhelm et al., 2010), *Drosophila melanogaster* (Malone and Oliver, 2011), *Saccharomyces cerevisiae* (Nookaew et al., 2012), *Caenorhabditis elegans* (Liu et al., 2007), mouse tissues (Liu et al., 2007; t Hoen et al., 2008), and rat tissues (Su et al., 2011). However, few studies have focused on human gene expression consistency between RNAseq and microarray methods. A few studies (Asmann et al., 2009; Cloonan et al., 2008; Marioni et al., 2008) have performed comparisons using human samples or cell lines, but the sample sizes of those studies were very limited.

The RNAseq technology specifically targeting miRNA is called miRNAseq. A previous study has shown that different miRNA microarray platforms produce reasonable reproducibility (Sato et al., 2009). To date, no study has focused on the reproducibility of miRNAseq technologies. To measure the repeatability of miRNAseq technology we designed a miRNAseq experiment using bone marrow mononuclear cells (BM-MNCs). Through the analysis of our miRNAseq data, we developed a statistical framework to identify the noise signal level for miRNAseq data. This noise signal level can be applied to future miRNAseq studies to reliably distinguish sequencing noise and true miRNA expression signal. Finally, we also evaluated the efficiencies of two popular miRNA isolation kits, TRIzol and mirVana.

2 Materials and methods

Bone marrow mononuclear cells (BM-MNCs) were isolated from a single, fresh, unsorted normal bone marrow aspirate specimen by the Ficoll method using Cellgro Lymphocyte Separation Medium (Corning, Manassas, VA) followed by lysis of residual red blood cells by ACK lysing buffer (KD Medical, Columbia, MD). After performing a cell count, the cells were re-pelleted from Dulbecco's phosphate buffered saline solution and resuspended in either 10% dimethylsulfoxide in fetal bovine serum (10% DMSO/FBS), 20% DMSO/FBS, or Gibco Recovery Freezing medium (Invitrogen, Grand Island, NY) and then frozen gradually (1 °C/min) for 16 hours, prior to storage in liquid nitrogen. Cells were frozen at a density of 5 or 10 million cells per mL. Total RNA (totRNA) was isolated using a mirVana miRNA isolation kit (Life Technologies, Grand Island, NY) or the TRIzol-mediated isolation protocol (Life Technologies, Grand Island, NY).

Library construction was performed on the total RNA from 6 samples obtained from a single bone marrow aspirate specimen (each of three different freezing media isolated by both mirVana and TRIzol methods) using the TruSeq Small RNA sample preparation kit (Illumina, San Diego, CA). The small RNA protocol specifically ligates RNA adapters to mature miRNAs that have a 5'-phosphate and 3'-hydroxyl group resulting from enzymatic cleavage by RNA processing enzymes like Dicer. In the first step, RNA adapters were ligated onto each end of the RNA molecules and a reverse transcription reaction was used to create single stranded cDNA. This cDNA was then PCR amplified with a universal primer and a second primer containing one of 48 uniquely indexed tags to allow multiplexing. The cDNA construct was then size-selected using the 3% gel cassette on the Pippin Prep (Sage Sciences) to reduce the library to the mature microRNAs and other regulatory RNAs in the 20–30 bp size range. The resulting cDNA libraries then underwent quality control by running on the Agilent Bioanalyzer HS DNA assay to confirm the final library size and on the Agilent Mx3005P qPCR machine using the KAPA Illumina library quantification kit to determine concentration. A 2 nM stock was created and samples pooled by molarity for multiplexing. From the pool 10 pM was loaded into each well for the flow cell on the Illumina cBot for cluster generation. The flow cell was then loaded onto the Illumina HiSeq 2500 utilising v3 chemistry and HTA 1.8. The raw sequencing reads in BCL format were processed through CASAVA-1.8.2 for FASTQ conversion and demultiplexing. The RTA chastity filter was used and only the PF (pass filter) reads are retained for further analysis.

The resulting miRNAseq FASTQ data were processed as follows. Due to the small size (22–25 base pairs) of miRNA and longer read length (50 base pairs), parts of the sequenced read did not represent miRNA but rather the adaptor. Those adaptor sequences were trimmed to obtain adaptor sequence-free FASTQ files. A majority of the sequenced reads from a miRNAseq experiment are the result of contamination from ribosomal RNA. We performed alignment against ribosomal RNA to identify and remove all these unwanted sequences. Even after decontamination, some remaining reads may be still sequenced from mRNA. Thus we aligned the rest of the reads against mRNA reference and eliminated likely mRNA sequences to obtain the most likely candidates for miRNA. A final alignment was performed against miRNA (1733 entries) and precursor miRNA (1424 entries) reference sequences downloaded from mirBASE (Kozomara and Griffiths-Jones, 2011). Read count and RPKM were generated for each miRNA and precursor miRNA for each sample.

We computed Spearman's correlation coefficients between repeated samples to evaluate the repeatability of miRNAseq technology. To evaluate the isolation efficiency of mirVana and TRIzol, we compared reads sequenced, miRNA detected, and correlations within and between these two capture kits. The noise baseline for miRNAseq technology was computed by comparing the RPKM level between miRNA detected in all three repeats within each isolation kit to miRNA detected in only 1 or 2 repeats.

3 Results and discussion

Both mirVana and TRIzol methods were employed in triplicate, using three different freezing media, to isolate miRNA from a single bone marrow aspirate specimen, which was then sequenced on an Illumina HiSeq 2500 sequencer. We observed on average 31.5 million reads for Mirvana-isolated samples and 17.7 million reads per TRIzol-isolated samples. After filtering our contamination reads, we identified on average 0.4 million reads for miRNA and 0.5 millions reads for precursor miRNA for mirVana isolated samples and 0.1 million reads and 0.1 millions reads for TRIzol-isolated samples (Figure 1, Table 1). MirVana-isolated samples clearly produced more sequencing reads than TRIzol-isolated samples for both miRNA and precursor miRNA. We then considered how many miRNA were identified by at least one read. Using this criterion, we observed more mature miRNA and precursor miRNA for mirVana-isolated samples than TRIzol-isolated samples (Table 2). For example, on average, 357 mature miRNA and 415 precursor miRNA were detected for mirVana-isolated samples while 299 mature miRNA and 351 precursor miRNA were observed for TRIzol-isolated samples. These results are roughly proportional to the read counts observed.

To test the reproducibility of miRNA and precursor miRNA expression within each miRNA isolation method, we performed Spearman's (Figure 2) and Pearson's (Figure 3) correlations between any two repeated samples. The average Spearman correlation coefficient across all samples was 0.955 for both mature miRNA and precursor miRNA with slightly higher coefficients observed for mirVana-isolated samples (0.977 for mature miRNA and 0.967 for precursor miRNA) compared to TRIzol-isolated samples (0.963 for both mature miRNA and precursor miRNA). Not surprisingly, the average Spearman correlation coefficient calculated between mirVana and TRIzol-isolated samples were lower (0.944 for mature miRNA and

0.949 for precursor miRNA) compared to the correlations computed within each isolation method.

To detect the baseline noise for miRNAseq technology, we computed the overlap of detected mature miRNA and precursor miRNA between repeats (Figure 4). The majority of the detected mature and precursor miRNA were detected in all three repeats (three different freezing conditions) within each isolation method. However, there were still a large portion of mature miRNA and precursor miRNA detected by only one or two of the freezing conditions. To better understand these singletons and doubletons, we checked their RPKM distribution (Figure 5). The expression level of mature miRNA and precursor miRNA detected by all three isolation method replicates within each capture kit is significantly higher than mature miRNA and precursor miRNA detected by one or two replicates. The maximum RPKM value observed in singleton or doubleton miRNAs was not any larger than the minimum RPKM value observed in the tripleton miRNA. Almost all of the singleton or doubleton miRNA had only 1 read aligned to them. This result strongly suggests that miRNA detected through a single mapping read is highly unreliable and should be excluded from further analysis.

4 Conclusion

Compared to regular messenger RNAseq technology, the efficiency of miRNAseq is low: only 1–2% of the total sequenced reads can be aligned to miRNA or precursor miRNA while the efficiency for messenger RNAseq technology can range from 20% to 50%. The low efficiency of miRNAseq is not due to technical limitations of the miRNAseq platforms or insufficient miRNA isolation. The low efficiency observed is due rather to the low quantities of miRNA contained in each sample. Compared to coding regions which have approximately 20,000 genes spanning over 30 million base pairs, there are only approximately 1700 identified miRNAs spanning over 40,000 base pairs. Even after miRNA-specific isolation, the fraction of miRNA is only enriched and is still tiny in the final sequencing library. Consequently only a small fraction of the total reads can be aligned to mature and precursor miRNA. Even though the fraction of the useful reads is small, usually 1–2% efficiency is enough to accurately profile the sample's miRNA expression. Unlike regular RNAseq technology which is rapidly replacing microarray as the platform of choice for expression profiling, due to the small number of miRNA that can be identified, miRNAseq has not replaced miRNA array at the same pace. As we have shown, the sequencing efficiency of miRNA is very poor. Therefore, currently microarray technology still has its value in miRNA-based studies.

Based on the current results, there is significant room to improve the sequencing efficiency by reducing the unwanted content in the final sequencing library. Standard mRNA-seq libraries use a poly-A selection resulting in a very specific target pool for the resulting cDNA library and sequencing. It is very common to have >80% of the reads in a mRNAseq experiment align and be informative for downstream analysis. The miRNAseq samples are uniquely challenging because of the need to generate libraries from very small RNA species in a pool of totRNA and then use a very tight size selection to reduce the complexity of the cDNA library for the small RNAs that are to be sequenced. The RNA adapter ligation steps

are used to avoid sequencing adapter dimers that would be ~120 bp in size from microRNA libraries that are 140–150 bp in final size. Ligation to ssRNA with RNA-ligase is inefficient and does not always target just the microRNA species of interest. In cell line and especially FFPE samples degraded RNAs or other small RNAs can fall into the size range selected. Further investigation on the possibility of rRNA reduction and more precise size selection of the final library may increase the efficiency of library preparation and therefore the data content of the sequencing read pool.

A recent study by Kim et al. has shown that specific loss of miRNA can happen during RNA preparation using TRIzol isolation kit (Kim et al., 2012). Their analyses have shown that structured small RNAs including mature miRNA, precursor miRNA, small interfering RNA (siRNA) duplexes, and transfer RNAs (tRNAs) with low GC content are recovered inefficiently when a small number of cells is used for RNA isolation with TRIzol. This finding has resulted in them retracting their earlier finding based on a TRIzol isolation kit. Our results also indicate that mirVana is a better miRNA isolation kit than TRIzol. MirVana produced more reads, detected more miRNA, and had better reproducibility than TRIzol. We therefore recommend use of the mirVana kit over the TRIzol method for the isolation of miRNA.

Finally, by comparing overlapping miRNA between samples, we concluded that miRNA identified by a single mapping read is highly unreliable. Granted, a small portion of tripleteon miRNA is still mapped by a single read, but there is no way for us to distinguish these 'hits' from noise. To increase the specificity of the study, we recommend excluding such miRNA reads in further analyses.

Biographies

Yan Guo is a Research Assistant Professor of Department of Cancer Biology at Vanderbilt University. He received his PhD in Computer Science from University of South Carolina.

Amma Bosompem is a Research Associate with Dr. Annette S. Kim. She has a Master's of Science degree from Clemson University.

Xue Zhong received his PhD in Human and Molecular Genetics from the University of Texas Health Science Center at Houston and MD Anderson Cancer Center, Houston, Texas in 2000. Currently, he is an Associate Professor in the Departments of Biomedical Informatics, Psychiatry, and Cancer Biology at Vanderbilt University Medical Center.

Travis Clark is the Technical Director of Sequencing at Vanderbilt Technologies for Advanced Genomics.

Yu Shyr is a Professor of Department of Biostatistics and he is the Director for Center for Quantitative Sciences.

Annette S. Kim, MD, PhD, is an Assistant Professor in the Department of Pathology, Microbiology, and Immunology.

References

- Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S, Oberg AL, Therneau TM, Smith DI, Poland GA, Wieben ED, Kocher JP. 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics*. 2009; 10:531. [PubMed: 19917133]
- Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009; 136:215–233. [PubMed: 19167326]
- Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*. 2009; 10:221. [PubMed: 19435513]
- Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*. 2007; 8:93–103. [PubMed: 17230196]
- Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*. 2008; 5:613–619. [PubMed: 18516046]
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4:249–264. [PubMed: 12925520]
- Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*. 2009; 37:D98–104. [PubMed: 18927107]
- Kim YK, Yeo J, Kim B, Ha M, Kim VN. Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. *Mol Cell*. 2012; 46:893–895. [PubMed: 22749402]
- Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*. 2011; 39:D152–157. [PubMed: 21037258]
- Kusenda B, Mraz M, Mayer J, Pospisilova S. MicroRNA biogenesis, functionality and cancer relevance. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*. 2006; 150:205–215. [PubMed: 17426780]
- Liu F, Jenssen TK, Trimarchi J, Punzo C, Cepko CL, Ohno-Machado L, Hovig E, Kuo WP. Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC Genomics*. 2007; 8:153. [PubMed: 17555589]
- Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*. 2011; 9:34. [PubMed: 21627854]
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008; 18:1509–1517. [PubMed: 18550803]
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008; 5:621–628. [PubMed: 18516045]
- Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlen M, Nielsen J. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res*. 2012; 40:10084–10097. [PubMed: 22965124]
- Sato F, Tsuchiya S, Terasawa K, Tsujimoto G. Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. *PLoS One*. 2009; 4:e5540. [PubMed: 19436744]
- Su Z, Li Z, Chen T, Li QZ, Fang H, Ding D, Ge W, Ning B, Hong H, Perkins RG, Tong W, Shi L. Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem Res Toxicol*. 2011; 24:1486–1493. [PubMed: 21834575]
- t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT. Deep sequencing-based expression analysis shows major advances in

robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.* 2008; 36:e141. [PubMed: 18927111]

Wilhelm BT, Marguerat S, Goodhead I, Bahler J. Defining transcribed regions using RNA-seq. *Nat Protoc.* 2010; 5:255–266. [PubMed: 20134426]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Overall read distribution for all six repeats by percentage (Top graph), reads in millions (Bottom table) (see online version for colours)

Figure 2. Reproducibility was evaluated using Spearman's correlation between any two repeats for (a) mature miRNA and (b) precursor miRNA. The overall reproducibility is high for miRNAseq. The mirVana isolation kit showed slightly better reproducibility than TRIzol isolation method (see online version for colours)

Figure 3. Reproducibility was evaluated using Pearson's correlation between any two repeats for (a) mature miRNA and (b) precursor miRNA (see online version for colours)

Figure 4.

The Venn diagram of overlapped, identified mature miRNA (a, b) and precursor miRNA (c, d) among the 3 repeats separate by miRNA isolation method. The mirVana isolation kit (a, c) identified more tripton miRNA reads than the Trizol-based isolation method (b, d) (see online version for colours)

Figure 5.

Median expression values of singleton, doubleton and tripleton (a) mature miRNA and (b) precursor miRNA for both mirVana and TRIzol isolation kits. The mirVana isolation kit had higher expression values than the TRIzol method for most freezing conditions. The doubleton and singleton miRNA have significantly lower expression compared to tripleton miRNA (see online version for colours)

Table 1

Read mapping statistics

	Un-mapped	Contam	Genomic	mir	Precursor	Total count
mirVana_rep1	24562174	855473	17106831	927011	1084514	44536003
mirVana_rep2	12442525	459672	7337388	313057	336871	20889513
mirVana_rep3	18754230	562302	9549453	93099	125728	29084812
TRIZOL_rep1	14397497	810027	8205640	118828	179316	23711308
TRIZOL_rep2	9715436	742174	5933026	83625	130199	16604460
TRIZOL_rep3	4253119	2923577	5454786	81959	123558	12836999

Table 2

miRNA detection

	No. of hits	% of hits
<i>Mature microRNA database (1733)</i>		
Mirvana_repeat1	446	26
Mirvana_repeat2	351	20
Mirvana_repeat3	273	16
TRIZol_repeat1	309	18
TRIZol_repeat2	282	16
TRIZol_repeat3	306	18
<i>Precursor microRNA database (1424)</i>		
Mirvana_repeat1	503	35
Mirvana_repeat2	402	28
Mirvana_repeat3	341	24
TRIZol_repeat1	347	24
TRIZol_repeat2	356	25
TRIZol_repeat3	351	25