Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 862130, 9 pages
http://dx.doi.org/10.1155/2015/862130

Research Article

RNAseq by Total RNA Library Identifies Additional RNAs Compared to Poly(A) RNA Library

Yan Guo,[1] Shilin Zhao,[1] Quanhu Sheng,[1] Mingsheng Guo,[1] Brian Lehmann,[2] Jennifer Pietenpol,[2] David C. Samuels,[3] and Yu Shyr[1]

[1]*Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, USA*
[2]*Department of Biochemistry, Vanderbilt University, Nashville, TN 37232, USA*
[3]*Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA*

Correspondence should be addressed to Yan Guo; yan.guo@vanderbilt.edu and Yu Shyr; yu.shyr@vanderbilt.edu

Received 8 December 2014; Revised 27 January 2015; Accepted 15 February 2015

Academic Editor: Xia Li

The most popular RNA library used for RNA sequencing is the poly(A) captured RNA library. This library captures RNA based on the presence of poly(A) tails at the 3′ end. Another type of RNA library for RNA sequencing is the total RNA library which differs from the poly(A) library by capture method and price. The total RNA library costs more and its capture of RNA is not dependent on the presence of poly(A) tails. In practice, only ribosomal RNAs and small RNAs are washed out in the total RNA library preparation. To evaluate the ability of detecting RNA for both RNA libraries we designed a study using RNA sequencing data of the same two breast cancer cell lines from both RNA libraries. We found that the RNA expression values captured by both RNA libraries were highly correlated. However, the number of RNAs captured was significantly higher for the total RNA library. Furthermore, we identify several subsets of protein coding RNAs that were not captured efficiently by the poly(A) library. One of the most noticeable is the histone-encode genes, which lack the poly(A) tail.

1. Introduction

With the advancement of high throughput sequencing technology, advanced data mining techniques have been developed for high throughput DNA sequencing data [1, 2]. Similar data mining techniques can be applied to RNAseq data. RNAseq technology can be categorized into three subclasses by the types of RNA sequenced: messenger RNA (mRNA or protein coding RNA), micro RNA (miRNA), and total RNA. The sequencing method is the same but each differs in the RNA species present for cDNA synthesis and subsequent library construction. The cDNA library for mRNAseq is made only from the poly(A) mRNA. Small RNAs are not captured during oligo-dT based mRNA enrichment. To date, the most popular application of RNAseq technology is mRNA sequencing because most researchers use RNAseq as a replacement for microarray to perform high throughput gene expression profiling [3–6] and coding regions remain the focus of human disease research.

Long noncoding RNA (lncRNA), on the other hand, was traditionally believed to be nonfunctional. However, many recent studies have shown evidence for the functionality of lncRNA [7, 8], such as roles in high-order chromosomal dynamics [9], embryonic stem cell differentiation [10], telomere biology [11], subcellular structural organization [12], and breast cancer [13, 14]. The interest in lncRNA grew considerably as the evidence of lncRNA's role in various biological contexts accumulated in the recent years. LncRNAs are usually defined as noncoding RNA with length more than 200 base pairs [7, 15]. Structurally, lncRNAs and mRNAs are very similar, as both can exhibit polyadenylation (poly(A)). The number of definable lncRNAs varies by study. An early study in 2007 estimated that there are 4 times more lncRNAs than protein coding RNA [16]. Another study claims to have identified 35,000 lncRNAs [17], and many of them have characteristics similar to mRNA such as 5′ capping, splicing, and polyadenylation, with the exception of open reading frames [17]. In the latest effort to quantify human

lncRNA, the Encyclopedia of DNA Elements (ENCODE) [18] project identified 13,333 lncRNAs and further categorized them into four subclasses: (1) antisense, (2) large intergenic noncoding RNAs (lincRNA), (3) sense intronic, and (4) processed transcripts.

While it is possible to study lncRNAs using traditional microarrays, RNAseq has been proven to be the superior technology for this purpose due to its greater sensitivity and the ability to detect novel lncRNAs [19, 20]. The rise in the popularity and affordability of RNAseq technology is primarily responsible for the growing interest in and understanding of lncRNAs as researchers explore the presence of these stowaways in their mRNA data sets. In mRNA sequencing, mRNAs are captured based on the presence of a poly(A) tail. LncRNAs can also be captured provided they have a poly(A) tail. According to a study in 2005, it is estimated that 40% of lncRNA transcripts are nonpolyadenylated [21]. An alternative library preparation method for studying lncRNA is the total RNA library. Only ribosomal RNA is removed leaving small RNAs, mRNAs, and all forms of lncRNAs. This library preparation method is the most inclusive of RNA species but requires more sequencing reads due to the multiple RNA species present in the library, and ribosomal RNA reduction does not completely remove ribosomal RNA from the library due to their high abundance.

Total RNA sequencing theoretically should detect more lncRNAs due to its RNA selection independent of the poly(A) tail. However, total RNAseq costs more than mRNA sequencing (mRNA $500 versus total RNA $650) and the question of how many more lncRNAs does total RNA sequencing capture compared to mRNA sequencing has not been answered. Moreover, whether the mRNAs captured in total RNA sequencing are comparable to mRNA sequencing also remains unknown. To answer these questions, we designed the following study. We hypothesized that total RNA sequencing generates more relevant data than mRNA sequencing for the purpose of lncRNA research. Total RNA and mRNA libraries of two breast cancer cell line samples were built and sequenced. We analyzed the sequencing data and compared their usability for lncRNA and mRNA research.

2. Methods

Total RNAseq on two breast cancer cell lines HS578T and BT549 was performed by the Vanderbilt Technologies for Advanced Genomics (VANTAGE) core. Total RNA was isolated with the Aurum Total RNA Mini Kit. All samples were quantified on the QuBit RNA assay. RNA quality was checked using Agilent Bioanalyzer. RNA integrity number (RIN) for both samples was 10. RNAseq data was obtained by first using the Ribo-Zero Magnetic Gold Kit (human/mouse/rat) (Epicentre) to perform ribosomal reduction on 1 μg total RNA following the manufacturer's protocol. After ribosomal RNA (rRNA) depletion, samples were then purified using the Agencourt RNAClean XP Kit (Beckman Coulter) according to the Epicentre protocol specifications. After purification, samples were eluted in 11 μL RNase-free water. Next, 1 μL ribosomal depleted samples were run on the Agilent RNA 6000 Pico Chip to confirm rRNA removal. After confirmation of rRNA removal, 8.5 μL rRNA-depleted sample was input into the Illumina TruSeq Stranded RNA Sample Preparation kit (Illumina) for library preparation. The libraries were sequenced on Illumina High HiSeq 2500 with paired-end 100 base pair long reads. Raw RNAseq sequencing data generated from the poly(A) library of the same two cell lines were downloaded from the Gene Expression Omnibus (GEO) (GSM1172877: 19.8 million reads and GSM1172855: 15.3 million reads) for comparative purpose. The poly(A) libraries were prepared using Illumina TruSeq RNA Sample Preparation kit. Poly(A) RNA was purified with oligo dT magnetic beads, and the poly(A) RNA was fragmented with divalent cations followed by reverse transcription into cDNA and ligation of Illumina paired-end oligo adapters to the cDNA fragments. More detail of poly(A) library construction can be found at GEO website.

The raw data quality was examined using QC3 [22]. Alignment against human genome reference HG19 was performed using TopHat2 [23]. Novel gene quantification was performed using Cufflinks [24]. Additional quality control was carried out at alignment level based on the alignment quality control concept described in [25]. ENSEMBL gene transfer format (GTF) version GRCh37.35 was used to annotate the gene expression. We categorized the RNA into three subclasses: protein coding RNA, lncRNA, and other RNAs. This GTF contains 20327 protein coding RNAs, 13346 lncRNAs, and 24100 other RNAs (such as pseudogene and antisense). Read count per RNA was computed using HTSeq [26]. To avoid variation caused by total reads sequenced, raw read counts were normalized to the total read count by sample. Log2 transformations were performed on normalized read counts. To avoid log of zeroes, all read counts were increased by 1 before taking the log transformation. Differential expression analyses and additional quality control were conducted between poly(A) capture method and total RNA method using MultiRankSeq [27] which embeds three different RNAseq differential expression analysis methods: DESeq2 [28], edgeR [29], and baySeq [30]. DEseq2's results were selected for further analysis due to its ability to take paired samples into consideration. Cluster analysis was performed using Heatmap3 [31]. Functional analysis was carried out using gene set enrichment analysis (GSEA) [32], and gene ontology (GO) analysis was conducted using WebGestalt [33].

3. Results

Even though the RNAseq data were generated from the same cell lines, there could be potential heterogeneity and batch effect because the cell lines were cultured at two different labs and sequenced at two different facilities. To test if there is potential heterogeneity and batch effect, we conducted a cluster analysis using Heatmap3 [31]. Unsupervised cluster results showed cluster of cell line type rather than sequencing batch (Figure 1) which suggested that the RNAseq data of

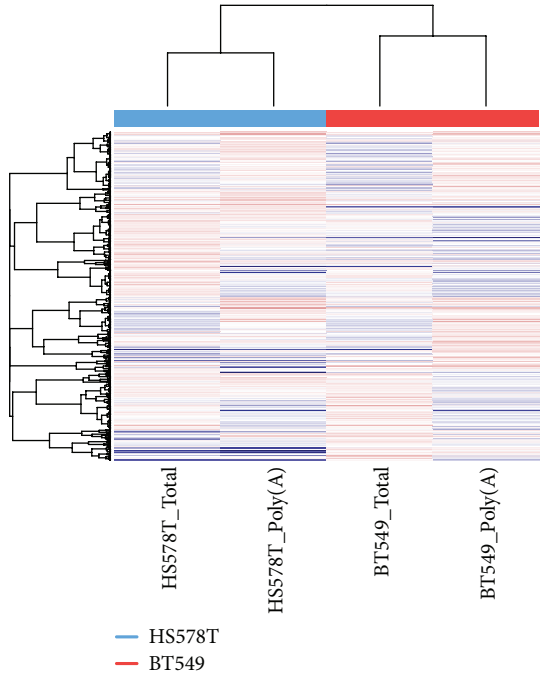

Figure 1: Cluster results of the two breast cancer cell lines. The poly(A) and total RNA libraries were constructed and sequenced by separated facilities. The samples clustered together by cell line type rather than library type or sequencing facility, which suggests that there is no severe heterogeneity of cell line and batch effect between sequencing.

these two cell lines were similar; no severe heterogeneity and batch effect were observed.

The sequencing data went through rigorous quality control. To account for variation in number of reads sequenced within the 4 samples, read counts were adjusted by normalizing the total read count of each sample. In terms of proportion of reads mapped to lncRNA, total RNA library samples (3.62% and 3.23%) had a higher proportion than poly(A) library samples (0.85% and 1.02%). For protein coding RNA, poly(A) library samples (96.34% and 95.38%) mapped a higher proportion of reads than total RNA samples (92.47% and 93.45%). For other species of RNAs, poly(A) library samples (2.81% and 3.59%) and total RNA library samples (3.91% and 3.32%) had similar proportion of reads aligned.

The distributions of read normalized read counts for protein coding RNA, lncRNA, and other RNAs can be seen in Figure S1 (in Supplementary Material available online at http://dx.doi.org/10.1155/2015/862130). All three types of RNAs were detected by both poly(A) and total RNA library building methods. To compare whether RNA expressions are comparable between the two RNA library building methods, we drew a scatter plot and computed their Pearson's correlation coefficients (Figure 2). All three types of RNA expression are highly agreeable between the two methods (protein coding RNA $r = 0.92$, lncRNA Pearson $r = 0.79$, and other RNAs $r = 0.69$). These results are consistent with previous findings [34] which suggest that RNA expression is consistently measured for poly(A) and total RNA sequencing library.

Next, we examine the number of RNAs detectable by each library construction method. To determine whether RNA is detected, a cutoff value of the normalized read count was applied. Because this cutoff is arbitrary, we choose several different thresholds for sensitivity analysis. An RNA is considered detected if its normalized read count is above the detection threshold. We used the following thresholds: >0.1, >0.5, >1, >1.5, and >2. Regardless of which threshold we applied, samples from the total RNA method consistently showed higher numbers of RNAs detected for all three types of RNAs (Figure 3). This suggests that without the restriction of poly(A) selection, the total RNA library is capable of identifying more expressed RNAs (lncRNA t-test $P < 0.0001$, protein coding RNA t-test $P < 0.0001$, and other RNAs t-test $P < 0.0001$). Furthermore, we compared the number of genes that are differentially expressed between the two libraries' construction methods and found there were much higher expressed RNAs (log2 fold change > 2) for total RNA library samples than poly(A) library samples (Figure 4). We also counted the potential novel transcripts identified from Cufflinks. The two poly(A) library samples detected 4122 and 6169 potential new transcripts, and the two total RNA samples detected 53282 and 58111 potential new transcripts, roughly a 10-fold increase.

It has been shown that not all mRNAs necessarily contain a poly(A) tail at their $3'$ ends [35]. For example, the mRNA that encodes histone proteins is nonpolyadenylated [36]. Another study has shown that a significant portion of the mRNA transcript has no poly(A) tail [37]. This can

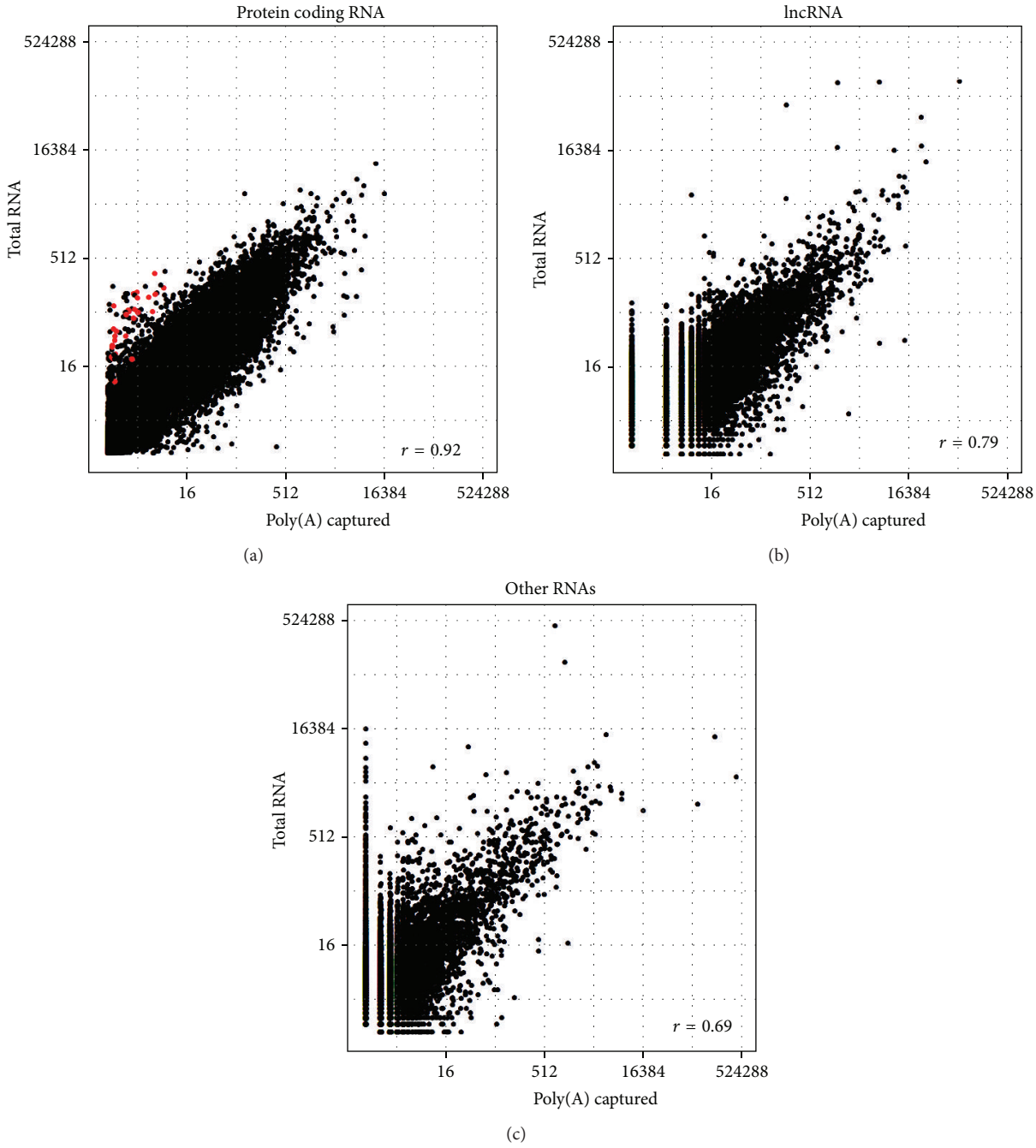

Figure 2: RNA expression level consistency between poly(A) and total RNA library samples. Read counts were normalized by total read count per sample and log2 transformed. (a) Consistency of expression of protein coding RNAs. The red color indicates histone-encoding genes. (b) Consistency of expression of lncRNAs. (c) Consistency of expression of other RNAs.

potentially explain why we observe more protein coding RNA detected by total RNA than the poly(A) method. To test this hypothesis, we searched through the ENSEMBL database and found 38 histone-encoding genes. We conducted enrichment analysis in GSEA using results from DESeq2 against the histone-encoding genes and found that our dataset was highly enriched (FDR < 0.0001) (Figure 5(a)). The expression value of the histone-encoding genes was clearly higher for total RNA library samples (Figure 5(b)). The GSEA showed that total RNA library samples captured histone-encoding genes at a much higher efficiency than the poly(A) library samples. Based on fold change results from DESeq2, there were 737 protein coding RNAs that have a log2 fold change greater than 2 (overexpressed in total RNA samples), which suggests that additional subsets of protein coding RNAs may be better captured using total RNA methods. To better categorize these potential subcategories of protein coding RNAs, we conducted GO analysis using WebGestalt (Figure S2) (Table 1).

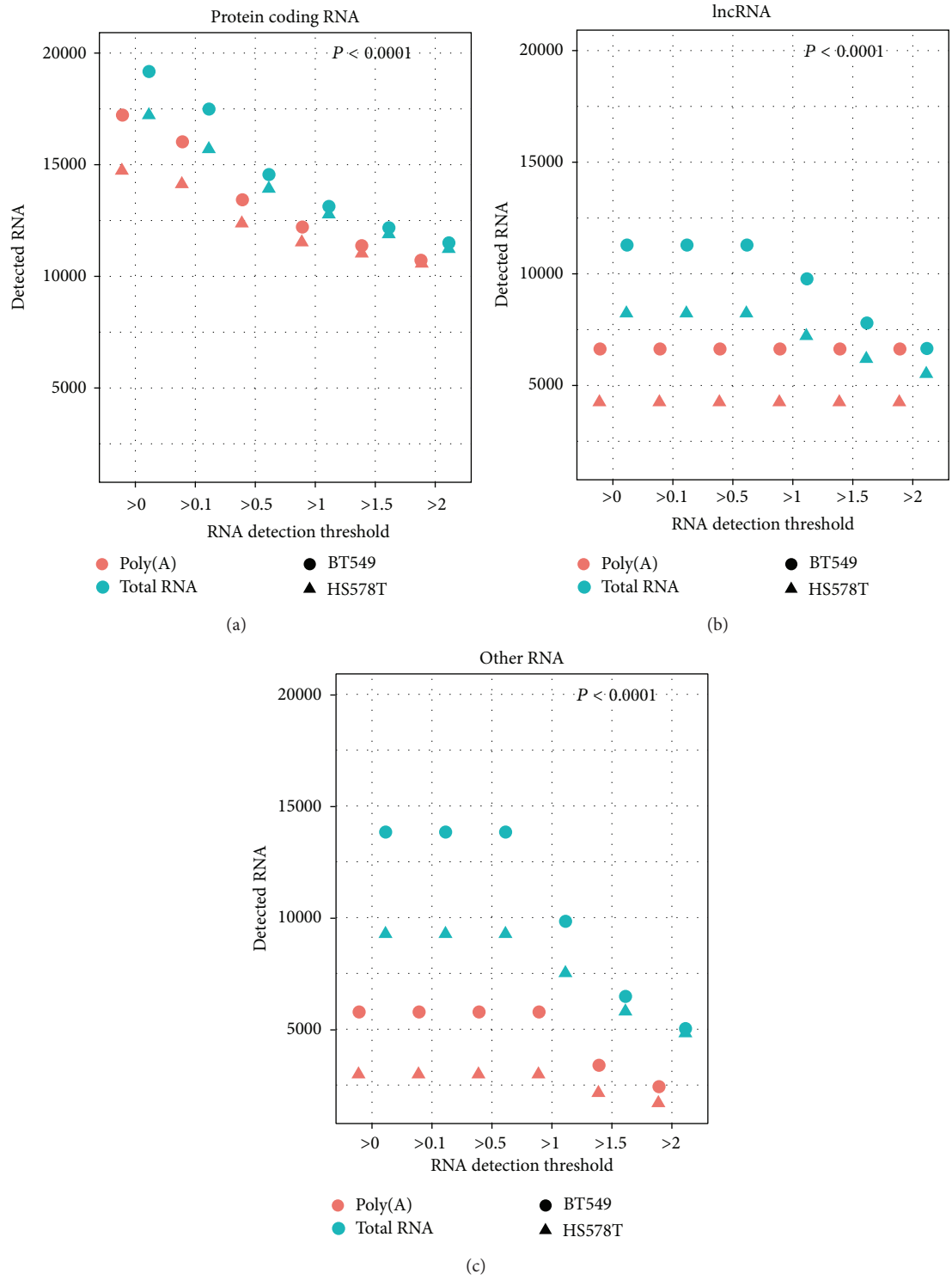

Figure 3: Number of RNAs detected at different detection thresholds for all three types of RNA. Total RNA library samples detected significantly more RNAs than poly(A) RNA library samples at all RNA detection thresholds. (a) Protein coding RNA. (b) lncRNA. (c) Other RNAs.

The top 10 subcategories of genes were found within all three big GO categories: biological process, molecular function, and cellular component. Eleven out of the 30 subcategories primarily consisted of histone-encoding genes. The other 19 subcategories were protein-DNA complex, chromatin, and so forth. No obvious pattern was recognizable. There were also 592 protein coding genes that were captured better by the poly(A) library samples (log2 fold change < -2). We also

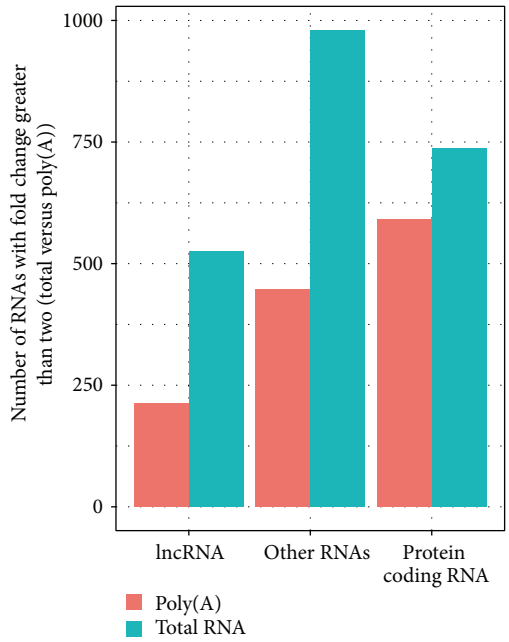

Figure 4: Using log2 fold change >2 as cutoffs, total RNA library samples had more RNAs with higher expression levels than poly(A) samples for all three types of RNAs.

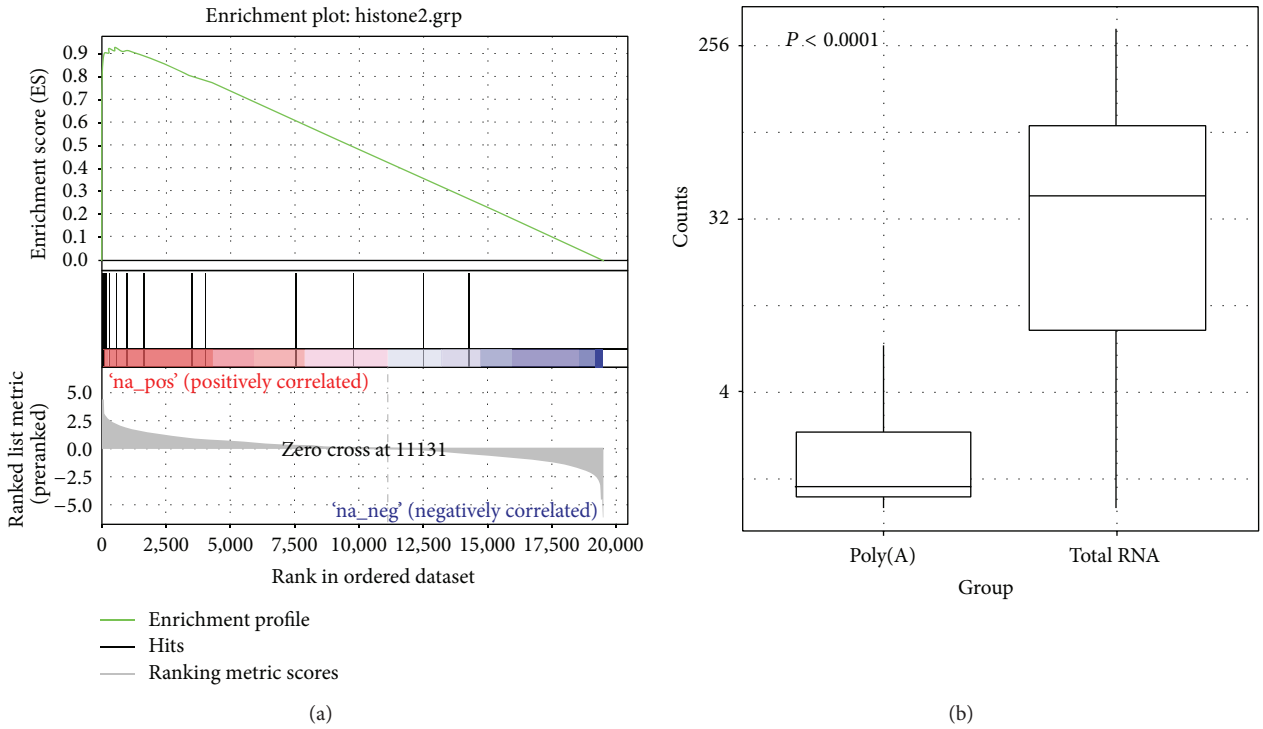

Figure 5: (a) Enrichment plot of histone-encoding genes from GSEA. Based on fold change ranked (total RNA versus poly(A)) gene list, histone-encoding genes were highly enriched (adjust $P < 0.0001$). (b) Normalized read count distribution of the 38 histone-encoding genes between poly(A) and total RNA libraries.

Table 1: Gene ontology results of genes that are captured more by total RNA library.

Major category	Subcategory	Number of genes	Adjusted P
Biological process	Nucleosome assembly (histone)	30	$4.81E-16$
	Protein-DNA complex assembly (histone)	33	$6.27E-16$
	Chromatin assembly (histone)	30	$2.58E-15$
	Protein-DNA complex subunit organization (histone)	33	$9.36E-15$
	Nucleosome organization (histone)	30	$2.36E-14$
	Chromatin assembly or disassembly (histone)	30	$1.81E-13$
	DNA packaging (histone)	30	$2.51E-12$
	DNA conformation change (histone)	30	$7.46E-10$
	Cellular macromolecular complex assembly (histone)	38	$6.50E-03$
	Detection of virus	3	$7.30E-03$
Molecular function	Protein heterodimerization activity (histone)	32	$5.00E-04$
	Ketosteroid monooxygenase activity	3	$4.00E-03$
	Phenanthrene 9,10-monooxygenase activity	3	$4.00E-03$
	cGMP binding	5	$4.00E-03$
	Oxidoreductase activity	7	$6.30E-03$
	Androsterone dehydrogenase activity	3	$6.30E-03$
	Dehydrogenase activity	3	$6.30E-03$
	Cyclic nucleotide binding	6	$1.48E-02$
	N,N-Dimethylaniline monooxygenase activity	3	$1.48E-02$
	Metal ion transmembrane transporter activity	25	$2.83E-02$
Cellular component	Nucleosome (histone)	29	$8.22E-23$
	Protein-DNA complex	30	$4.93E-17$
	Chromatin	35	$1.22E-08$
	Chromosomal part	40	$1.00E-04$
	Chromosome	41	$2.60E-03$
	Extracellular region part	59	$1.33E-02$
	Axoneme	9	$2.28E-02$
	Platelet dense tubular network membrane	3	$6.32E-02$
	Platelet dense tubular network	3	$1.15E-01$
	Desmosome	4	$1.57E-01$

performed GO analysis on these genes (Figure S3) (Table 2). No clear gene pattern was detected.

4. Discussion

In this study, we examined the difference between the RNAs captured through poly(A) and total RNA libraries. Our study was also designed with several limitations. First, we were only able to collect two samples with sequencing data from both RNA libraries. The small sample size might limit our ability to identify true signals. Also, the sample type is limited to breast cancer cell lines. Other tissue types might behave differently.

Using sequencing data from two breast cancer cell lines captured using both libraries, we found that, in terms of expression level, both libraries were highly correlated and the correlation was the highest for protein coding RNAs. This suggests that both methods of RNA library construction are capable of generating consistent data for studying protein coding RNAs. For the three types of RNA we defined: protein coding RNA, lncRNA, and other RNAs; at all gene detection thresholds, total RNA library samples consistently identified more RNAs than poly(A) library samples which suggests that the total RNA library is capable of detecting additional RNA not detected by the poly(A) library. Through gene set enrichment analysis we were able to identify that histone-encoding genes were not captured efficiently by the poly(A) RNA library due to their lack of poly(A) tails. This finding is consistent with previous reports [36, 37]. Through gene ontology analysis we identified several additional subgroups of RNA which were better captured by the total RNA library. This could be explained in several ways. First, the results could be due to random variation, thus not holding any biological significance. Second, the poly(A) tails might have degraded prior to the construction of the poly(A) RNA library. Third, some unknown mechanisms may prevent proper capture of such RNAs through poly(A) identification.

Total RNA library construction costs around $150 more than a poly(A) library, but it allows the detection of additional RNAs. Whether the extra cost is justifiable should be decided during the experimental design stage of RNAseq study. If the goal is to study lncRNA, then it is better to use total RNA library; if the goal is to study protein coding RNAs, then total

Table 2: Gene ontology results of genes that are captured more by poly(A) RNA library.

Major category	Subcategory	Number of genes	Adjusted P
Biological process	RNA metabolic process	174	$1.30E-03$
	Nucleic acid metabolic process	188	$4.50E-03$
	Cellular macromolecule metabolic process	260	$5.30E-03$
	Positive regulation of cell development	17	$5.50E-03$
	Transcription from RNA polymerase II promoter	75	$5.50E-03$
	Cellular component organization	174	$5.80E-03$
	Cellular component organization or biogenesis	177	$6.90E-03$
	Positive regulation of cell morphogenesis	7	$6.90E-03$
	Negative regulation of viral entry into host cell	3	$7.00E-03$
	Regulation of transcription, DNA-dependent	126	$7.00E-03$
Molecular function	Chromatin binding	27	$1.00E-03$
	Protein binding	276	$1.60E-03$
	Binding	400	$3.90E-02$
	D-Erythro-sphingosine kinase activity	2	$5.57E-02$
	Transcription cofactor activity	28	$5.57E-02$
	Lipid kinase activity	3	$5.57E-02$
	Sphinganine kinase activity	2	$5.57E-02$
	Transcription factor binding transcription factor activity	28	$6.34E-02$
	Nucleic acid binding	127	$9.22E-02$
	Protein binding transcription factor activity	28	$9.22E-02$
Cellular component	Nucleus	251	$4.99E-09$
	Membrane-bounded organelle	349	$1.44E-06$
	Intracellular membrane-bounded organelle	349	$1.44E-06$
	Intracellular organelle	375	$5.83E-06$
	Organelle	375	$5.83E-06$
	Nuclear lumen	128	$6.62E-06$
	Nuclear part	139	$8.80E-06$
	Intracellular organelle lumen	144	$3.83E-05$
	Organelle lumen	145	$4.64E-05$
	Nucleoplasm	77	$5.05E-05$

RNA library might not be necessary unless histone-encoding genes are of interest.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Yan Guo and Yu Shyr were supported by P30 CA68485. Jennifer Pietenpol was supported by National Institute of Health, Grants CA95131 and RC2CA148375. Brian Lehmann was supported by Komen for the Cure Foundation, Grant KG262005. The sequencing of the two total RNA library samples was supported by Vanderbilt Institute for Clinical and Translational Research, Grant VR8688. The authors would also like to thank Margot Bjoring for editorial support.

References

[1] D. C. Samuels, L. Han, J. Li et al., "Finding the lost treasures in exome sequencing data," *Trends in Genetics*, vol. 29, no. 10, pp. 593–599, 2013.

[2] F. Ye, D. C. Samuels, T. Clark, and Y. Guo, "High-throughput sequencing in mitochondrial DNA research," *Mitochondrion*, vol. 17, pp. 157–163, 2014.

[3] Y. Guo, C.-I. Li, F. Ye, and Y. Shyr, "Evaluation of read count based RNAseq analysis methods," *BMC Genomics*, vol. 14, supplement 8, article S2, 2013.

[4] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Article ID e71462, 2013.

[5] J. Shendure, "The beginning of the end for microarrays?" *Nature Methods*, vol. 5, no. 7, pp. 585–587, 2008.

[6] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[7] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[8] M. E. Dinger, P. P. Amaral, T. R. Mercer, and J. S. Mattick, "Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications," *Briefings in Functional Genomics and Proteomics*, vol. 8, no. 6, pp. 407–423, 2009.

[9] P. P. Amaral and J. S. Mattick, "Noncoding RNA in development," *Mammalian Genome*, vol. 19, no. 7-8, pp. 454–492, 2008.

[10] M. E. Dinger, P. P. Amara, T. R. Mercer et al., "Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation," *Genome Research*, vol. 18, no. 9, pp. 1433–1445, 2008.

[11] S. Schoeftner and M. A. Blasco, "Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II," *Nature Cell Biology*, vol. 10, no. 2, pp. 228–236, 2008.

[12] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick, "Specific expression of long noncoding RNAs in the mouse brain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 2, pp. 716–721, 2008.

[13] A. Bhan, I. Hussain, K. I. Ansari, S. A. M. Bobzean, L. I. Perrotti, and S. S. Mandal, "Bisphenol-A and diethylstilbestrol exposure induces the expression of breast cancer associated long noncoding RNA HOTAIR *in vitro* and *in vivo*," *The Journal of Steroid Biochemistry and Molecular Biology*, vol. 141, pp. 160–170, 2014.

[14] A. Bhan, I. Hussain, K. I. Ansari, S. Kasiri, A. Bashyal, and S. S. Mandal, "Antisense transcript long noncoding RNA (lncRNA) HOTAIR is transcriptionally induced by estradiol," *Journal of Molecular Biology*, vol. 425, no. 19, pp. 3707–3722, 2013.

[15] J. M. Perkel, "Visiting 'noncodarnia'," *BioTechniques*, vol. 54, no. 6, pp. 301–304, 2013.

[16] P. Kapranov, J. Cheng, S. Dike et al., "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Science*, vol. 316, no. 5830, pp. 1484–1488, 2007.

[17] P. Carninci, T. Kasukawa, S. Katayama et al., "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, 2005.

[18] S. Djebali, C. A. Davis, A. Merkel et al., "Landscape of transcription in human cells," *Nature*, vol. 489, no. 7414, pp. 101–108, 2012.

[19] L. Han, K. C. Vickers, D. C. Samuels, and Y. Guo, "Alternative applications for distinct RNA sequencing strategies," *Briefings in Bioinformatics*, 2014.

[20] K. C. Vickers, L. A. Roteta, H. Hucheson-Dilks, L. Han, and Y. Guo, "Mining diverse small RNA species in the deep transcriptome," *Trends in Biochemical Sciences*, vol. 40, no. 1, pp. 4–7, 2015.

[21] J. Cheng, P. Kapranov, J. Drenkow et al., "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution," *Science*, vol. 308, no. 5725, pp. 1149–1154, 2005.

[22] Y. Guo, S. Zhao, Q. Sheng et al., "Multi-perspective quality control of Illumina exome sequencing data using QC3," *Genomics*, vol. 103, no. 5-6, pp. 323–328, 2014.

[23] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, article R36, 2013.

[24] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[25] Y. Guo, F. Ye, Q. Sheng, T. Clark, and D. C. Samuels, "Three-stage quality control strategies for DNA re-sequencing data," *Briefings in Bioinformatics*, vol. 15, no. 6, pp. 879–889, 2014.

[26] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[27] Y. Guo, S. Zhao, F. Ye, Q. Sheng, and Y. Shyr, "MultiRankSeq: multiperspective approach for RNAseq differential expression analysis and quality control," *BioMed Research International*, vol. 2014, Article ID 248090, 8 pages, 2014.

[28] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, article 550, 2014.

[29] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[30] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, article 422, 2010.

[31] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced heat map and clustering analysis using heatmap3," *BioMed Research International*, vol. 2014, Article ID 986048, 6 pages, 2014.

[32] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[33] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," *Nucleic Acids Research*, vol. 41, pp. W77–W83, 2013.

[34] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou, "Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling," *Acta Veterinaria Scandinavica*, vol. 15, article 419, 2014.

[35] W. F. Marzluff, E. J. Wagner, and R. J. Duronio, "Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail," *Nature Reviews Genetics*, vol. 9, no. 11, pp. 843–854, 2008.

[36] W. F. Marzluff, "Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts," *Current Opinion in Cell Biology*, vol. 17, no. 3, pp. 274–280, 2005.

[37] L. Yang, M. O. Duff, B. R. Graveley, G. G. Carmichael, and L.-L. Chen, "Genomewide characterization of non-polyadenylated RNAs," *Genome Biology*, vol. 12, no. 2, article R16, 2011.