Effects of mRNA Library Preparation Methods on Next Generation Sequencing (NGS) Data Analysis

Chao-Kai Chou[1], Wells W. Wu[1], Je-Nie Phue[1], Changyi Lin[1], Rong-Fong Shen[1]

[1]Facility for Biotechnology Resources, Office of the Director, Center for the Biologics Evaluation and Research
U.S. Food and Drug Administration

Abstract

Next generation sequencing (NGS) has revolutionized the detection and quantification of messenger RNA (mRNA) through the unbiased measurement of mRNA in a single assay. It significantly impacts biomedical research and promotes applications of personalized medicine. One of the key processes in NGS mRNA sequencing (mRNA-Seq) is the enrichment of mRNA for the subsequent library construction. While various library preparation methods have been developed, different mRNA enrichment methods appear to yield non-identical sequencing reads (DNA fragments for sequencing) which led to inconsistent mRNA-Seq results. We compared two major mRNA isolation methods, poly-A selection and exome capture, to better understand their impacts on sequencing outcomes, with the hope to identify potential factors that might contribute the observed discrepancies. An identical amount of human tissue total RNA (from brain, liver, and testis) was used to prepare mRNA-Seq libraries, which were then subjected to NGS analysis. The results showed about 5% of protein-encoding transcripts were affected by different library preparation methods. The reads variation did not seem to arise randomly, notably that most genes encoding those transcripts share some common features. A relatively large gene size and absence in the poly-A tail were the two more prominent features that we observed. In addition, the analysis of alternatively-spliced transcripts indicated that quantification of transcript isoforms was greatly influenced by the library preparation method used. Sequencing reads distribution was highly correlated to the differences in isoform expression. In conclusion, a fraction of the profiled transcripts have the tendency to be affected by the library construction method employed. We caution that the interpretation of differences in observed gene expression profiles should take into consideration of possible artifacts caused by library construction methods.

Introduction

mRNA-Seq has been widely used for both basic and clinical research. It largely replaces microarray in gene expression analysis and is utilized in studies for identifying new drug targets, searching biomarkers, and developing precision medicine. To date, several library preparation methods have been developed specifically for mRNA-Seq. The ribosomal RNA (rRNA) depletion method is a general direct approach for removing rRNA by capture probes and enriching other types of RNA, including mRNA, before sequencing. Although the advantage of this method is to analyze coding and non-coding RNA simultaneously, it is not the best approach if the project is focused on mRNA.

Currently, the most common method is to capture mRNA through its poly-A tail region. It can effectively remove rRNA and other poly-A absence RNA to enrich mRNA, which is a straightforward and convenient method to generate mRNA-seq library. However, poly-A selection highly depends on the quality of total RNA. Low quality and degraded RNA fragments are not suitable for poly-A selection due to lacking poly-A region in the mRNA fragments for enrichment. To circumvent this issue, an exome-capture approach, which utilizes known exon probes to enrich the mRNA, was developed to perform mRNA-Seq for highly degraded RNA. The unique advantage of exome-capture is to bypass the RNA quality requirement for generating the mRNA-Seq library. However, the method is limited for sequencing probes captured mRNA only.

In this report, we compared the mRNA sequencing results generated by Illumina TruSeq Stranded mRNA (poly-A selection library; PAL) and Illumina TruSeq Exome (exome capture library; EL) kits using intact tissue RNA. Sequencing libraries generated from brain, liver, and testicular tissue RNA by the two different methods revealed inconsistent expression for about 5% of coding genes. The sequencing reads amount and/or distribution of those affected genes were obviously different in PAL and EL. Gene length, exon number, and poly-A are critical factors that contributed to the discrepancy. Our study demonstrated that library preparation methods significantly affect certain genes' mRNA-Seq results and are likely to influence data interpretation.

Study Design

We compared mRNA-Seq performance and reproducibility with libraries prepared as PAL and EL, sought to understand whether two methods yield comparable results. Using identical total RNA samples from human tissues, we performed quality metrics and gene quantification analysis with prepared sequencing libraries. To understand whether the differences are common in various types of tissue, we utilized total RNA from brain, liver, and testis for analysis. For PAL, mRNA was selected by poly-dT conjugated beads for binding to mRNA poly-A tail. Once mRNA was captured, the cDNA was then generated and amplified for sequencing. Unlike poly-A selection, exome capture was carried out at the cDNA level before conducting amplification. To ensure each library has the same sequencing depth, we randomly sampled 60M reads from each library and processed the data analysis by the CLC Genomics Workbench®. Obtained sequencing reads were mapped to the human reference genome GRCh38 for comparison.

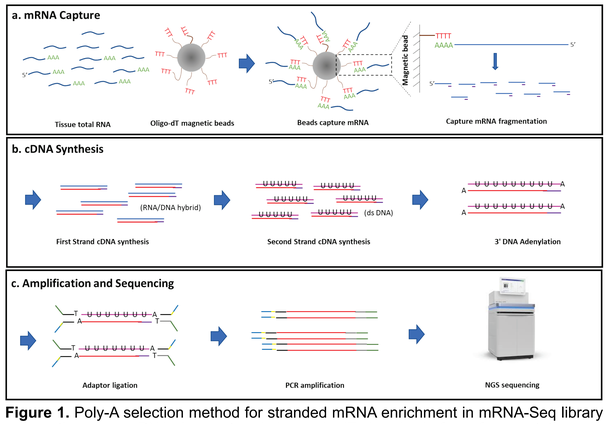

Figure 1. Poly-A selection method for stranded mRNA enrichment in mRNA-Seq library preparation. **a.** mRNA is captured by poly-dT beads before fragmentation. **b.** Fragmented mRNA undergoes cDNA synthesis. **c.** cDNA proceeds for amplification and sequencing.

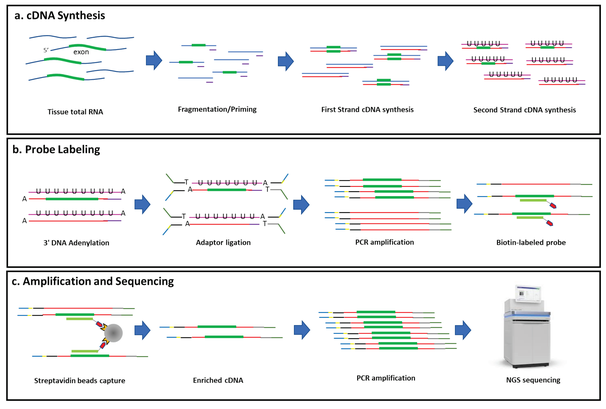

Figure 2. Exon selection method for stranded mRNA enrichment in mRNA-Seq library preparation. **a.** cDNA are synthesized from RNA. **b.** Exome probes label cDNA. **c.** Gene cDNA are enriched by beads, follow by the amplification and sequencing process.

Results

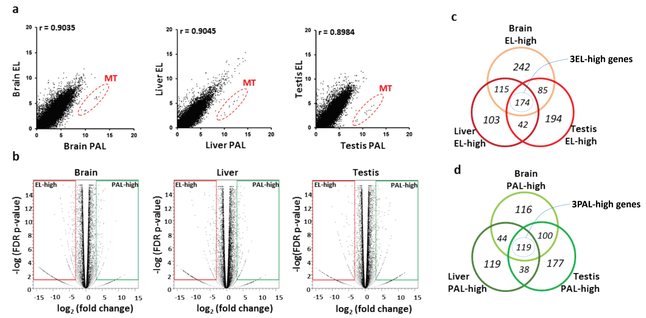

Figure 3. Differential gene expression analysis of poly-A selection (PAL) and exon capture (EL) mRNA-Seq data. **a.** Concordance of gene expression between PAL and EL in different tissues' mRNA-Seq. "r" value is correlation coefficient value. The numbers in the x- and y- axis represent the $\log_2$(CPM+1). CPM, counts per million. Genes in the red circle are affected genes in mitochondrial (MT). **b.** PAL vs.. EL volcano plots of differentially expressed genes in brain, liver, and testis. Genes in the red box are affected genes with EL high expression (EL-high genes), which are $\log_2$ (fold change) < -3, FDR p-value < 0.05; Genes in the green box are PAL-high genes, $\log_2$ (fold change) > 3, FDR p-value < 0.05. **c.** and **d.** Venn diagram of EL-high and PAL-high genes in three different tissues and the number of overlapping genes. 174 3EL-high and 119 3PAL-high genes were identified.

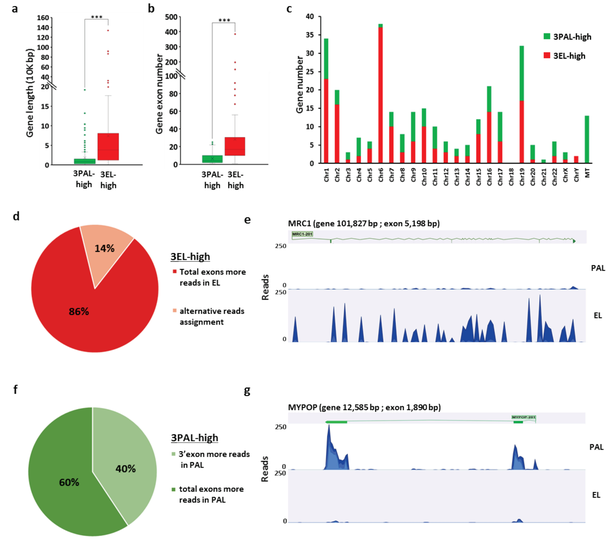

Figure 4. Analysis of factors involved in libraries discrepancy. PAL- and EL-high genes that found in all three tissue types (3PAL-high and 3EL-high genes) were analyzed. **a.** A comparison of gene length between 3PAL- and 3EL-high genes. **b.** Average exon numbers in 3PAL- and 3EL-high genes. **c.** Distribution of 3PAL- and 3EL-high genes in chromosomes. **d.** Pie chart of the factors that involved in 3EL-high genes. **e.** Sequencing reads distribution of *MRC1* in PAL and EL, an example of 3EL-high genes. **f.** Pie chart of the factors that involved in 3PAL-high genes. **g.** Reads distribution of a 3PAL-high gene, *MYPOP*, in PAL and EL. ***T-test p value <0.001.

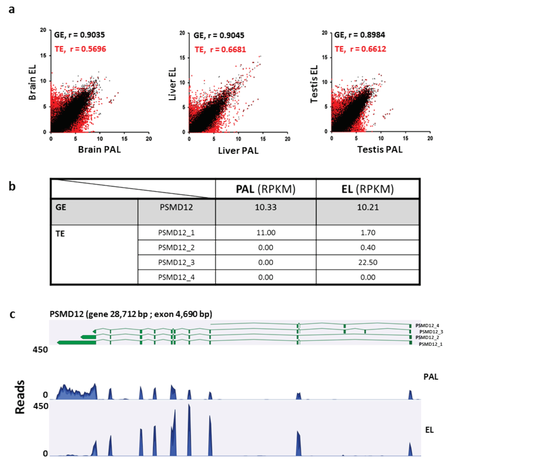

Figure 5. Differential mRNA isoforms expression of PAL and EL mRNA-Seq data. **a.** Correlation of mRNA isoforms transcriptome expression (TE) in different tissues' PAL and EL mRNA-Seq libraries. Red dots, the correlation of isoforms. Black dots, the correlation of genes (GE). r, correlation coefficient. The numbers in the x- and y- axes represent the $\log_2$(CPM+1). **b.** Quantification of *PSMD12,* an example of showing discrepancy in isoforms (TE) but not in (GE) level. **c.** Sequencing reads distribution of *PSMD12* in PAL and EL.

Conclusion

After comparing mRNA-Seq libraries that prepared by either poly-A selection or exome-capture approach, we found 95% of coding genes' mRNA-Seq results were highly correlated. However, the other 5% of genes were showing certain degrees of discrepancies, which might potentially affect the data interpretation. The likely nonrandom discrepancies are attributed to several factors, such as missing poly-A tails or variations in gene lengths. In addition, the measurement of mRNA isoforms were heavily affected by the discrepancy of reads distribution. Therefore, when comparing mRNA-Seq data, the interpretation of differences in gene expression profiles should take into consideration of the potential artifacts caused by different library construction methods.

Note

Our report is an informal communications and represent our best judgement. These comments do not bind or obligate FDA.

For questions and comments, please contact
Chao-Kai Chou (chao-kai.chou@fda.hhs.gov)
Wells Wu (Wells.Wu@fda.hhs.gov)