RNA Sequencing Data Processing

Bioinformatics data processing and analysis was performed by Biostatistics & Bioinformatics Shared Resource, Sylvester Comprehensive Cancer Center, University of Miami.

1) Differential expression analysis

Raw sequence paired-ended reads data in FASTQ format were assessed for quality with FastQC (v.11.5). Trimmomatic (v.0.32) was then used to remove adapters, Illumina-platform specific sequences, and low quality leading and trailing bases from reads. Then STAR (v.2.5.0) was utilized to mapping reads to the reference transcriptome (UCSC hg38 knownGene database). After that, the mapping data were processed by samtools (v.0.1.19) for assignment to genomic features with FeatureCounts function in the Subread package (v.1.5.0)

The transcript quantification (Fragments Per Kilobase of transcript per Million mapped reads) was performed with RSEM (v.1.2.31) with a reference transcriptome (gencode.v26.p10.h38)

2) Motif Analysis

The up-regulated-, down-regulated-, and no-significant-change protein sets were identified by LC-MS/MS. The 5’and 3’UTRs features of the proteins among such sets were fetched from reference transcriptome (ensembl\_GRCh38\_gcv26). The consensus motifs present among the UTR groups associated with each group of proteins were discovered by meme program within the Multiple Em for Motif Elicitation suite (MEME v.4.11.4) as a local (Linux-Ubuntu 16.10) application with minimal length of 8 nucleotide sequence and any number of repeats. Then the percentage of the UTR carrying the specific motif among all the UTR with length > 20 were calculated and demonstrated.

3)The correlation analysis of UTR length and gene expression

The Spearman correlation between the UTR length corresponding to the proteins in the up-regulated and down-regulated sets and the protein expression fold changes were performed in stats package in R (v.3.3.2)