Analyzing Single Cell Small RNA

Yasodha Kannan Sivasamy, Srikanthi Ramachandrula, Hemant Kumar Adil

Strand Life Sciences Pvt. Ltd.

Overview

Strand NGS empowers research biologists by providing a comprehensive and flexible workflow for the analysis of small RNA-Seq data comprising of Alignment, Novel Small RNA detection, Class prediction and a great number of visualization options. To add more power to the tool now, with the introduction of UMI handling and more dynamic plots, single cell sequencing data can be effortlessly analyzed. This support for single cell data analysis is not only limited to RNA-Seq data but extends to small RNA-Seq data as well. This Application Note highlights the robustness of Strand NGS in analyzing one such dataset involving small RNA-Seq data from Human Embryonic Single cells.

Introduction

Small-Seq[1], a method for analyzing small RNA from single cell was employed for studying hESC (naive and primed human Embryonic Stem Cells) and HEK (Human Embryonic Kidney) cells. We reanalyzed these samples using small RNA-Seq analysis workflow in Strand NGS and identified a large number of novel small RNAs in these cells. There remains so much to discover about these novel small RNAs, especially those that are differentially expressed in the hESC cells. We further built a prediction model based on the small RNA expression values and were able to predict accurately the cell type, indicating that these small RNAs can act as biomarkers for hESC and HEK cells.

Sample Dataset

The Single Cell Small RNA dataset was downloaded from NCBI (GEO: GSE81287[1]). A total of 539 cells including Primed and Naive hESCs, HEK293FT, Glioblastoma cells (JM3, JM4, KS4, U87 cell lines) were sequenced by various protocols and analyzed for small RNA expression. These were Directional single ended samples with Unique Molecular Identifiers (UMIs) of length 8bp and stem length of 2bp.

Data analysis methodology

All analyses reported here were performed using Strand NGS bioinformatics software version 3.2. FASTQ files from GSE81287 dataset were imported into Strand NGS and Pre-alignment Quality Control for these samples reported that the average read quality is ~40.

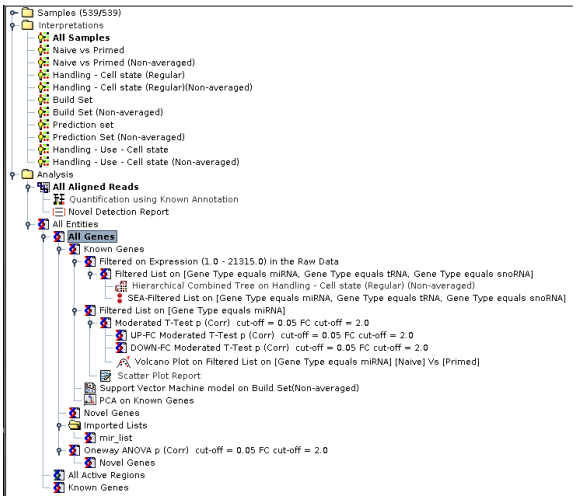


Figure 1: Workflow for Small RNA-Seq analysis

- Samples were aligned to hg19 reference using Strand NGS aligner based on Burrows-Wheeler Transform (BWT) approach. The small RNA reference was annotated with gene annotations comprising of miRNA from miRbase (v20), tRNA from tRNAscan-SE and snoRNA, snRNA, piRNA from Ensembl (e75) which are readily available from Strand NGS. Alignment parameters were set to 1 mismatch allowed, gaps ≤ 0 and mismatch penalty of 4. Adapters were trimmed and UMIs were used for de-duplication of reads.

- Post Alignment, small RNA analysis experiment was created and Quantification (Quantile Normalization) was performed. During Quantification, Novel genes were also detected.

- Genes were filtered after quantification to retain only those genes which have at least one read aligning to it. These genes were then clustered both on gene and group level with Clustering Algorithm as Hierarchical, Similarity Measure as Euclidean, and Linkage Rule as Wards.

- Also, Parametric test (ANOVA) with Asymptotic p-value computation and Benjamin Hochberg FDR multiple testing correction was performed to find the differentially expressed genes.

- From the known small RNAs, the miRNAs were retained and other small RNAs were filtered out and a scatter plot was created. Here, entities with fold change ≥ 2.0 were also identified. From Faridani, O. et.al[1], list of miRNA reported to be differentially expressed were imported and highlighted in the scatter plot.

- In the Experimental Grouping, the samples, handled “Regular”, were grouped into build and validation data set and Interpretations were created accordingly. Out of 204 samples, assigning randomly, 90% samples were used for model building (Support Vector Machine algorithm with N-fold validation step and kernel type as polynomial) and 10% were used for the Model prediction step.

Results and Discussion

Alignment QC

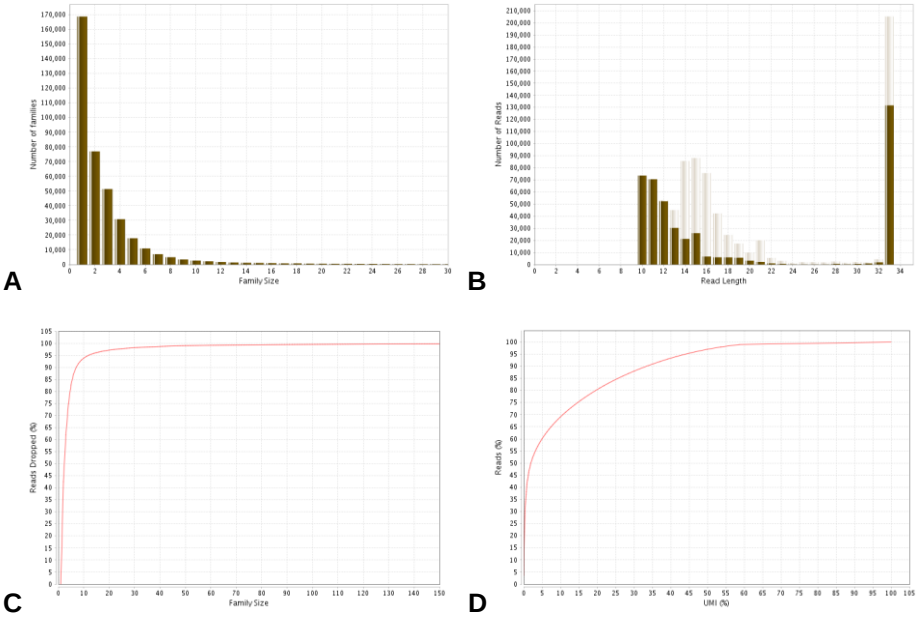


Figure 2: For the sample SRR3495421 **A**) UMI distribution (%) **B**) Read Length Distribution **C**) Cumulative Family Size Distribution **D**) Family Size Distribution

Alignment against hg19 reference genome was completed for all the samples in the dataset. After trimming adapters, UMIs used during the library preparation were used for the de-duplication of reads.

In Figure 2A, the pre-alignment QC plot illustrates the distribution of UMIs in a representative sample and similar distribution was observed in all the samples. Figure 2B indicates that shorter reads are probably of miRNAs and longer reads are composed of tRNA and snoRNA and their derived classes. Figure 2C and 2D are the post-alignment QC plots depicting the family size distribution and % of reads dropped according to family size.

Genic Region QC

In small RNA Analysis experiment, Figure 3, shows the Genic Region QC plot for a representative sample. Depending on the cell type, count of small RNA gene types varied, but within a cell type, the samples had almost the same small RNA distribution.

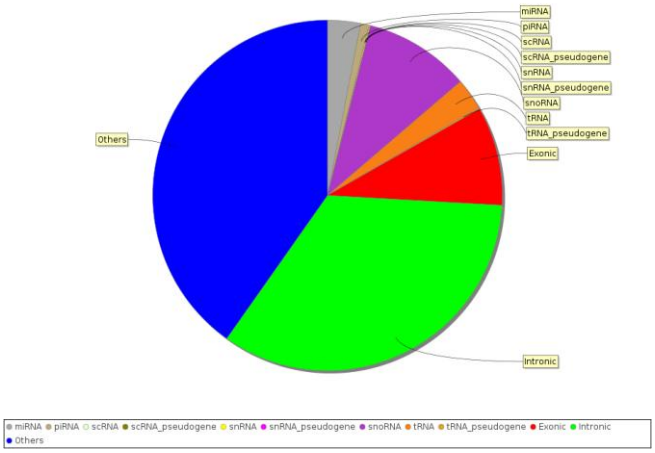
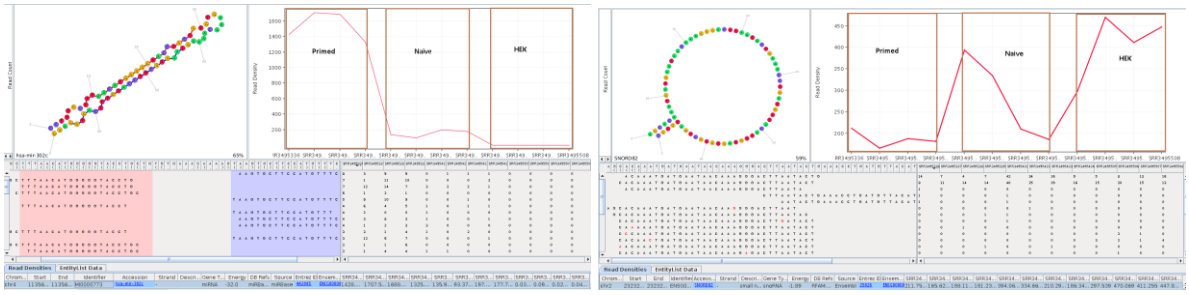


Figure 3: Genic Region QC plot for the sample SRR3495421

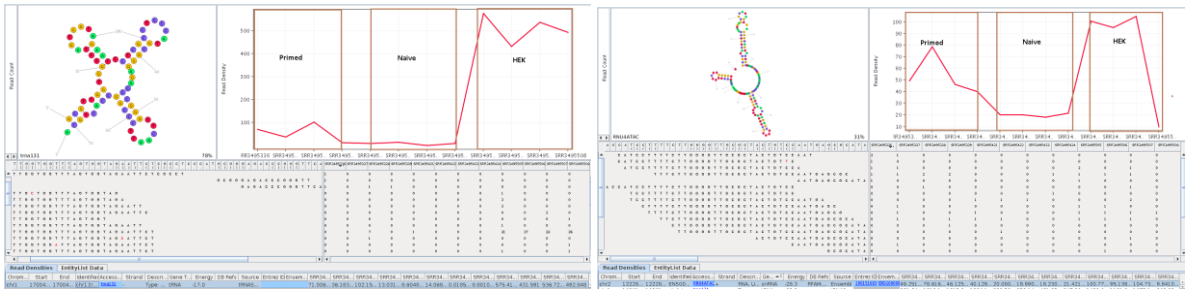
Small RNA Gene View

Gene View was launched with Known Genes entity list and interpretation of “Regular” handling with cell type’s Naive, Primed and HEK cells. In Figure 4, the gene view for different small RNA types can be visualized with different profile plots and read counts across samples.



A

B

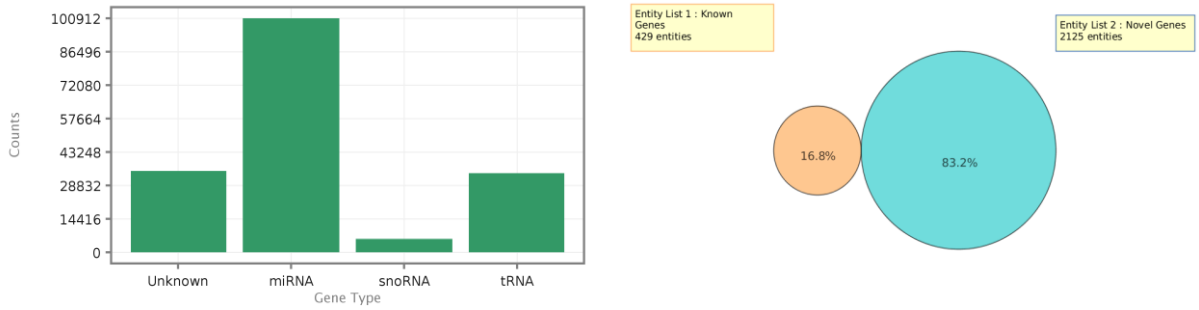


C

D

Figure 4: Gene view launched with a smaller dataset of Primed, Naive and HEK cells for **A**) miRNA (hsa-mir-302c) **B**) snoRNA (SNORD82) **C**) tRNA (tRNA131) **D**) snRNA (RNU4ATAC)

Novel Small RNA Detection



A **B**

Figure 5: A) Novel small RNAs detected from all the samples **B**) Differentially expressed genes mainly comprising of Novel Genes

Using an in-built algorithm developed based on the approach by Langenberger D et.al[2], Strand NGS detects regions corresponding to novel genes, and classifies them into one of the small RNA types based on their read distribution patterns. Based on this algorithm, 17,60,600 novel small RNAs were detected during quantification from one or more of the 539 cells. In Figure 5, the novel genes classified as Unknown, miRNA, snoRNA, and tRNA can be observed. After performing the parametric test (ANOVA), it could be seen that 2,125 out of 2,554 differentially expressed genes were novel small RNAs as seen in Figure 5B. This leads to a conclusion that more investigation is needed with respect to these novel genes and they could act as potential biomarkers for distinguishing cell types.

Clustering Analysis

Dendrogram in Figure 6 depicts that the naive and primed cells can be distinguished from the HEK cells based on their expression profiles, indicating that these small RNAs could act as biomarkers for identifying cell types.

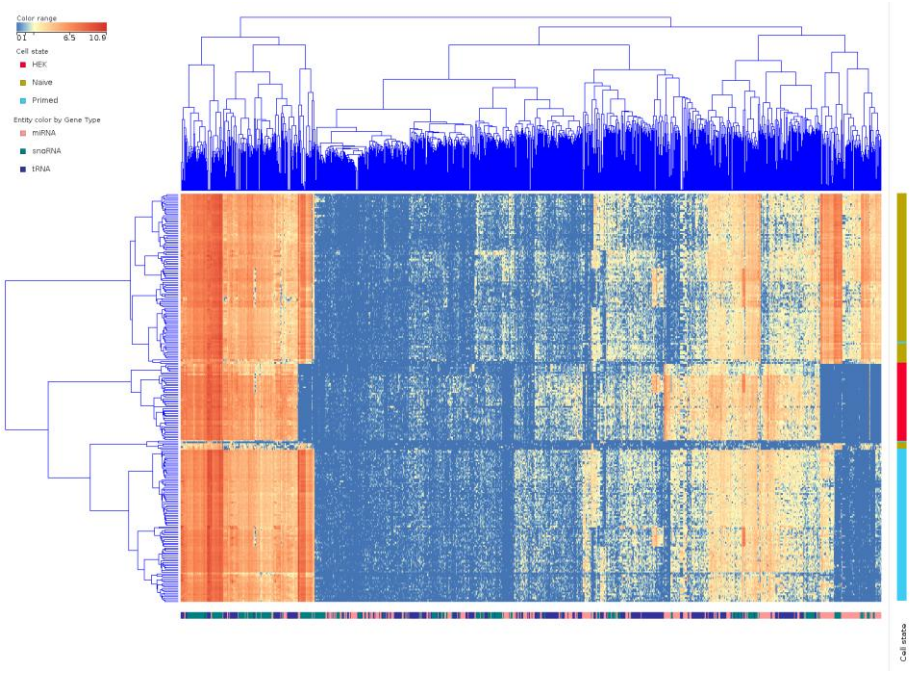


Figure 6: Clustering of entities based on cell types Naive, Primed and HEK cells.

Scatter Plot

Scatter plot in Figure 7 depicts the expression level of known miRNAs in Naive vs Primed samples. The miRNAs highlighted in green are reported by the authors[1] as differentially expressed and our results match theirs.

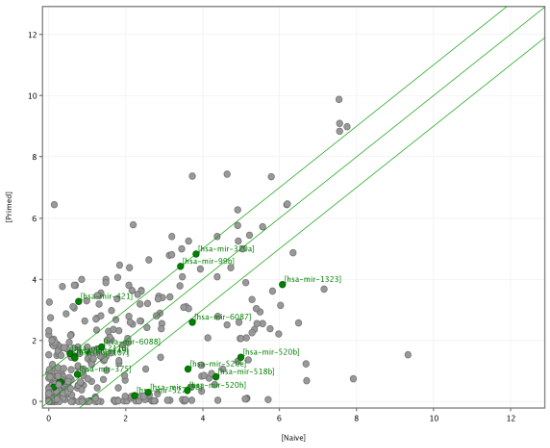


Figure 7: Scatter plot of Naive vs Primed

Prediction Model

Classifier model was built with Support Vector Machine algorithm and it had an overall accuracy of 100%. PCA plot in Figure 8 captures the cell type separation achieved by known small RNAs. The validation data set was predicted accurately using the classifier built model.

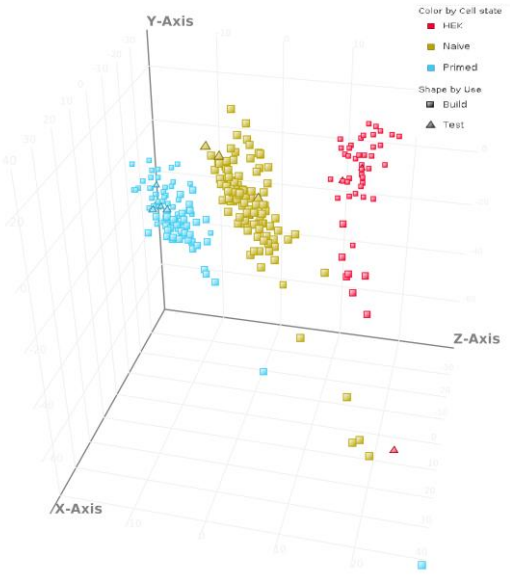


Figure 8: PCA plot of samples used for class prediction

Conclusions

The novel small RNA detection capability of Strand NGS helped us in understanding the complexity of the small RNA transcriptome in hESCs and highlighted how most of it remains uncharacterized. The small RNAs can be used as potential biomarkers to distinguish between Naive, Primed and HEK cells based on their expression profiles.

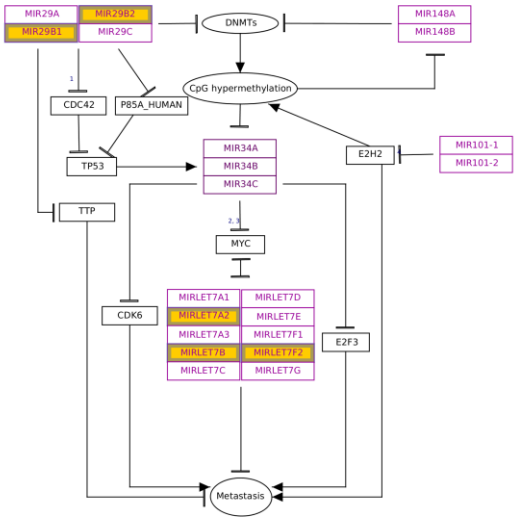
System specifications

Alignment, Quantification, and Analysis were performed on a machine with 32 core processor, 64 GB RAM and 12 TB of storage space.

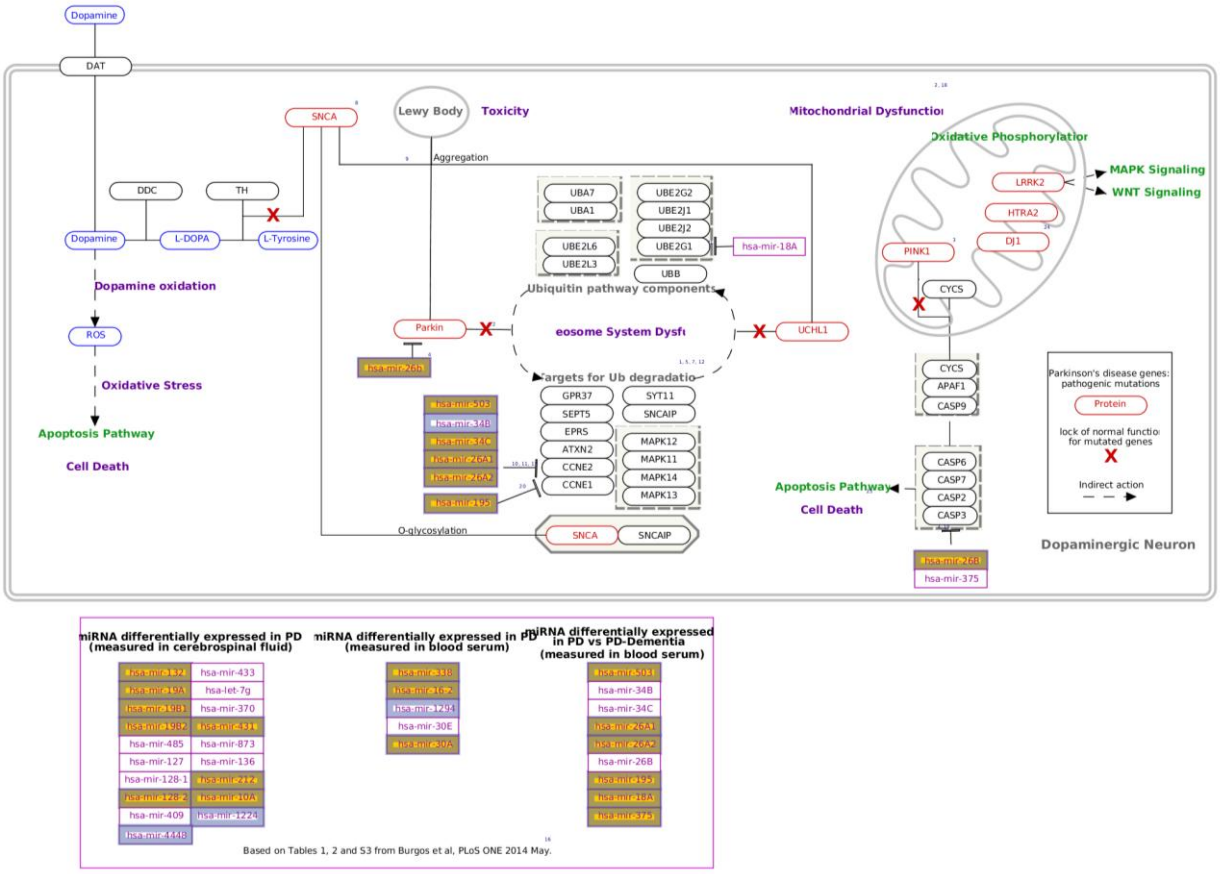
References

1. Faridani, O., Abdullayev, I., Hagemann-Jensen, M., Schell, J., Lanner, F., & Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nature Biotechnology, 34(12), 1264-1266. doi: 10.1038/nbt.3701
2. Langenberger, D., Bermudez-Santana, C., Stadler, P., & Hoffmann, S. (2009). Identification and Classification of Small RNAs in transcriptome sequence data. Biocomputing 2010, 80-87. doi: 10.1142/9789814295291_0010

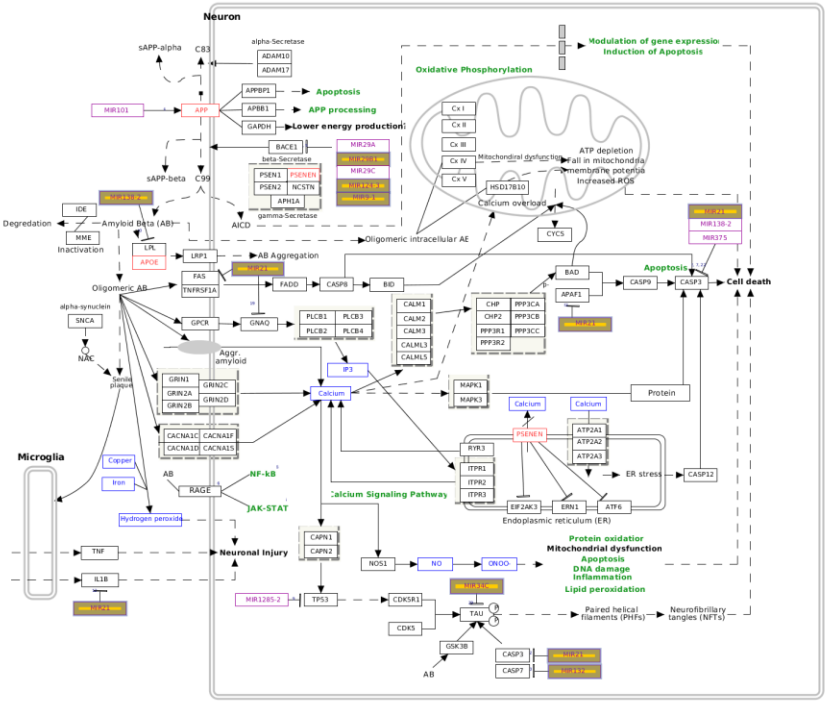
Supplementary Figures:

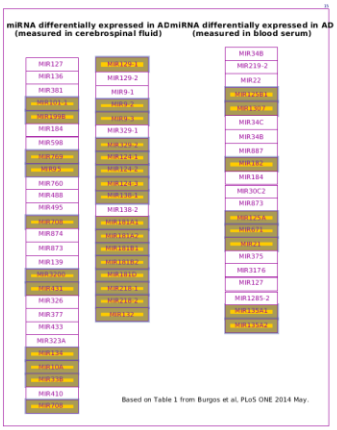


A



B





C

Figure 1: Pathway Analysis ($p \leq 0.01$) on *Homo sapiens* WikiPathways using differentially expressed miRNAs, tRNAs and snoRNAs (Highlighted in Yellow) **A**) Metastatic brain tumor pathway ($p \leq 1.26E-4$) **B**) Parkinsons Disease pathway ($p \leq 5.91E-14$) **C**) Alzheimers Disease pathway ($p \leq 2.80E-24$)