README

for analysis seen in "Defining the breast cancer exosomal proteome: Strategy to unbiasedly characterize exosome vesicles"

code by Will Fondrie

Decription

This collection of scripts were used to analyze the proteomics data from the manuscript titled, "Defining the breast cancer exosomal proteome: Strategy to unbiasedly character exosome vesicles" by David J Clark et al. The analysis performed by these scripts uses the peptide ratios that were calculated using Quantimore (formerly IsoQuant), calculates the protein ratios and performs SVM multivariate analysis to cluster proteins into groups likely to be of exosomal origin or not from the exosome.

Experiment Detail Shotgun proteomics data was acquired using tri-plexed tandem mass tags (TMT), specifically TMT-129,130 and 131 in 2 replicates. The 3 tags represent different points in a traditional exosome isolation strategy, where TMT-129 is a 10,000 x g pellet, TMT-130 is a 100,000 x g supernatant, and TMT-131 is the exosome fraction from an Optiprep density gradient. The raw mass spectra were searhed using Comet. Peptide IDs were validated using PeptideProphet, keeping all peptides above 0.8 PeptideProphet probability (1% FDR) for quantitation. ProteinProphet was used to infer protein evidence using all identified peptides. Peptide quantitation was performed on the filtered PeptideProphet results using QuantiMORE. These results are used as the input for these scripts.

Inputs for analysis

Peptide Quantiation Results from QuantiMORE
"noC" indicates that we used Excel search a replace to replace all commas with semi-colons in order so that they don't interrupt parsing.

- all_peptide_Rep1_noC.txt
- all_peptide_Rep2_noC.txt

Results from ProteinProphet
These are the results from ProteinProphet exported as a excel spreadsheet then saved as csv.

- ProteinProphet_Combined.csv

FASTA Protein Database
This is needed to get the protein descriptions.

- UniProtKB-human-2014_DECOY.fasta

The Exosome and Non-Exosome Markers
These are the markers we hand-picked as a training set for our SVM analysis.

- markers.csv

DAVID Results
*After running **ClusterAnalysis.R**, a lists of gene names for the Exosome cluster and Non-Exosome cluster are generated (GN_Exo.txt and GN_Non.txt, repectively). These were entered into DAVID and the "Functional Annotation Chart" was saved as DAVID_Exo.txt and DAVID_Non_Exo.txt. Additionally, GoTerms.csv is needed to define the selected gene ontology cellular component terms*

- DAVID_Exo.txt
- DAVID_Non_Exo.txt
- GoTerms.csv

R scripts

1. **ProtQuant.R** - *Depends on ProtQuant_Functions.R.* This script is used to calculate the protein ratios for each replicate, independently based on the peptides assigned to each protein by ProteinProphet. The protein ratio is calculated as the average of the peptide ratios assiged to it, weighted by the number of quantified PSMs for each peptide. An equation to represent this calculation is shown below, n represents the number of peptides quantified for a given protein:

$$ProteinRatio = \frac{\sum_{i=1}^{n} PSMs_i * PeptideRatio_i}{\sum_{i=1}^{n} PSMs_i}$$

2. **ProtQuant_Functions.R** - Contains the functions needed to read in the ProteinProphet files and perform the protein quantitation in *ProtQuant.R.*
3. **ClusterAnalysis.R** - Performs the SVM cluster analysis on the protein ratios obtained from *ProtQuant.R.* For our final anlysis, the SVM parameters were optimized over 100 iterations (time=100) and 5x cross-validation (xval=5). This was performed based on the pRoloc tutorial on Bioconductor.
4. **Make_Plots.R** - *Depends on Plot_Functions.R.* Creates figures to represent our data using ggplot2. These were further annotated using Adobe Illustrator to to yield the figures seen in the published manuscript. Note that results from this script are written to a "Figures" folder in your working directory
5. **Plot_Functions.R** - Contains the R code to creat most of the presented figures using ggplot2.
6. **Make_Tables.R** - Creates intelligible tables from the dataframes used in this analysis. These tables are the unformatted versions of the tables seen in the published manuscript. Note that the results from this script are written to a "Tables" folder in your working directory.

Order to Run R Scripts To repeat our analysis, you should run the R scripts in the order shown below. Additionally the scripts must be located in your working directory.
ProtQuant.R -> **ClusterAnalysis.R** -> **Make_Plots.R** -> **Make_Tables.R**

Final Outputs

Figures

1. **ClusterPlot.pdf** and **ClusterPlot.tiff** - A plot of our SVM cluster analysis results with the Exosome and the Non-Exosome markers shown in black.
2. **PM_MarkerPlot.pdf** and **PM_MarkerPlot.tiff** - A plot of our SVM cluster analysis results with the plasma membrane markers from pRoloc shown in black.
3. **ScatterPlot.pdf** and **ScatterPlot.tiff** - A scatter plot of the Log2 protein ratios for all of our quantified proteins.
4. **ScatterPlot_M.pdf** and **ScatterPlot_M.tiff** - A scatter plot of the Log2 protein ratios for all of our quantified proteins with Exosome and Non-Exosome markers shown in color.

Tables

1. **SI-1_Proteins.csv** - contains a detailed protein list with protein quantitation information.
2. **SI-2_Peptides.csv** - contains a peptide list with the assigned protein and peptide quantitation information.
3. **SI-3_Krona.csv** - contains a mapping of proteins to selected gene ontology cellular component terms used in our Krona plots (seen in published manuscript).
4. **table_1_ExosomeMarkers.csv** - contains the list of exosome and non-exosome markers with protein quantitation information.
5. **table_S1_AllProteins.csv** - contains a simplified protein list.
6. **table_S2_PMMarkers.csv** - contains a list of the plasma membrane markers used from the pRoloc markers data set.
7. **Krona_GO_Counts.csv** - contains the number of gene hits for each selected gene ontology cellular component term used in our Krona plots (seen in published manuscript).