Github README file

Depository of all R files for Extensive protein dosage compensation in aneuploid human cancers

This is a repository for the Genome Research article "Extensive protein dosage compensation in aneuploid human cancers" by Klaske M Schukken and Jason M Sheltzer

The READ\_ME\_R.xlsx is an overview of which R file generates which datasets and graphs, also listed below.

Many of the code files need additional datasets retrieved from outside sources-- The source of the datasets are often listed in the ## Get data ## section of the R files. If data source is not listed in the R code itself, it will be credited in the "Extensive protein dosage compensation in aneuploid human cancers" article. Several intermediate datasets are made available for ease of use.

All code was build on R version 3.6.3 GUI 1.70 El Capitan build, and R studio version 1.1.383.

In order to run the code, open the .R files, download the indicated datasets, set the proper working directory, and run through the code.

ORGANIZATION

R CODE:

Protein\_RNA\_filtered\_CellLine.R

* All figures
* Filter protein expression and RNA expression data to isolate cell lines with RNA expression, protein expression, and arm call data

Protein\_filtered\_analysis\_3cat.R

* Figure 1, S1
* Calculate protein expression difference for all genes upon chrm arm gain or loss (ex. gain chrm 5, protein expression diff in chrm 1,2,3,4,5,6, etc. )

RNA\_filtered\_analysis\_3cat.R

* Figure 1, S1
* Calculates RNA expression difference for all genes upon chrm arm gain or loss (ex. gain chrm 5, RNA expression diff in chrm 1,2,3,4,5,6, etc. )

ChrmArm.Scatterplot.Protein.RNA.R

* Figure 1
* Protein and RNA expression difference per chrm arm scatterplot (ex. Gain chromosome 5q: protein expression difference on chrm 5q, and on other chromosomes)

DNA\_Protein\_RNA\_filtered.R

* Figure 1, S1
* Make graphs of chrm arm difference from neutral-ploidy, for specific cells (ex. MCF7)

Protein\_RNA\_Heatmap\_filtered\_min10percategory.R

* Figure 1
* Generate heatmaps: protein/RNA expression difference upon chrm arm gain/loss, and additional summary heatmap

Protein\_RNA\_expression.PerCell\_v3.R

* Figure 2,3,5,6, S2, S3, S4, S5, S8
* Difference upon chrm gain/loss dataset(s), categorize (scaling, buffering, anti-scaling). Boxplots of RNA or Protein expression vs DNA copy number category (gain, neutral, loss). Near diploid vs near triploid difference analysis. High aneuploidy score and low aneuploidy score difference analysis. Non-coding RNA (microRNA) analysis.

Protein\_RNA\_1d.plot\_v2

* Figure 3, 6, S2, S6
* Plot 1dimentional density plot of gene expression changes per group (Protein and RNA). Also look at low vs high aneuploidy cells only

gProfile\_Quantile\_Bargraph\_v2.R

* Figure 3, 4, S5, S6, S7, S8
* Make graphs of key Gene Ontology enrichment terms (for multiple analysis')

Yeast\_proteomics\_v2.R

* Figure S6
* Aneuploid yeast analysis (Protein & RNA)

DownSyndrome\_analysis\_v2.R

* Figure S6
* Down syndrome fibroblast analysis (Protein and RNA)

Stingele\_Protein\_filter\_v2.R

* Figure S6
* Stable aneuploid cell line analysis (Protein and RNA)

Aneuploid\_Score.R

* Figure 4, S7
* Generate cellular aneuploidy scores

Protein\_expression\_data\_GeneScore.R

* Figure 4, S7, S8
* Protein expression correlate with cellular aneuploidy, and high/low aneuploid cell only analysis

CCLE\_RNA\_Analysis\_geneAn.R

* Figure S7
* RNA aneuploidy score correlation analysis

Variance\_aneuploidy.R

* Figure 5
* Calculate RNA and Protein neutral-ploidy variance

Protein\_buffering\_factors\_v2.R

* Figure 5
* Protein buffering factors. Calculate ROC AUC and make boxplots of scores/values

TSG.OG\_CNV\_Difference\_v2.R

* Figure 6
* Oncogene and tumor suppressor gene difference upon chromosome copy number changes and gene copy number changes

TCGA\_ProteinDiff\_Ovarian.R

* Figure 7, S9
* Analysis of ovarian tumor data from TCGA.

ADDITIONAL FILES:

arm\_calls\_data\_bendavid.csv

* Aneuploid arm calls per chromosome arm per cell line analyzed. Used to generate aneuploidy score per cell line, and used to identify which cell lines have chromosome arm gains and losses per chromosome arm. Data from Cohen-Sharir et al. 2021 Nature, from the lab of Dr. Uri Ben-David.

bp\_per\_arm.csv

* number of basepairs per chromosome arm. Used to generate a basepair-based aneuploidy score

Score2.aneuploid.cell\_line.csv

* Cellular aneuploidy scores per cell line. Multupiple forms of the aneuploidy score are calculated, the gene\_ploidy\_score was used in the manuscript unless otherwise indicated.

RNA\_Protein\_Shared\_cells.csv

* Cell lines with both RNA expression data and protein expression data. Not all of these cells are also present in the Cohen-Sharir arm call data.

Protein\_location\_info.csv

* Information about the chromosome arm location per gene.

Protein\_ID\_info.csv

* Protein\_IDs matched with gene symbol and uniprot accesion

RNA.Protein\_loss.neutral.gain\_Difference\_Pvalue\_min10points.csv

* Dataset of gene expression difference on cells with chromosome gain or loss, relative to cells with a neutral ploidy for chromosome arm a gene is located on.

RNA.Protein\_loss.neutral.gain\_Difference\_Pvalue\_min10points\_3cat.csv

* Dataset of gene expression difference on cells with chromosome gain or loss, relative to cells with a neutral ploidy for chromosome arm a gene is located on.  
  with gene classifications as either "Anti-Scaling", "Buffering" or "Scaling" upon chromosome gain and loss

RNA.Protein\_loss.neutral.gain\_Difference\_Pvalue\_\*\*\*\*.csv

* Dataset of gene expression difference, as above, for a subset of cells and/or genes corresponding to the indicated control dataset. genes classifications as either "Anti-Scaling", "Buffering" or "Scaling" upon chromosome gain and loss.

Trisomy 21 comparison.xlsx

* Protein expression difference between Down Syndrome (Ts21) fibroblasts and control fibroblasts, set relative to cancer cell line protein expression difference. Yeast data from: Liu et al. 2017, Nature Communications. Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells

yeast-human aneuploidy.xlsx

* Yeast protein expression difference upon chromosome gain. Data originated from: Dephoure et al. 2014, eLife; 3:e03023 Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast

AmplificationRatio\_perGene1.75.csv

* Calculation of the percent of cells in the filtered dataset with gene amplifications per gene. Used in initial factor analysis, but not included in the final manuscript. Data from CCLE gene copy number data.

Protein.AllFactors.csv

* A list of all genes, difference upon chromosome gain or loss, and factor scores/values per gene. 5' UTR length and 3' UTR length were not included, because there are frequently multiple 5' or 3' UTR lengths per gene, and this duplication can skew data when investigating other factors. See Protein\_buffering\_factors\_v2.R, to incorporate 5' UTR length and 3'UTR length if desired.

Protein.AllFactors\_Ovarian.csv

* A list of all genes, difference upon chromosome gain or loss, and factor scores/values per gene for the ovarian tumor dataset.

Factor.ROCAUC.correlationscore.csv

* A list of all investigated factors, their AUC ROC values with regard to buffering, scaling and anti-scaling genes, and the correlation coefficients with RNA and Protein expression difference.