1. TG\_GATEs\_2rd code is responsible for preprocessing raw data,

It extracts all the file names for raw data at the beginning

Then it uses rat\_invivo\_liver\_single\_2rd.sh (and other .sh scripts) and get\_exprs\_2rd.R with these file names to get differential gene expression data.

Then it merges all the differential gene expression data for different sources (rat liver singe, rat liver repeat)

Then it finds correspond label for combined differential gene expression data and stores them as csv file.

And it also is used to combined enrichment scores.(see below)

1. Get gene set enrichment score for average DE data

Use rat\_inVivoLiverSingle\_enrichmentScore\_2rd.sh and enrichmentScore\_2rd.R to get enrichment scores for all gene sets across all conditions:

For enrichment\_2rd, it calculates enrichment scores for each compound in different time and different dose levels. Enrichment scores of all gene sets for every drug, time and dose level conditions is a sample in the final data

Then the code block in TG\_GATEs\_2rd code is used to combined all the enrichment scores for all the drugs in to one matrix and extract corresponding pathology labels for them

Use rat\_inVivoLiverRepeat\_enrichmentScore\_2rd.sh and enrichmentScore\_2rd.R to get enrichment scores for all gene sets across all conditions ……

1. Get gene set enrichment score for EachRepli DE data

Use rat\_inVivoLiverSingle\_enrichmentScore\_ EachRepli \_2rd.sh and enrichmentScore\_EachRepli \_2rd.R to get enrichment scores for all gene sets across all conditions:

Then the code block in TG\_GATEs\_2rd code is used to combined all the enrichment scores for all the drugs in to one matrix and extract corresponding pathology labels for them

Use rat\_inVivoLiverRepeat\_enrichmentScore\_ EachRepli \_2rd.sh and enrichmentScore\_ EachRepli \_2rd.R to get enrichment scores for all gene sets across all conditions ……

1. Get gene set enrichment score for molecular network modules and co-expression newtowrks

Use Disease\_CoExp \_rat\_inVivoLiverSingle\_enrichmentScore\_2rd.sh and Disease\_CoExp \_enrichment\_2rd.R to get enrichment scores for all gene sets across all conditions:

Then the code block in TG\_GATEs\_2rd code is used to combined all the enrichment scores for all the drugs in to one matrix and extract corresponding pathology labels for them

Use Disease\_CoExp \_rat\_inVivoLiverRepeat\_enrichmentScore\_2rd.sh and Disease\_CoExp \_ enrichment\_2rd.R to get enrichment scores for all gene sets across all conditions ……

1. Get\_normalizedData\_new.R and human\_invitro\_normlized\_new.sh are used to get normalized data,i.e. the absolute expression data ----these scripts may not be right as it log raw data twice
2. Get\_AveNormalizedData\_2rd.R and rat\_inVivoLiverSingle\_AveNormalized\_2rd.sh are used to get average normalized data, i.e. the average absolute expression data

Then the code block in TG\_GATEs\_2rd code is used to combined all average expression data for all the drugs in to one matrix and extract corresponding pathology labels for them

1. Get\_EachNormalizedData\_2rd.R and rat\_inVivoLiverSingle\_EachNormalized\_2rd.sh are used to get normalized data for each sample. And during preprocess step, avoid gene filtering to preserve as many genes as possible for later analysis(Such as CIBERSORT)

Then the code block in TG\_GATEs\_2rd code is used to combined all average expression data for all the drugs in to one matrix and extract corresponding pathology labels for them

1. CIBERSORT\_BIOQC.R file is used to get enrichment scores for certain immune gene sets for different pathologies.
2. Get\_corrs\_eachCom\_new.R is used by compound corrs\_new.sh. It uses normalized expression data for each compound from above and then uses get\_time\_interCorr\_new.R/get\_interCorr\_new.R script to calculate intra-correlation of certain gene set at certain time by Camera
3. Old\_data.R compare results with David’s work
4. Tao\_rnaseq\_pipeline.R is used to preprocess rna seq data and get quantile norm and log 2 tpm data and raw count data .
5. Tao\_rnaedgeR.R is used to get differential gene expression data from rna seq data(both EdgeR and voom data)
6. tao\_rnaedgeR\_enrichment.R
7. Tao\_rnaseq\_pipeline\_LTA4H.R is used to preprocess rna seq data and get quantile norm and log 2 tpm data and raw count data .
8. Tao\_rnaedgeR\_LTA4H.R is used to get differential gene expression data from rna seq data(both EdgeR and voom data)
9. tao\_rnaedgeR\_LTA4H\_EachRepli.R
10. tao\_rnaedgeR\_LTA4H\_CAIA.R
11. tao\_rnaedgeR\_LTA4H\_CAIA\_enrichmet.R
12. tao\_rnaedgeR\_LTA4H\_CAIA\_enrichmet\_filtergenes.R
13. tao\_GSE47875\_microarray.R and get\_exprs\_GSE47875.R are used to get DE for microarray data from GSE47875
14. get\_exprs\_ GSE70559.R are used to get DE for microarray data from GSE47875
15. drugmatrix. R is to preprocess data from drugmatrix database and get the final DE data/EachRepli DE data
16. get\_exprs\_drugmatrix.R is used to get DE data for each compound for drugmatrix data
17. get\_exprs\_drugmatrix\_EachRepli.R is used to get EachRepli DE data for each compound for drugmatrix data

rat\_drugmatrix\_EachRepli.sh

1. get\_exprs\_drugmatrix\_EachRepli\_enrichment.R is used

rat\_drugmatrix\_EachRepli\_enrichment.sh

1. get\_exprs\_drugmatrix\_EachRepliAllgenes.R is used to get EachRepli DE data for each compound for drugmatrix data

rat\_drugmatrix\_EachRepliAlgenes.sh

1. get\_exprs\_drugmatrix\_EachRepli\_enrichmentAllgenes.R is used

rat\_drugmatrix\_EachRepli\_enrichmentAllgenes.sh

1. get\_exprs\_drugmatrix\_EachRepli\_enrichment.R

rat\_drugmatrix\_EachRepli\_enrichment.sh

1. get\_exprs\_drugmatrix \_EachNormalizedData.R is usd to get …..
2. get\_exprs\_EachRepli.R is used to get DE data for each replicate (each exp condition has 3 replicates) for TG-GATEs data ,rum by rat\_invivo\_liver\_single\_EachRepli.sh and rat\_invivo\_liver\_repeat\_EachRepli.sh
3. get\_exprs\_EachRepli\_Kidney.R is used to get DE data for each replicate (each exp condition has 3 replicates) for TG-GATEs kidney data ,rum by rat\_invivo\_kidney\_single\_EachRepli.sh and rat\_invivo\_kidney\_repeat\_EachRepli.sh
4. TG\_GATEs\_Drugmatrix\_combat.R is used to remove batch effect for TG\_GATEs and drugmatrix data
5. last\_TG\_GTATs\_Drugmatrix\_comat
6. TG\_GATEs\_drug\_analysis\_EachRepli.R is used to TG-GATEs drug information as drug toxicity from Each Repli De data
7. Overcomparison\_results\_10floder\_analysis.R is used to analyze the results from 10 folder with/with prediction results analysis
8. last\_model\_comparison\_10folder\_analysis.Rmd is used to analysis the model for all kinds of 10 folded data analysis
9. last\_evaluation\_comparison\_10folder\_analysis.Rmd is used to analysis the evaluation results for all kinds of 10 folded data analysis
10. last\_ASO\_check.Rmd is used to check and preprocess raw data
11. Data sources and link

“ReactomePathways.gmt” is downloaded from https://reactome.org/download-data/Specialized data formats/[Reactome Pathways Gene Set](https://reactome.org/download/current/ReactomePathways.gmt.zip)

“ReactomePathwaysRelation.txt” and “ReactomePathways.txt” are downloaded from <https://reactome.org/download-data/Pathways>

“goa\_human.gaf” is download from http://geneontology.org/page/download-go-annotations

Goa/gaf formart: <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/README>

“go-basic.obo” is downloaded from http://www.geneontology.org/page/download-ontology

1. go\_preprocess\_tree.R is used for go data preprocess and tree function test
2. Comparisons with other tools/algorithms
3. Gene\_pathway\_matrix.R is used to construct gene pathway matrix and get raw go gene list(raw\_go\_list.rda)

4. “Module\_Identification\_GO\_REA\_UNI\_2rd.R” is used to test new model. I.e. map to hierarchical trees and draw heatmap of modules for reactome terms

5. . “Module\_Identification\_funcs.R” is used to store all the supplementary functions for module identification project

6. “Module\_Identification\_GO\_REA\_UNI\_all\_2rd.R ” is used to get module identification for all modules

7. “Module\_Identification\_GO\_REA\_PPI\_STRING\_test.R” is used to test new model after 08.14 meeting

8. “Module\_Identification\_GO\_REA\_PPI\_STRING.R” is used to get module identification for all modules

9. Module\_Identification\_suplementary\_figures.R is used to draw supplementary figures as described in :https://docs.google.com/document/d/1w-RcyMFWbx3fictyHthqdfu2sDgf0pvdnb0H-Z\_5iNY/edit#

Module\_Identification\_suplementary\_figures\_sup.R is used to save some old code for Module\_Identification\_suplementary\_figures.R script

Interpret\_TGGATEs.R is used to interpret TG-GATEs model based on the regression based pathway selection methods

regression\_selected\_pathways\_funcs.R is used to write a function to use the model

last\_interpret.Rmd

## Supplementary figures

Potential figures (to be discussed)

* Schematic of method (if helpful, maybe not necessary...)
* Reactome classes associated with modules (analogous to [Fig. 3f](https://docs.google.com/document/d/1tbtZud654W91xKWLOclvqbsxAFB7XaoKvWII5YGEP-M/edit))
* Maybe: For each module associated with a Reactome class (columns), count of module genes of a given UniProt class (in rows) (analogous to [Fig. 3g](https://docs.google.com/document/d/1tbtZud654W91xKWLOclvqbsxAFB7XaoKvWII5YGEP-M/edit))
* Maybe: relationship between regression coefficients and Fisher’s exact p-values. Ideas:
  + Distribution of (a) #significant terms (FDR corrected P-value from noncentral hypergeometric) and (b) #selected terms (regression). E.g., two boxplots or histograms.
  + Scatterplot: #significant terms (Fisher) vs. #selected terms for each module
  + Show that the selected terms are less redundant than the ones from Fisher: E.g., for each significant term, compute max Jaccard index across all other significant terms of that module. Show distribution of these max Jaccard indexes for all significant terms and modules. Do the same for the selected terms.

If using elasticnet, similar go terms in go ontology tree may eliminate similar reactome term sin reactome forest

[http://www.informatics.jax.org/marker/MGI:88596](https://www.google.com/url?q=http://www.informatics.jax.org/marker/MGI:88596&sa=D&source=hangouts&ust=1532521115447000&usg=AFQjCNGHmdFPU7QjBeVpLJ5X2WqB1ThKzQ)

[http://www.informatics.jax.org/gotools/MGI\_GO\_Slim.html#p](https://www.google.com/url?q=http://www.informatics.jax.org/gotools/MGI_GO_Slim.html%23p&sa=D&source=hangouts&ust=1532521543746000&usg=AFQjCNGu1dj1cv7_NB03QzgfAITLj9in2A)

Two tasks (1) make a heatmap of modules/GO root terms (or Reactome root terms) for supplementary figure (2) find out whether we can re-use the 'category' information