RNA-sequence and Differential Expression Analysis Reproduction

The Wings Group:

Dandan Zhang, Basanta Bista, Ping Kang, Michael Murphy, and Samantha Snodgrass

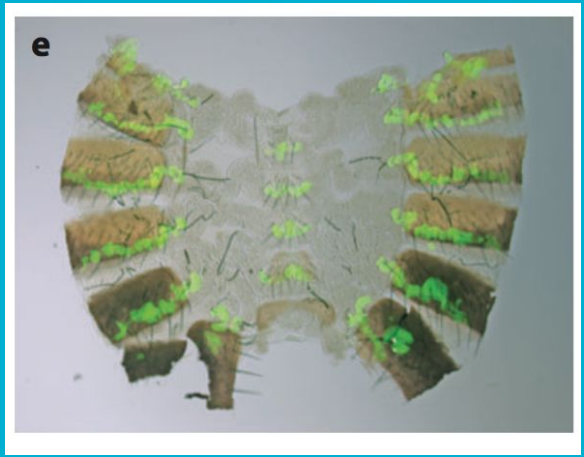
“Ribo-tag translatomic profiling of *Drosophila* oenocyte reveals down-regulation of peroxisome and mitochondria biogenesis under aging and oxidative stress”

- ★ Posted to bioRxiv February 2018
- ★ Test reproducibility within a lab
- ★ Identifying differentially expressed genes using command line tools
- ★ Visualizing differentially expressed genes using R packages

https://en.wikipedia.org/wiki/File:Drosophila_melanogaster.jpg

About the biology

- ★ What is an oenocyte?
- ★ What is Ribo-tagging?
- ★ What was the goal of this study?



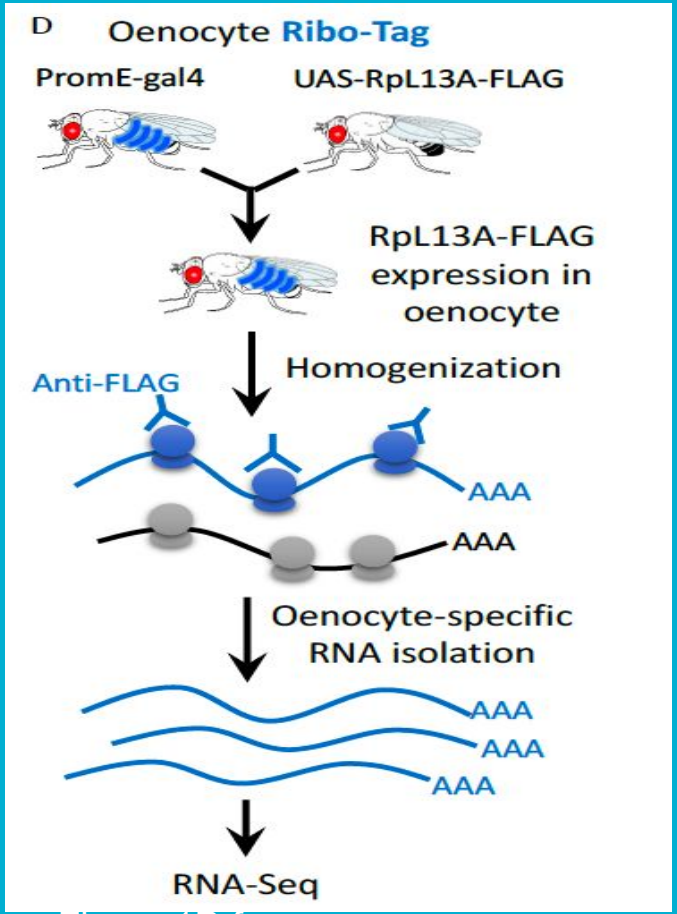


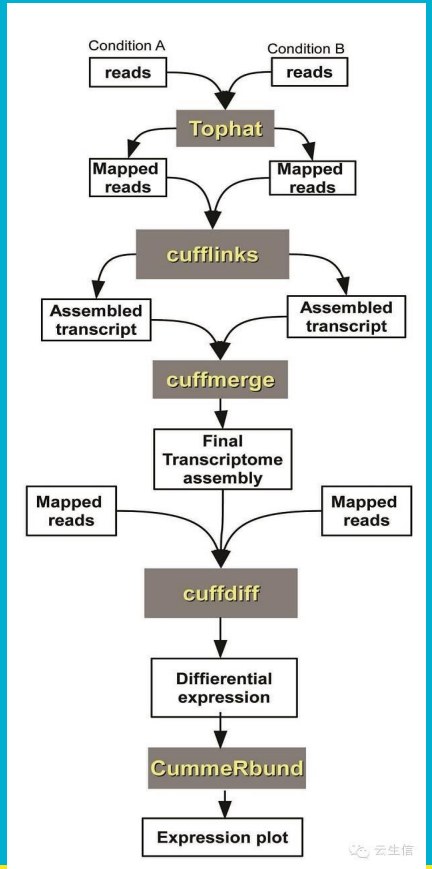
Figure 1D from paper

Description of our group's workflow

- ★ Basanta
  - RNA-seq mapping
  - TopHat through CuffDiff
- ★ Samantha
  - cummeRbund analysis of CuffDiff
- ★ Ping
  - PCA and scatterplots
- ★ Dandan
  - Heatmaps
- ★ Michael
  - GSEA analysis

Description of our group's workflow

- ★ Differential expression
  - Re-running TopHat and CuffLinks packages on Condo
  - Extracting differential expression information from cuffdiff output files
  - Comparing the number of genes identified as differentially expressed between groups
- ★ Recreating the figures
  - PCA results of groups
  - Scatterplots
  - Hierarchical clustering
  - Heat maps

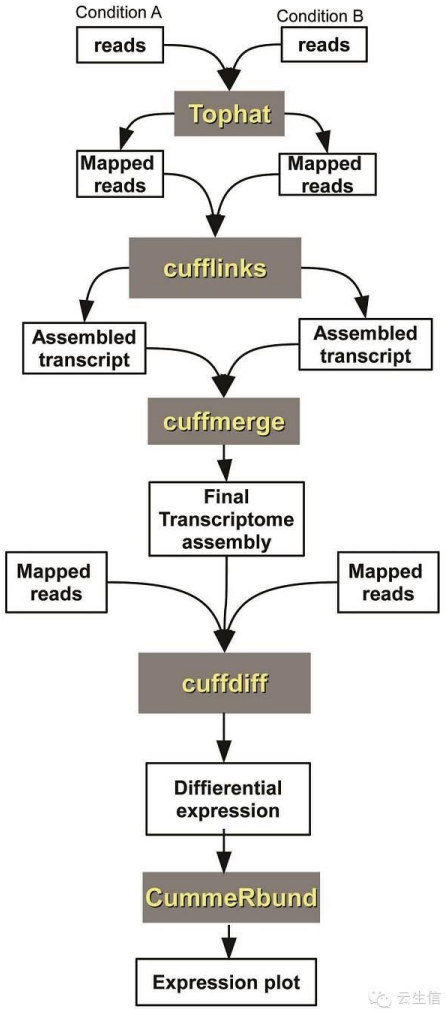


Cufflinks Pipeline

Cufflinks Pipeline run in Condo

Author used Galaxy

web-based platform for data intensive biomedical research



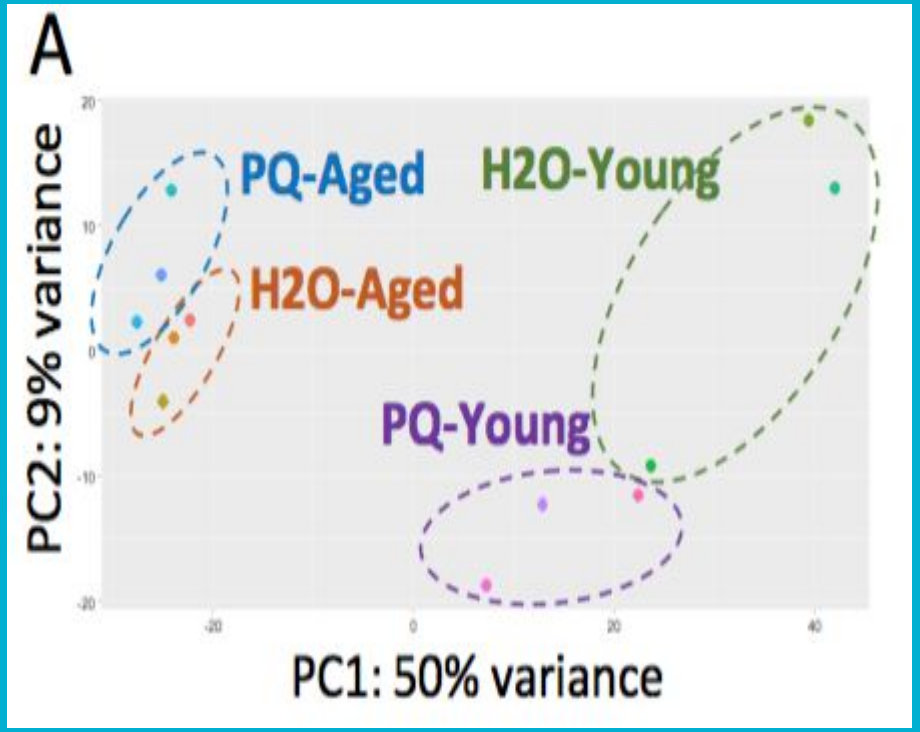
Differential Expression: CuffDiff output files

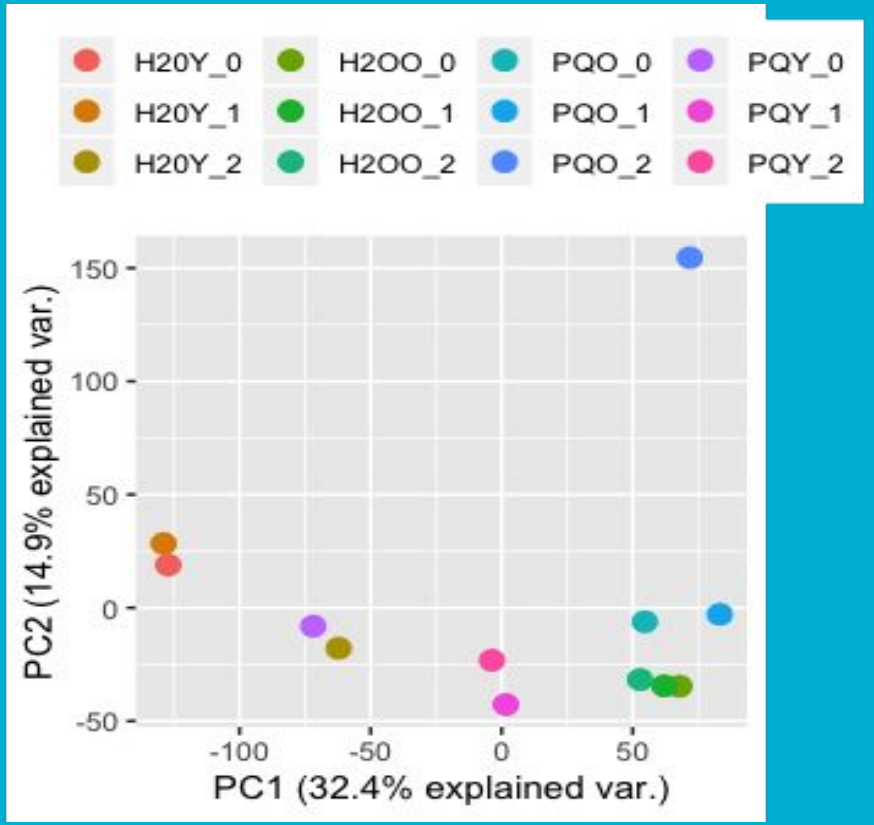
cummeRbund

- ★ R package that can navigate and interpret the many output files from CuffDiff
- ★ Parse out tables for differential expression of particular condition combinations and significance levels
- ★ Requires a gtf (annotation) file to include gene names

gene_id	gene_short_	locus	sample_1	sample_2	status	value_1	value_2	log2_fold_ch	test_stat	p_value	q_value	significant
XLOC_000001	JYalpha	NC_004353.4	H2OO	H20Y	NOTEST	0.12423	0.342782	1.46428	0	1	1	no
XLOC_000001	JYalpha	NC_004353.4	H2OO	PQO	NOTEST	0.12423	0.0721351	-0.784239	0	1	1	no
XLOC_000001	JYalpha	NC_004353.4	H2OO	PQY	NOTEST	0.12423	0.293155	1.23865	0	1	1	no
XLOC_000001	JYalpha	NC_004353.4	PQO	PQY	NOTEST	0.0721351	0.293155	2.02289	0	1	1	no
XLOC_000001	JYalpha	NC_004353.4	H20Y	PQY	NOTEST	0.342782	0.293155	-0.225628	0	1	1	no
XLOC_000001	JYalpha	NC_004353.4	H20Y	PQO	NOTEST	0.342782	0.0721351	-2.24852	0	1	1	no
XLOC_000002	CR45124	NC_004353.4	H2OO	H20Y	OK	2.77952	2.3933	-0.215835	-0.115781	0.8494	0.921666	no
XLOC_000002	CR45124	NC_004353.4	H2OO	PQO	OK	2.77952	2.31319	-0.264951	-0.133616	0.84485	0.919147	no
XLOC_000002	CR45124	NC_004353.4	H20Y	PQO	OK	2.3933	2.31319	-0.0491161	-0.0263429	0.9724	0.984897	no
XLOC_000002	CR45124	NC_004353.4	H20Y	PQY	OK	2.3933	3.06021	0.354627	0.212002	0.71765	0.845256	no
XLOC_000002	CR45124	NC_004353.4	PQO	PQY	OK	2.31319	3.06021	0.403743	0.223784	0.71345	0.842487	no
XLOC_000002	CR45124	NC_004353.4	H2OO	PQY	OK	2.77952	3.06021	0.138791	0.0769429	0.8895	0.943628	no

Principal Component Analysis(PCA) on four oenocyte translatomes

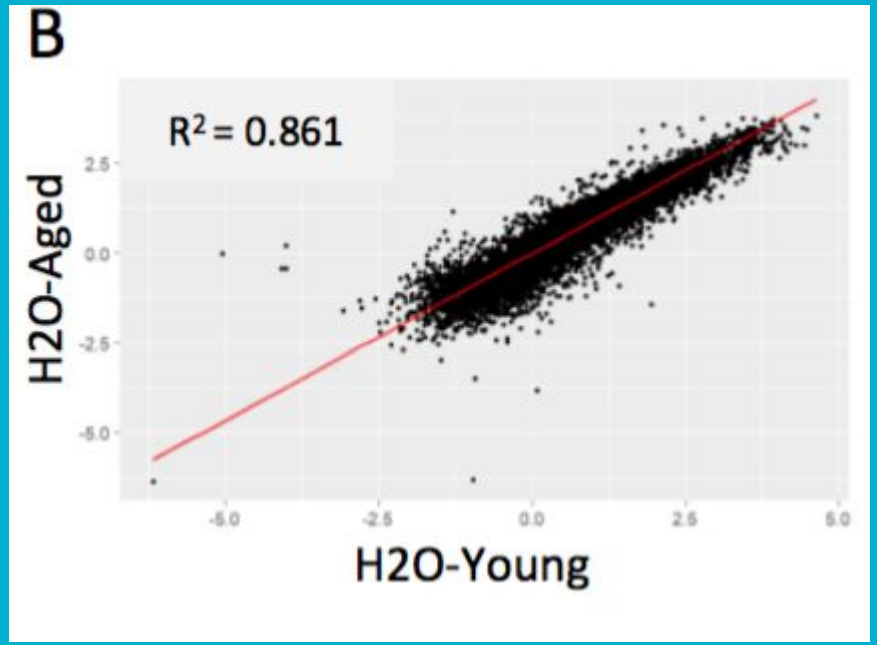


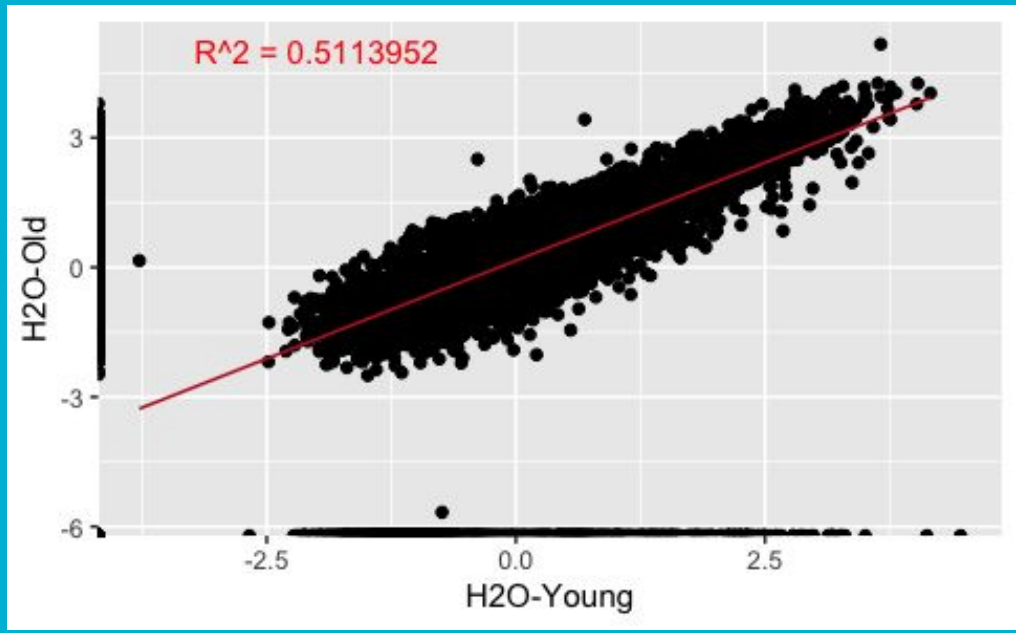


Original

Reproduction

Correlation analysis on the gene expression between H2O-Young and H2O-Old

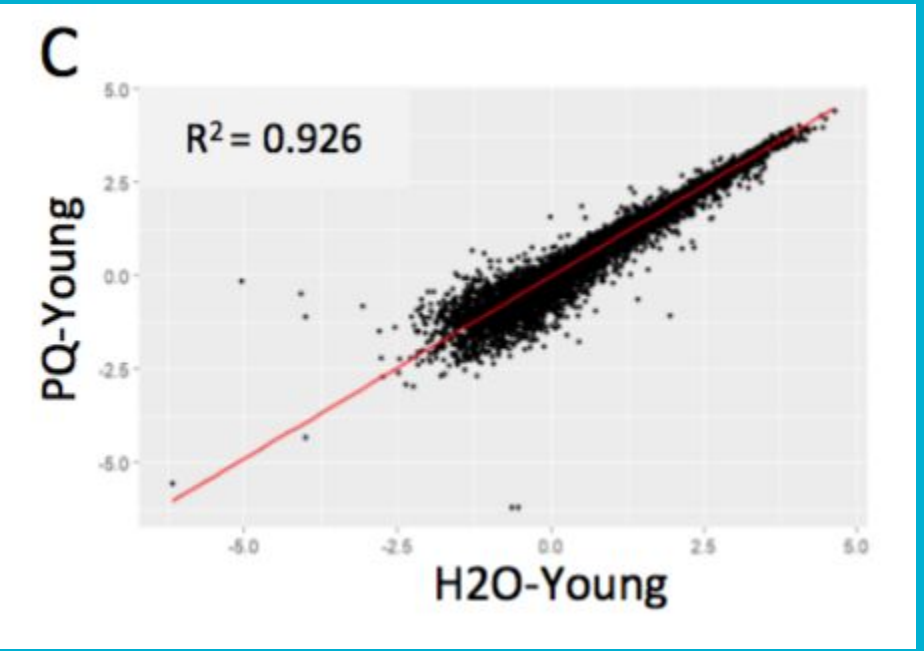




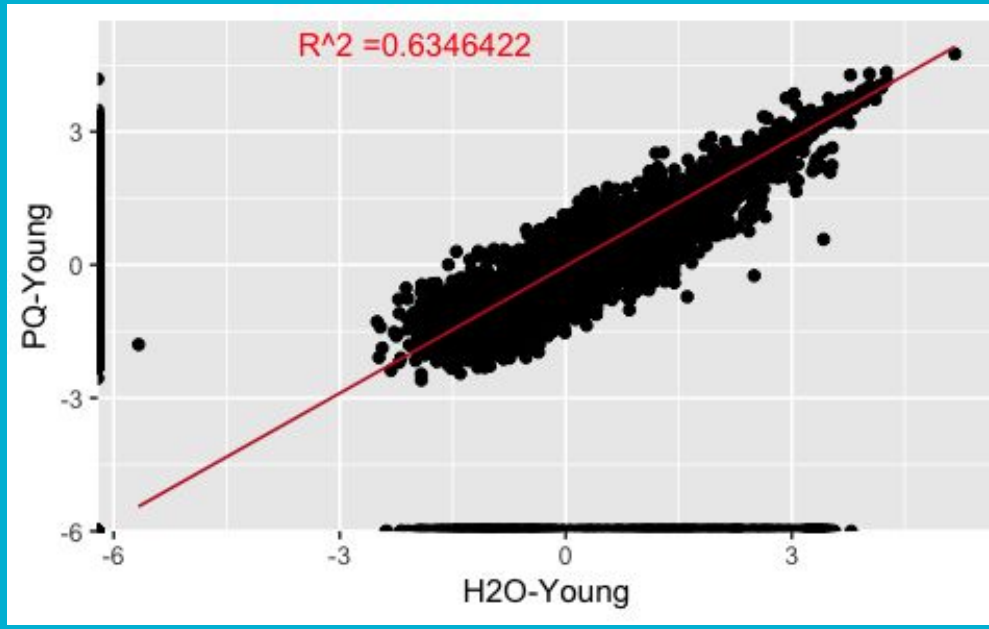
Original

Reproduction

Correlation analysis on the gene expression between H2O-Young and PQ-Young

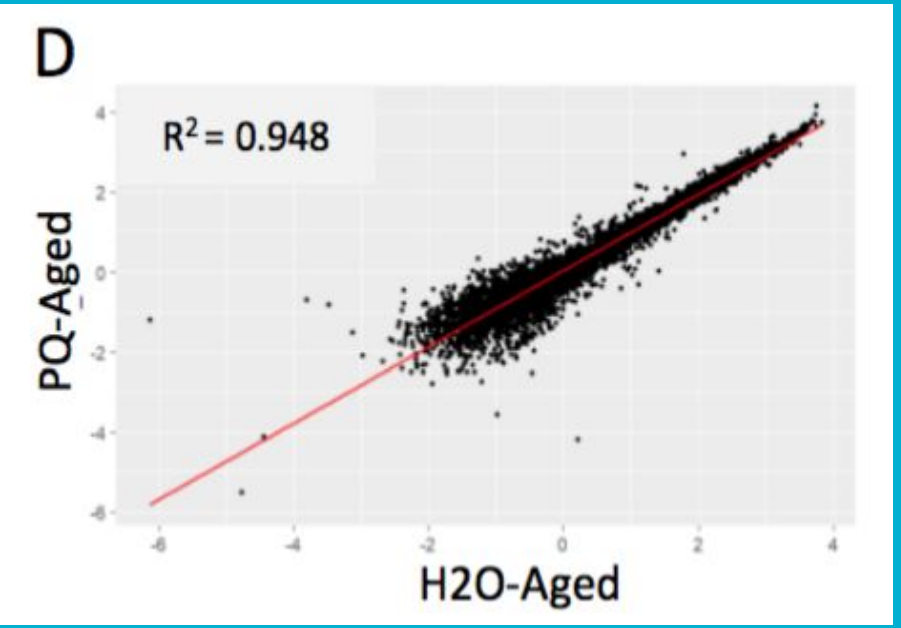


Original

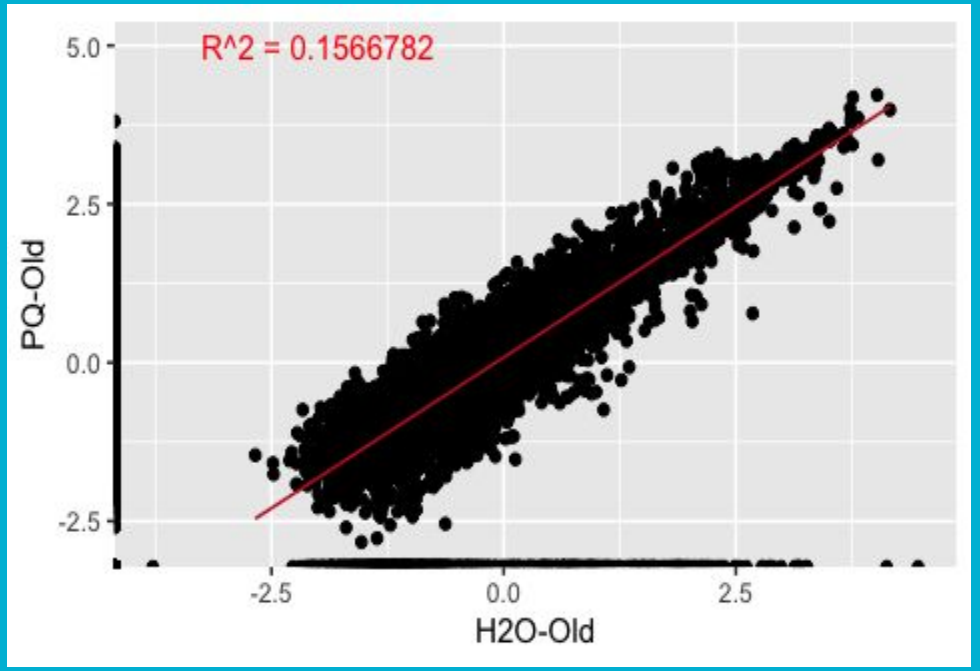


Reproduction

Correlation analysis on the gene expression between H2O-Old and PQ-Old



Original

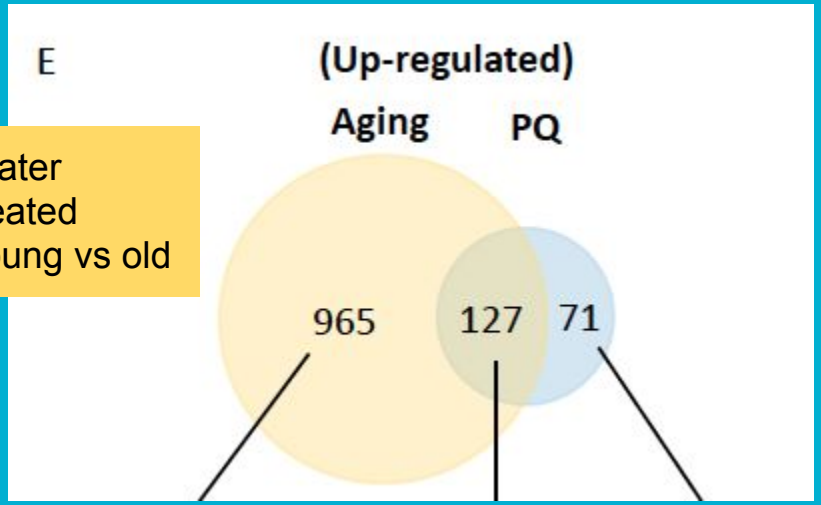


Reproduction

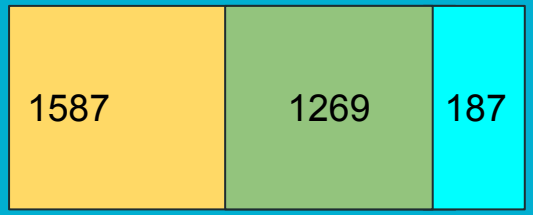
Differential Expression: venn diagrams

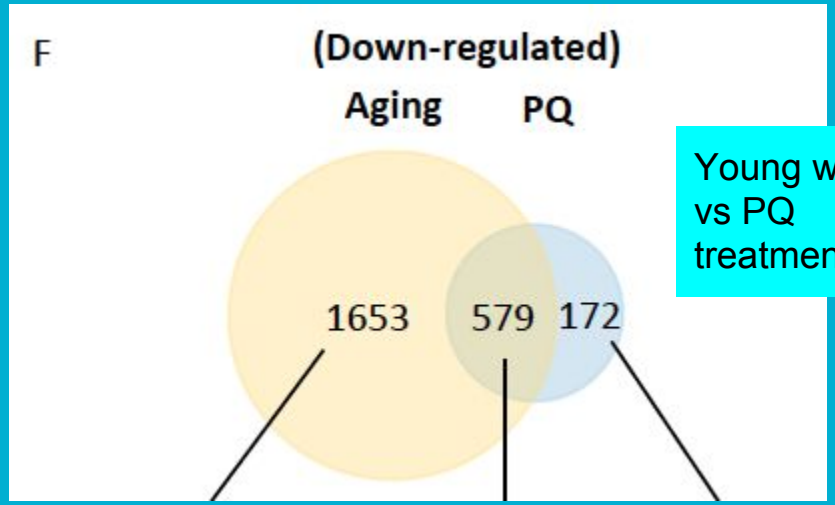
Original

Water treated young vs old



Reproduction

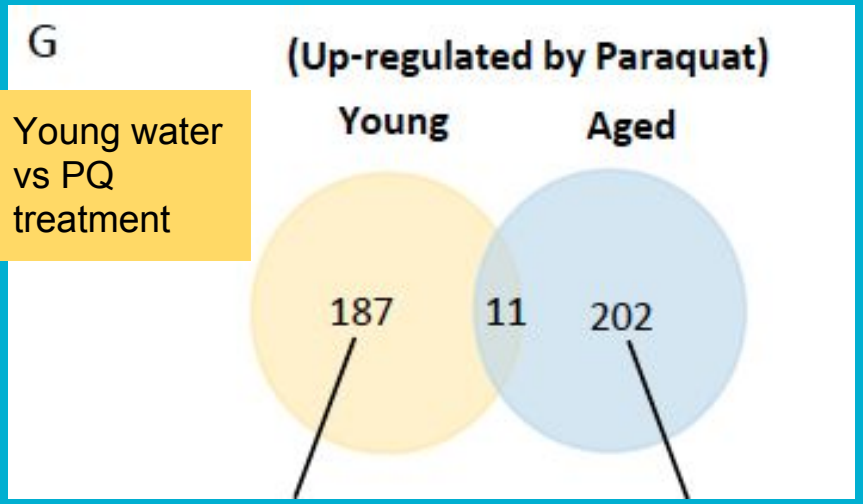




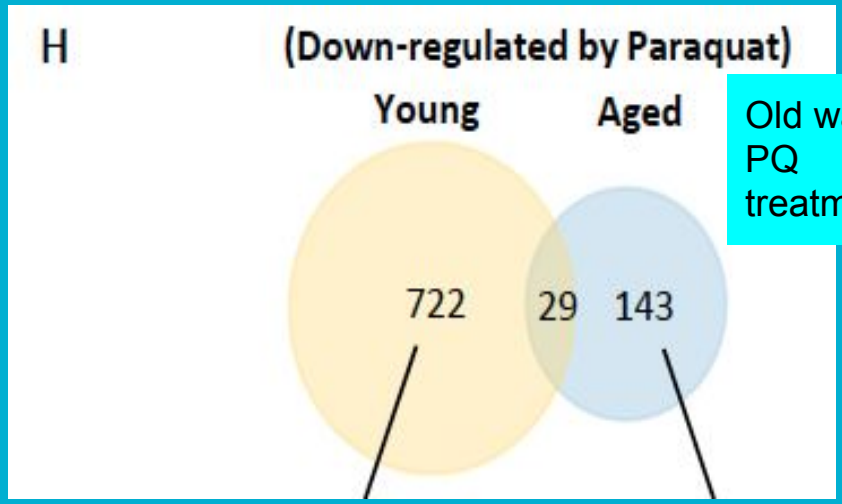
Young water vs PQ treatment

Differential Expression: venn diagrams

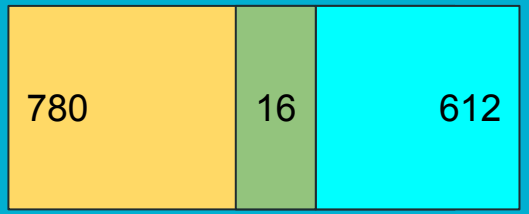
Original



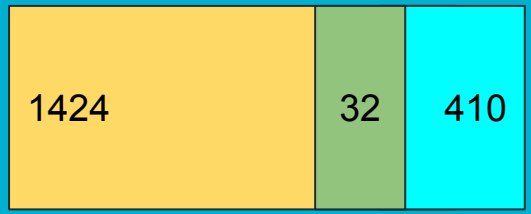
Young water vs PQ treatment



Old water vs PQ treatment



Reproduction



Up-Regulated Pathways

Original

A **Up-regulated in aged oenocytes**

NAME	ES	FDR
MISMATCH REPAIR	-0.65	0.025
DNA REPLICATION	-0.59	0.028
BASE EXCISION REPAIR	-0.61	0.035
NUCLEOTIDE EXCISION REPAIR	-0.53	0.037
FANCONI ANEMIA PATHWAY	-0.56	0.039
GLUTATHIONE S TRANSFERASE	-0.38	0.055

NAME	ES	FDR
MAPK SIGNALING PATHWAY	-0.43	0.028
ENDOCYTOSIS	-0.32	0.065
GLUTATHIONE METABOLISM	-0.32	0.062
DRUG METABOLISM 1	-0.32	0.065

Reproduction

Down-Regulated Pathways

Original

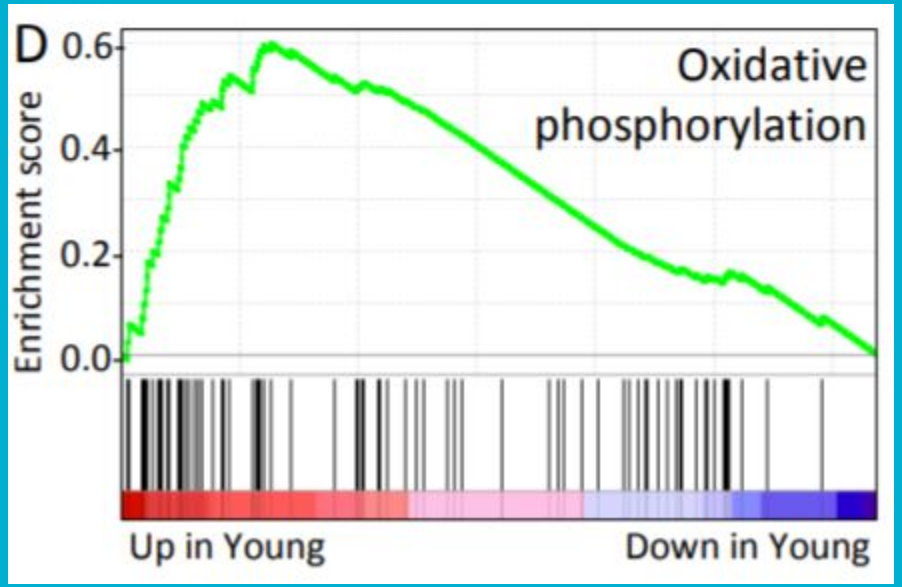
B **Down-regulated in aged oenocytes**

NAME	ES	FDR
OXIDATIVE PHOSPHORYLATION	0.60	0.001
RIBOSOME	0.57	0.001
PROTEASOME	0.65	0.002
CARBON METABOLISM	0.52	0.007
THIAMINE METABOLISM	0.70	0.008
PEROXISOME	0.53	0.008
PENTOSE PHOSPHATE	0.66	0.013
NEUROACTIVE LIGAND-RECEPTOR INTERACTION	0.55	0.015
GALACTOSE METABOLISM	0.59	0.016
GLYCOLYSIS	0.53	0.033
FATTY ACID METABOLISM	0.54	0.033
GLYOXYLATE METABOLISM	0.56	0.044
GLYCINE METABOLISM	0.56	0.045
FATTY ACID ELONGATION	0.63	0.049
CYTOCHROME P450	0.36	0.095

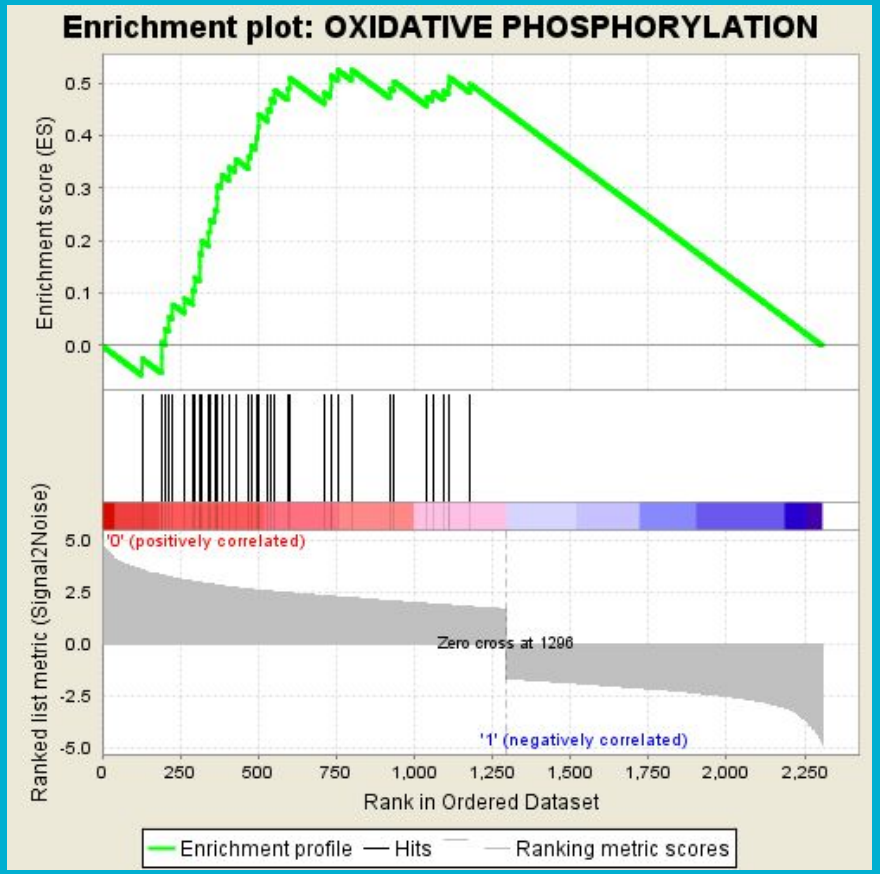
NAME	ES	FDR
OXIDATIVE PHOSPHORYLATION	0.53	0.000
PROTEASOME	0.61	0.000
CARBON METABOLISM	0.55	0.000
RIBOSOME	0.47	0.000
PEROXISOME	0.46	0.001
METABOLIC PATHWAYS	0.22	0.002
FATTY ACID METABOLISM	0.42	0.039

Reproduction

Enrichment Profiles

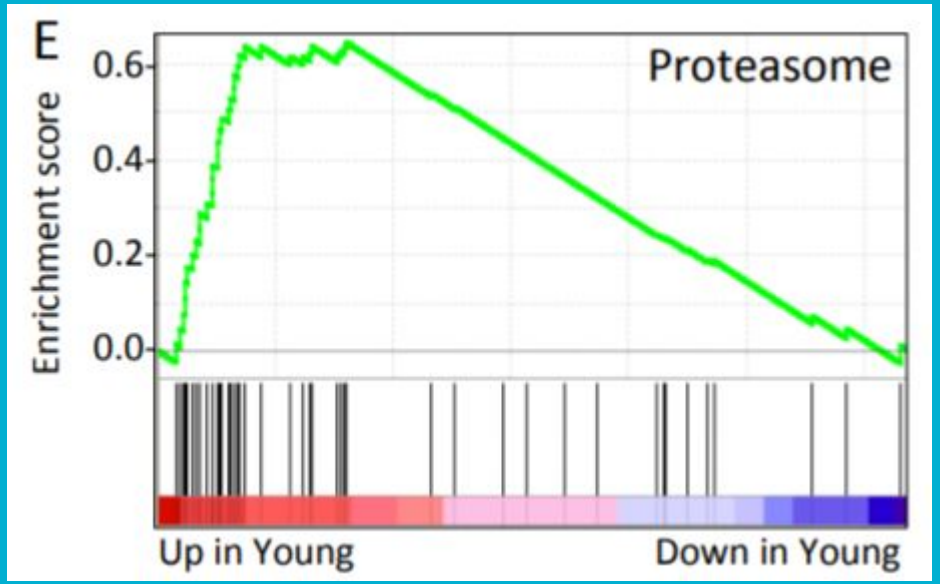


Original

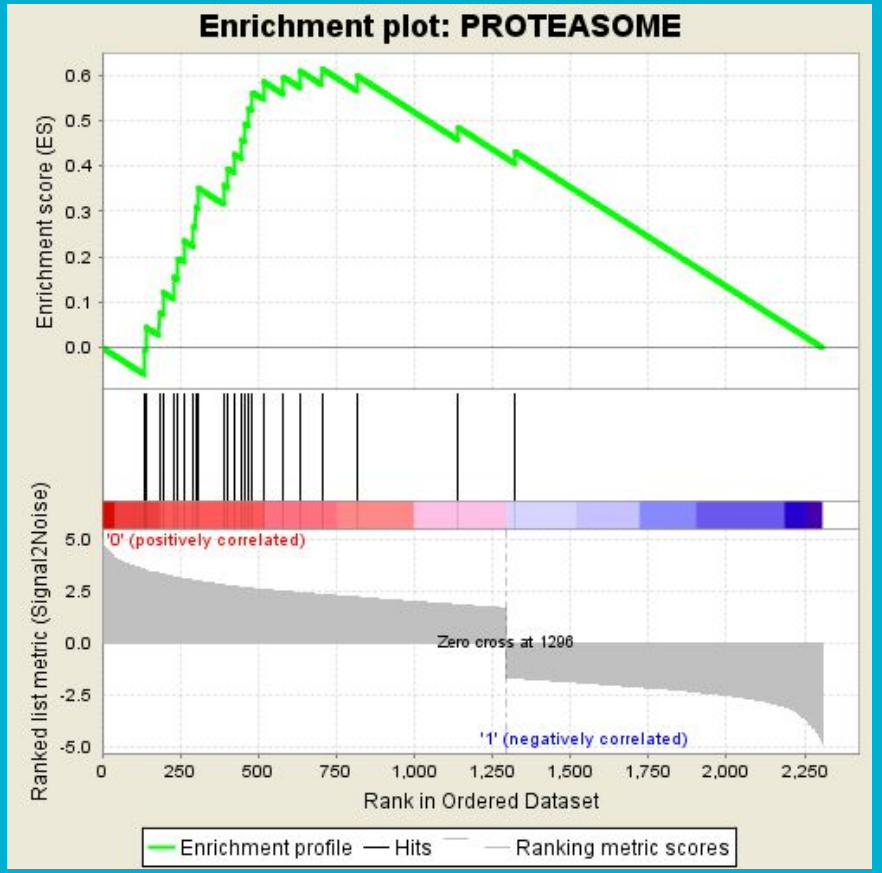


Reproduction

Enrichment Profiles

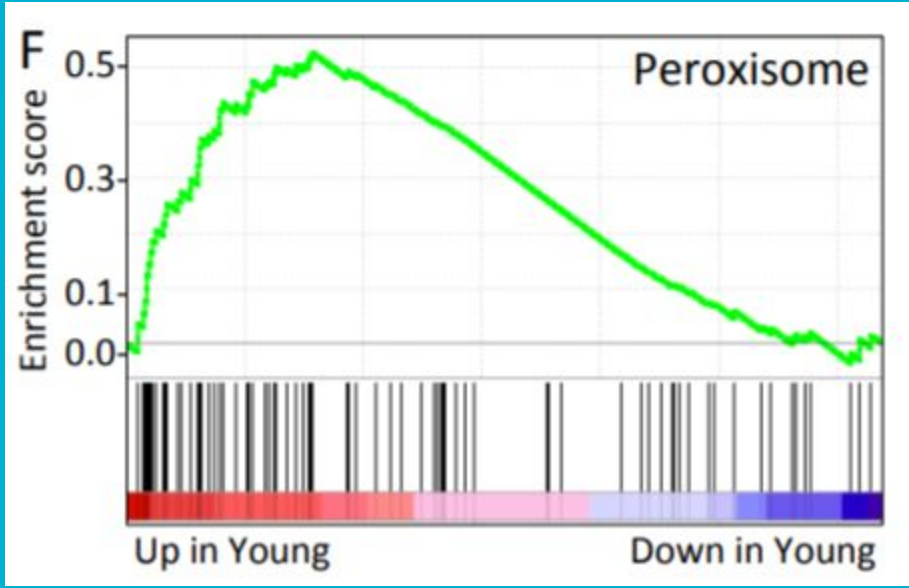


Original

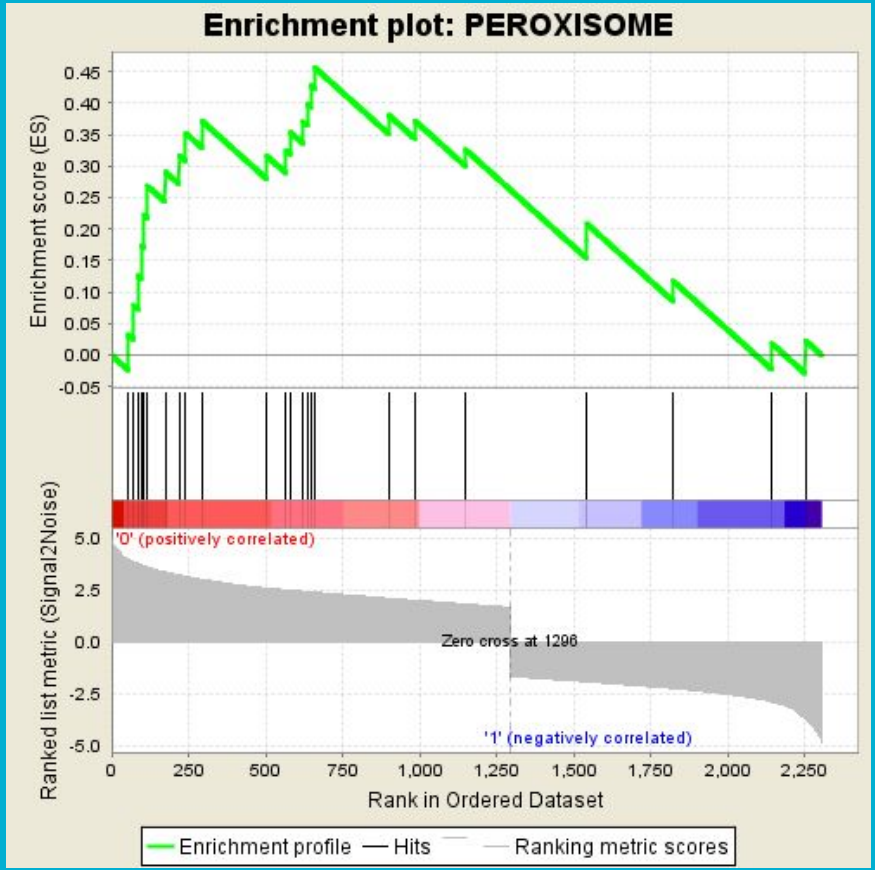


Reproduction

Enrichment Profiles

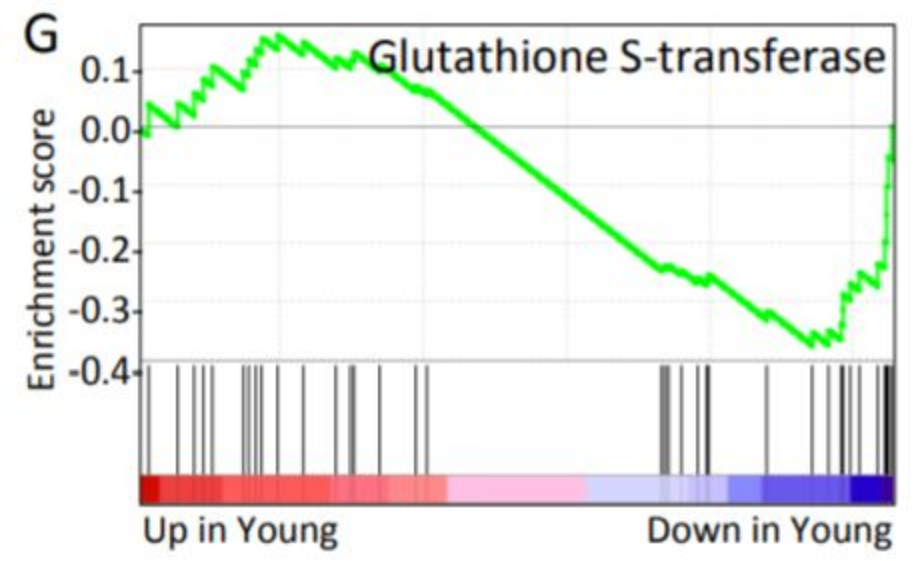


Original

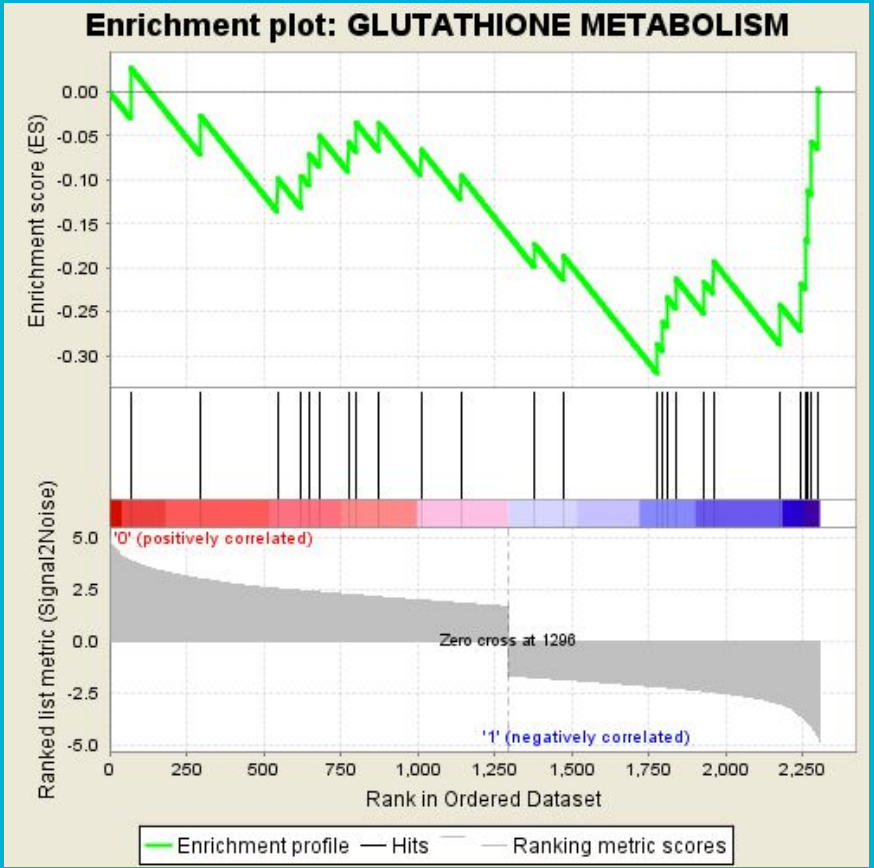


Reproduction

Enrichment Profiles



Original

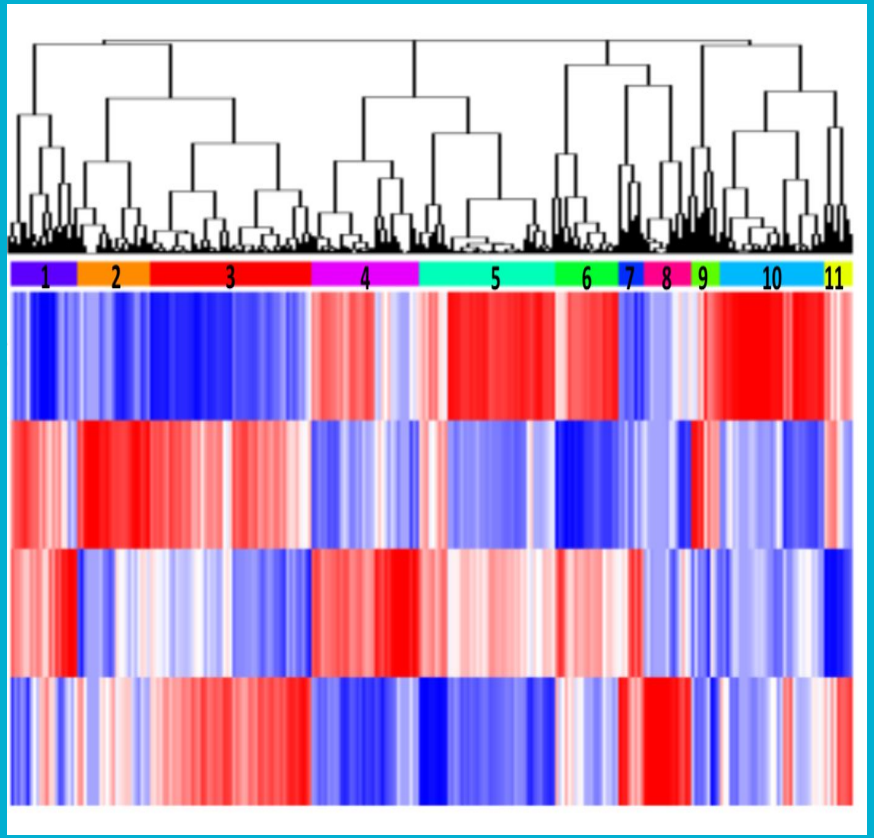


Reproduction

Observations

- ★ Enrichment profiles showed fewer matching genes (higher FDR values)
- ★ In GSEA, statistical significance is determined by comparing the observed gene set distribution to **random permutations** of the ranked gene list.
  - As a result, repeated analysis of the same data set in the GSEA program would yield slightly different results each time -- sometimes even reporting a different number of significant gene sets!

Heatmap Figures – fig.3I (different)



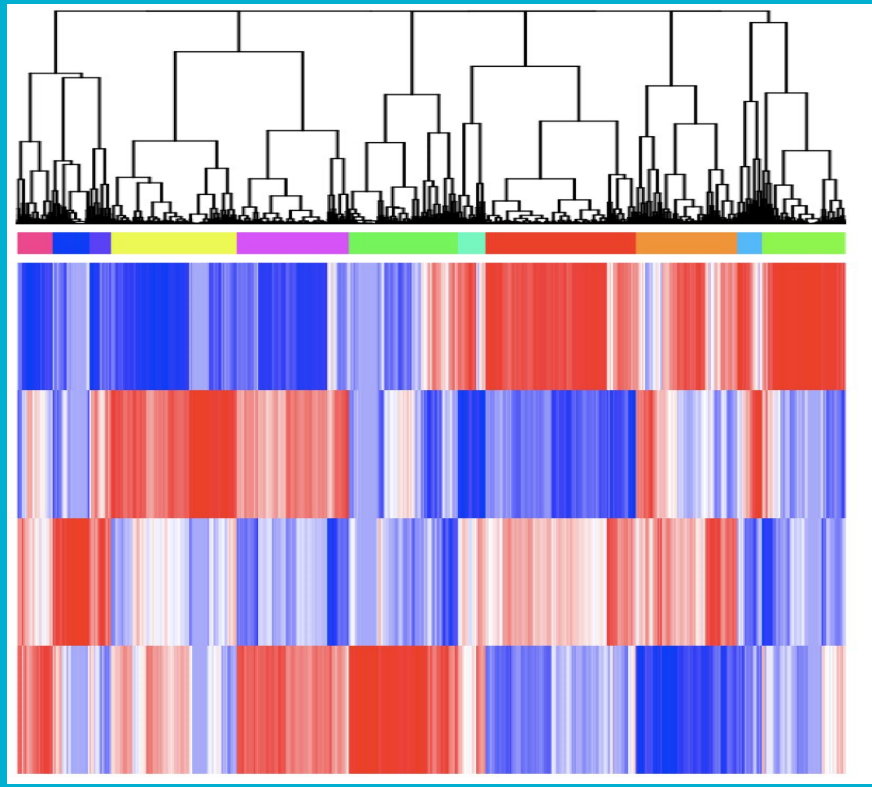
Original

H2O_Y

H2O_O

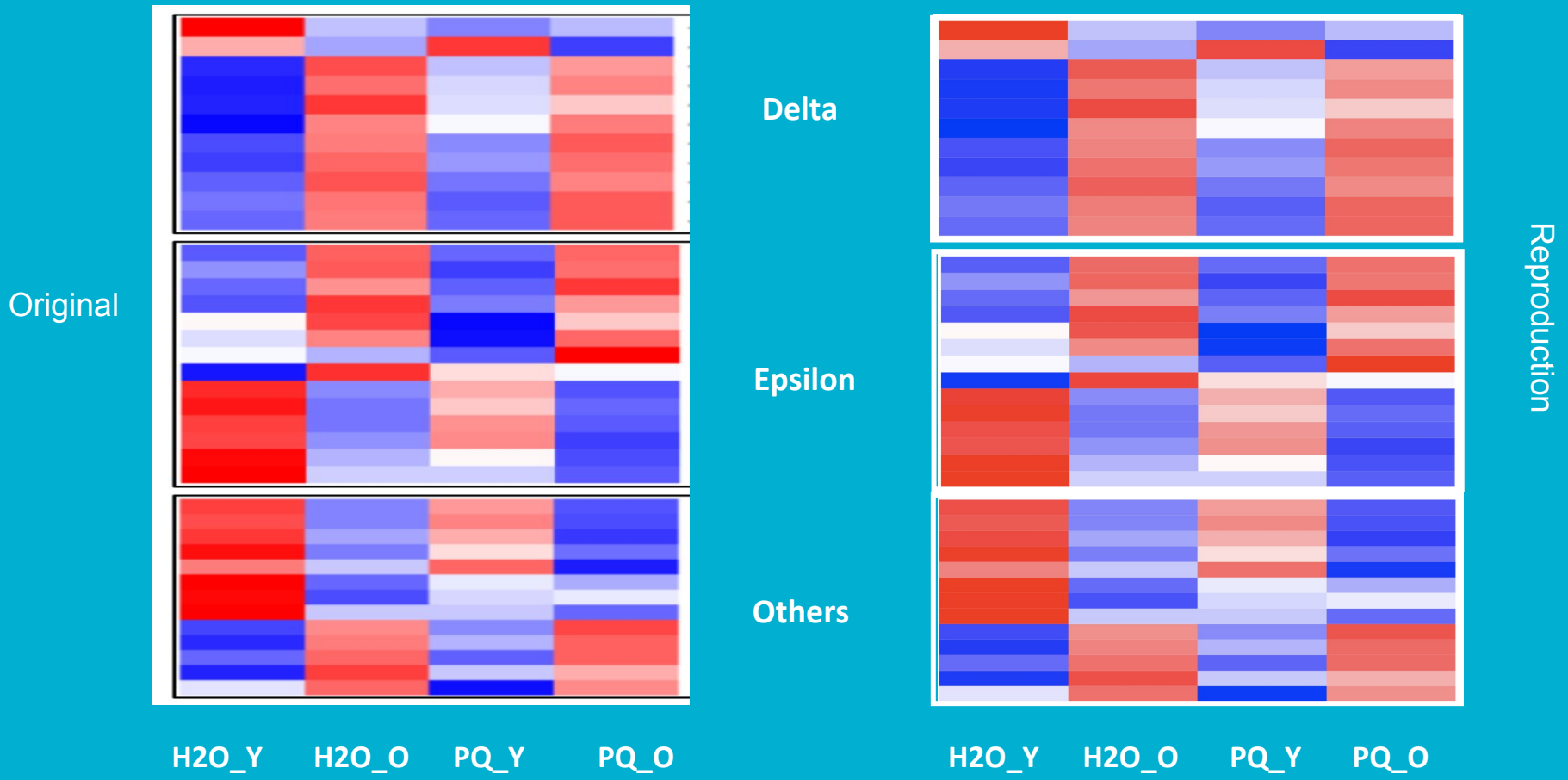
PQ_Y

PQ_O



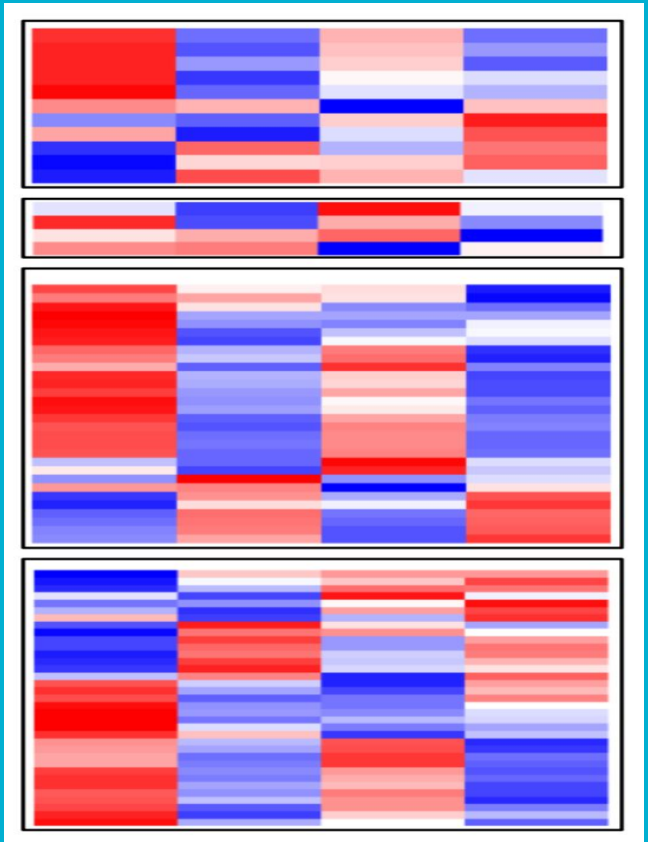
Reproduction

Heatmap Figures – fig.4J (same)



Heatmap Figures – fig.4K (different)

Original



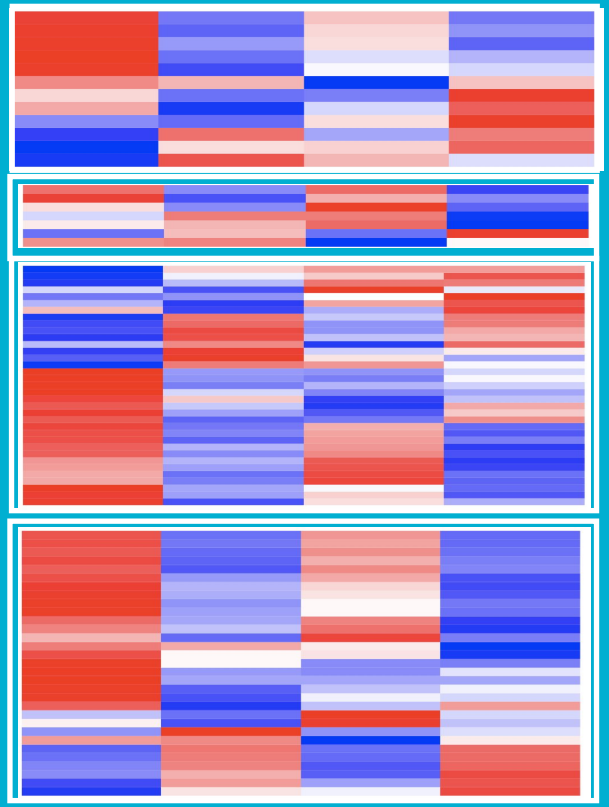
H2O_Y H2O_O PQ_Y PQ_O

Mito_Clan

Clan2

Clan3

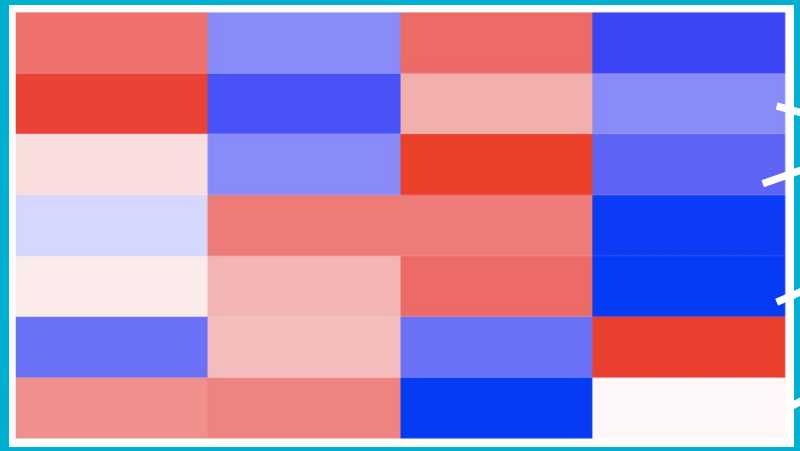
Clan4



H2O_Y H2O_O PQ_Y PQ_O

Reproduction

A closer look at Clan 2 subgroup



Reproduction

Original

H2O.Young	H2O.Old	PQ.Young	PQ.Old	X.	Mito_clan
35.919200	15.4485000	26.88680	17.562800	13	Clan 2 (7)
199.203000	95.1611000	204.04800	78.634300	17	Clan 2 (7)
0.802437	0.5188050	1.28753	0.448307	18	Clan 2 (7)
7.677150	7.8947800	3.70457	6.258480	16	Clan 2 (7)
1.486720	1.6592000	1.92506	0.809600	15	Clan 2 (7)
2.693440	3.4981700	3.49173	1.963300	14	Clan 2 (7)
0.000000	0.0179292	0.00000	0.031395	19	Clan 2 (7)

Potential reasons for inconsistency

★ Unknown criteria for filtering dataset

```
heat<-filter(forheat,ID_Symbol !="NA" & rowSums(forheat[3:6]) != 0) # filter data to get rid
of ID_Symbol without names and FPKM in 4 conditions all with value equal to 0
```

heat	14279 obs. of 6 variables

```
reads2 <- reads[1:13926, 1:5]
rownames(reads) <- make.names(reads$Flybase.ID, unique = TRUE)

#clean the first column and log2 transform
reads$Flybase.ID <- NULL
matrix <- data.matrix (reads) + 1
log_matrix <- log2(matrix)
matrixnew <- log_matrix[1:13925, 1:4]
```

Conclusions

- ★ Analysis pipeline consisted of well-established techniques and applications
- ★ Output of one step frequently did not match the input of the next
- ★ Very little detail provided on intermediate data transformations required between steps
- ★ Caused the results of our analysis to gradually diverge from the paper

Questions?