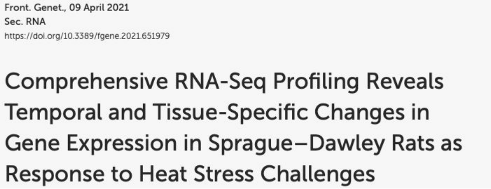
RNAseq species comparison of heat stress responses: orthologs, domains, GO terms, & GSEA

Hot Seq Summer

Jenn, Sarah, Kyle, Carin, Dmitry, Eric

2022-10-24

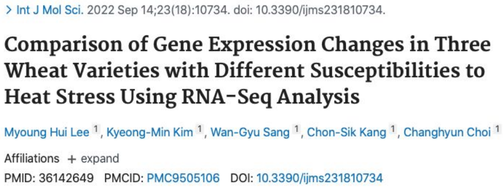
Development / tissue



Front. Genet., 09 April 2021
Sec. RNA
https://doi.org/10.3389/fgene.2021.651979

Comprehensive RNA-Seq Profiling Reveals Temporal and Tissue-Specific Changes in Gene Expression in Sprague–Dawley Rats as Response to Heat Stress Challenges

Single species adaptation



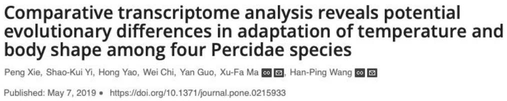
› Int J Mol Sci. 2022 Sep 14;23(18):10734. doi: 10.3390/ijms231810734.

Comparison of Gene Expression Changes in Three Wheat Varieties with Different Susceptibilities to Heat Stress Using RNA-Seq Analysis

Myoung Hui Lee [1], Kyeong-Min Kim [1], Wan-Gyu Sang [1], Chon-Sik Kang [1], Changhyun Choi [1]

Affiliations + expand
PMID: 36142649 PMCID: PMC9505106 DOI: 10.3390/ijms231810734

Related species adaptation



Comparative transcriptome analysis reveals potential evolutionary differences in adaptation of temperature and body shape among four Percidae species

Peng Xie, Shao-Kui Yi, Hong Yao, Wei Chi, Yan Guo, Xu-Fa Ma, Han-Ping Wang

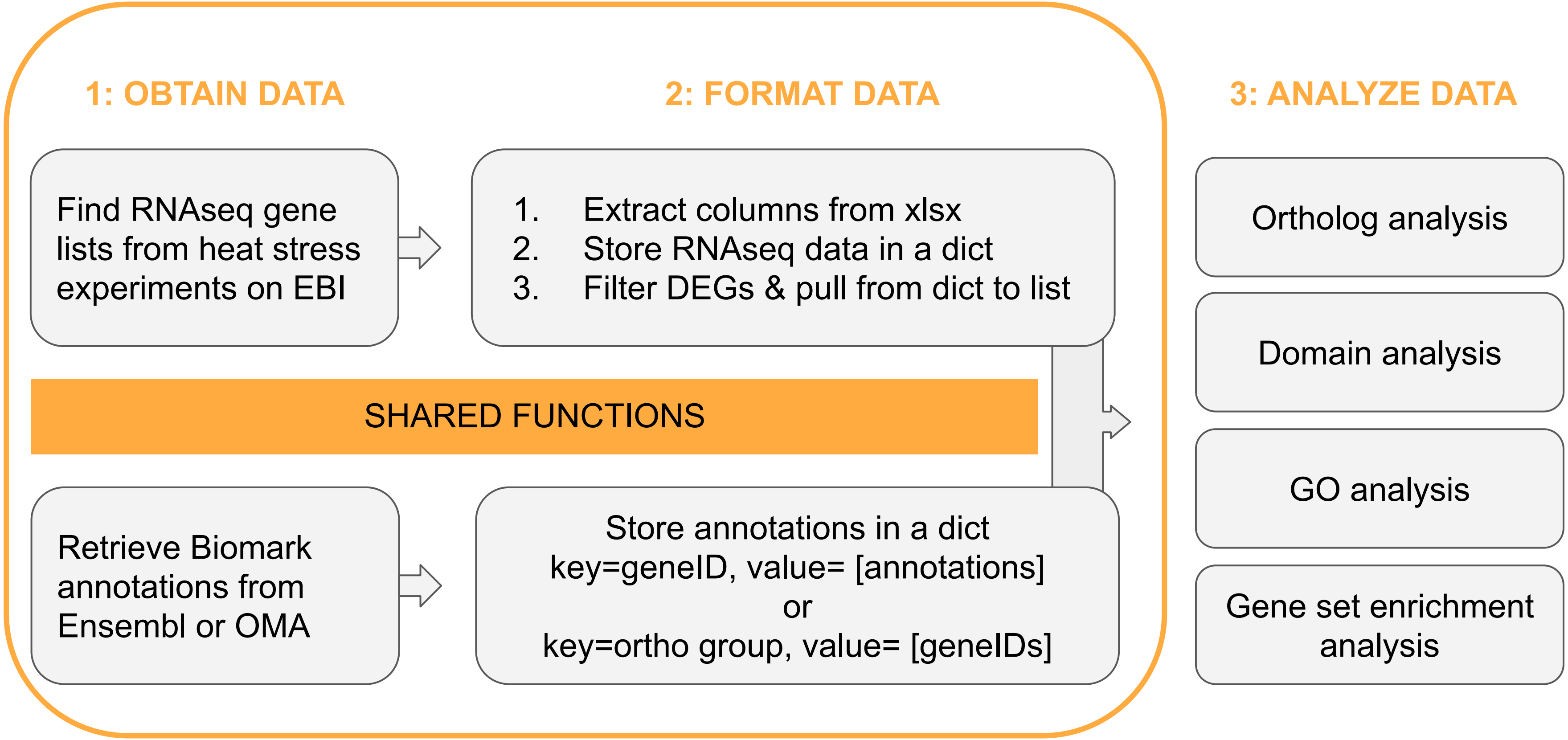
Published: May 7, 2019 • https://doi.org/10.1371/journal.pone.0215933

1-1 species/cell line comparison

Unrelated species comparison:

What gene functions & protein motifs are *shared* in heat stress responses across species?

Workflow to find shared heat response motifs



1: OBTAIN DATA

Find RNAseq gene lists from heat stress experiments on EBI

Retrieve Biomark annotations from Ensembl or OMA

2: FORMAT DATA

1. Extract columns from xlsx
2. Store RNAseq data in a dict
3. Filter DEGs & pull from dict to list

Store annotations in a dict
key=geneID, value= [annotations]
or
key=ortho group, value= [geneIDs]

SHARED FUNCTIONS

3: ANALYZE DATA

Ortholog analysis

Domain analysis

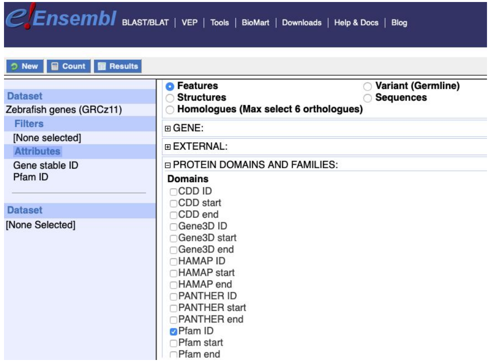
GO analysis

Gene set enrichment analysis

Shared functions for data retrieval and formatting

`species_annotation_df()`

Retrieves datasets from Ensemble using functions from pybiomart


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
New
Count
Results

Dataset
Zebrafish genes (GRCz11)
Filters
[None selected]
Attributes
Gene stable ID
Pfam ID

Dataset
[None Selected]

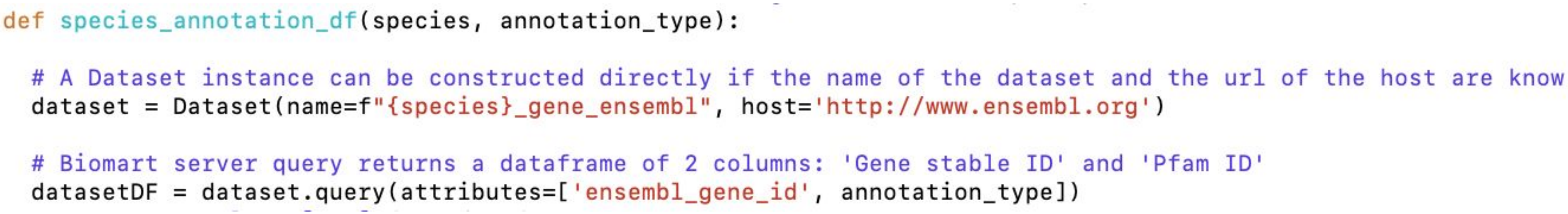
● Features
○ Structures
○ Homologues (Max select 6 orthologues)
○ Variant (Germline)
○ Sequences

⊞ GENE:

⊞ EXTERNAL:

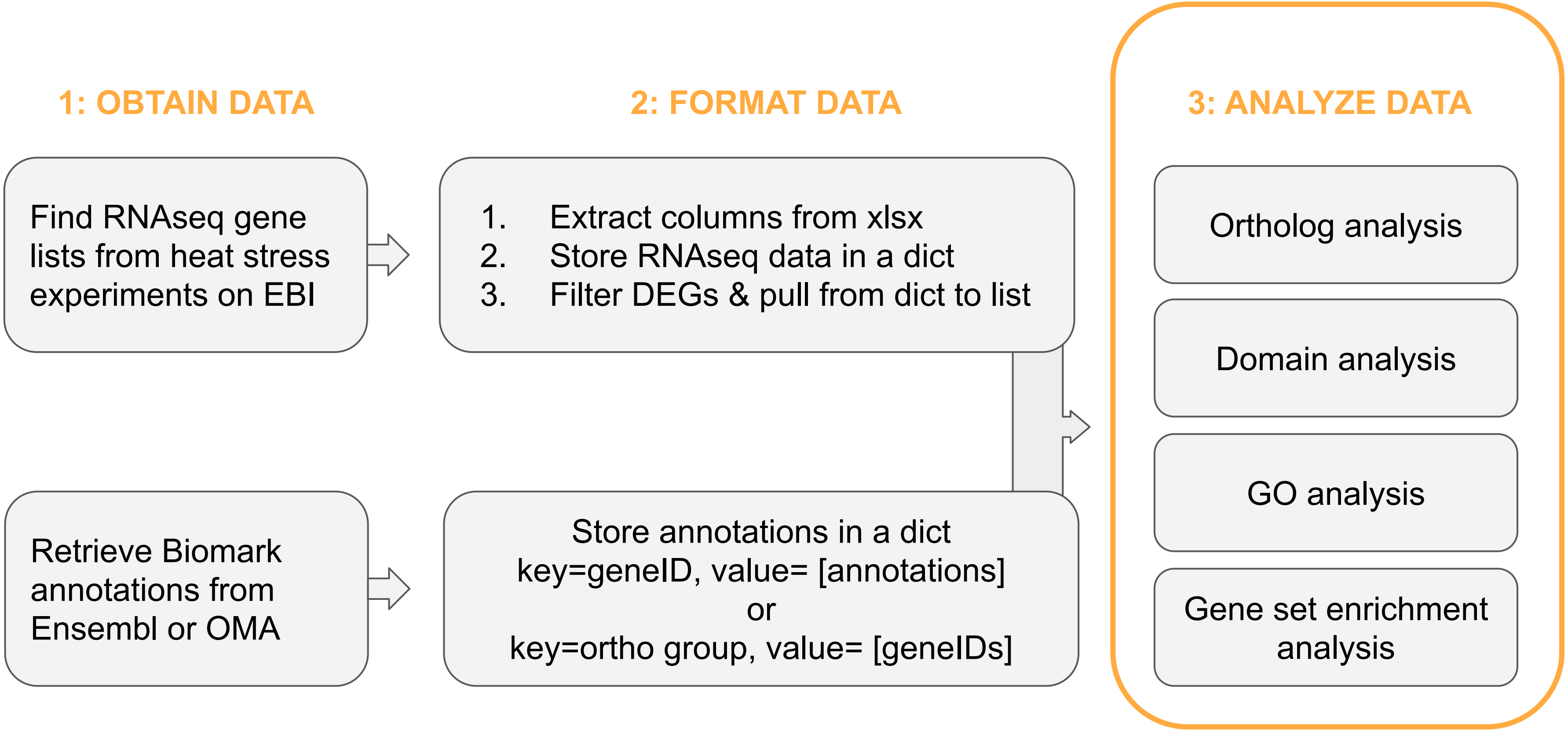
⊟ PROTEIN DOMAINS AND FAMILIES:

Domains
□ CDD ID
□ CDD start
□ CDD end
□ Gene3D ID
□ Gene3D start
□ Gene3D end
□ HAMAP ID
□ HAMAP start
□ HAMAP end
□ PANTHER ID
□ PANTHER start
□ PANTHER end
☑ Pfam ID
□ Pfam start
□ Pfam end



```
def species_annotation_df(species, annotation_type):

  # A Dataset instance can be constructed directly if the name of the dataset and the url of the host are know
  dataset = Dataset(name=f"{species}_gene_ensembl", host='http://www.ensembl.org')

  # Biomart server query returns a dataframe of 2 columns: 'Gene stable ID' and 'Pfam ID'
  datasetDF = dataset.query(attributes=['ensembl_gene_id', annotation_type])
```

Different analyses to find heat response motifs



1: OBTAIN DATA

Find RNAseq gene lists from heat stress experiments on EBI

Retrieve Biomark annotations from Ensembl or OMA

2: FORMAT DATA

1. Extract columns from xlsx
2. Store RNAseq data in a dict
3. Filter DEGs & pull from dict to list

Store annotations in a dict
key=geneID, value= [annotations]
or
key=ortho group, value= [geneIDs]

3: ANALYZE DATA

Ortholog analysis

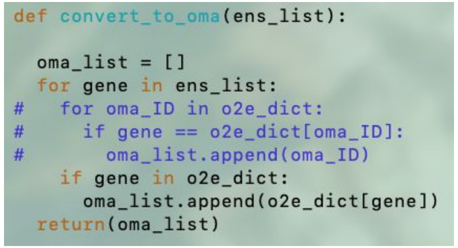
Domain analysis

GO analysis

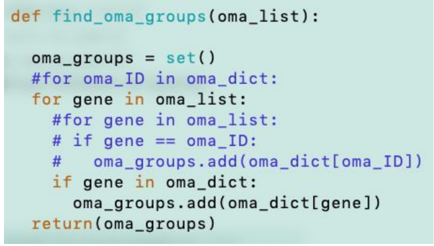
Gene set enrichment analysis

What orthologous groups are commonly up- or down-regulated after heat shock?

- Uses OMA (Orthologous Matrix) database
  - Oma database entered into dict
  - Oma_ID to Ensembl_ID list entered into dict
- General steps:
  - Pulls DEGs from 3 species into a dictionary (`deg_list()`)
  - Convert Ensembl_IDs to Oma_IDs (`convert_to_oma()`)
  - Finds each DEG's ortholog group (`find_oma_groups()`)
  - Gets the unique and shared ortholog groups between all species

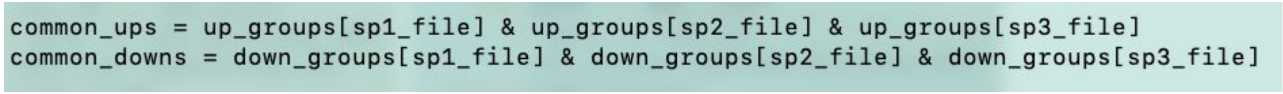


```
def convert_to_oma(ens_list):

  oma_list = []
  for gene in ens_list:
#   for oma_ID in o2e_dict:
#     if gene == o2e_dict[oma_ID]:
#       oma_list.append(oma_ID)
    if gene in o2e_dict:
      oma_list.append(o2e_dict[gene])
  return(oma_list)
```



```
def find_oma_groups(oma_list):

  oma_groups = set()
  #for oma_ID in oma_dict:
  for gene in oma_list:
    #for gene in oma_list:
    # if gene == oma_ID:
    #   oma_groups.add(oma_dict[oma_ID])
    if gene in oma_dict:
      oma_groups.add(oma_dict[gene])
  return(oma_groups)
```

Full script



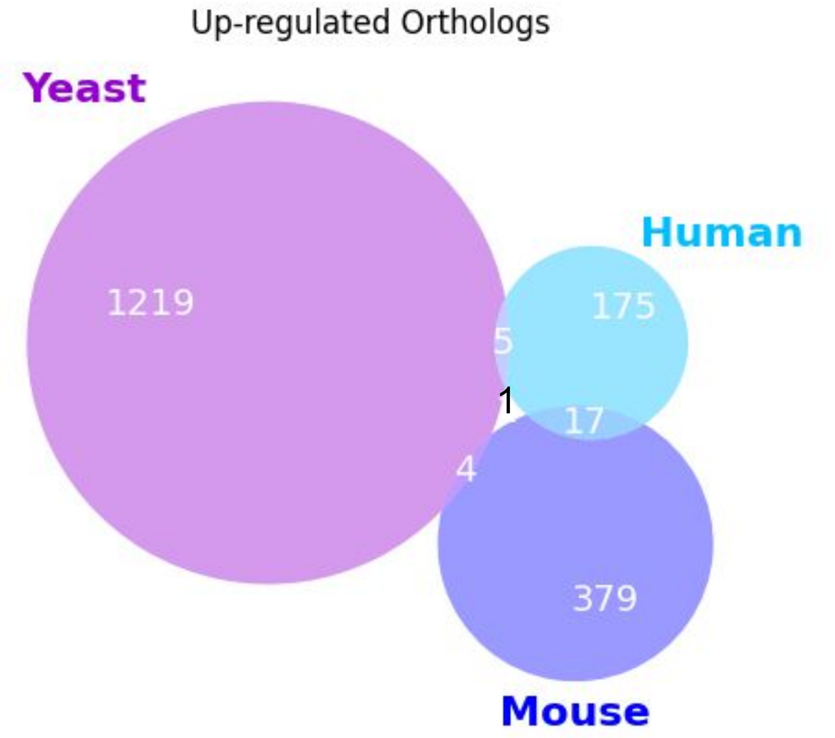
```
common_ups = up_groups[sp1_file] & up_groups[sp2_file] & up_groups[sp3_file]
common_downs = down_groups[sp1_file] & down_groups[sp2_file] & down_groups[sp3_file]
```

Orthology results

Common up-regulated oma group: 1105264

Heat shock protein (Yeast HSP10)

No common down-regulated oma groups



Full script

Domain analysis

Overarching questions: in the differentially expressed genes for all three species:

(1) which functional domains are particularly enriched across DEGs?

(2) which enriched domains are shared between all three species?

Domain analysis

Which functional domains are particularly enriched across DEGs?

- enriched_domains.py
  - Inputs from command line: *diff_direction* (“up”/”down” regulated genes), *species*
  - Steps:
    - Makes (1) DEG dictionary {ensemble_ID : [logFC, pvalue]} for one species and (2) an annotation dictionary {ensembl_ID : pfam_ID}
    - Use the two dictionaries to make a list of all domains present in DEGs
    - Use hypergeometric probability calculation to determine which functional domains are overrepresented (above chance) in the DEGs.
  - Outputs tab delimited .txt with two columns: pfam_ID, probability

script on github

$$\Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

When you pull K marbles from a bowl of N marbles, what is the probability of pulling exactly k green marbles when there are n green marbles in the bowl?

Domain analysis

Which enriched domains are shared between all three species?

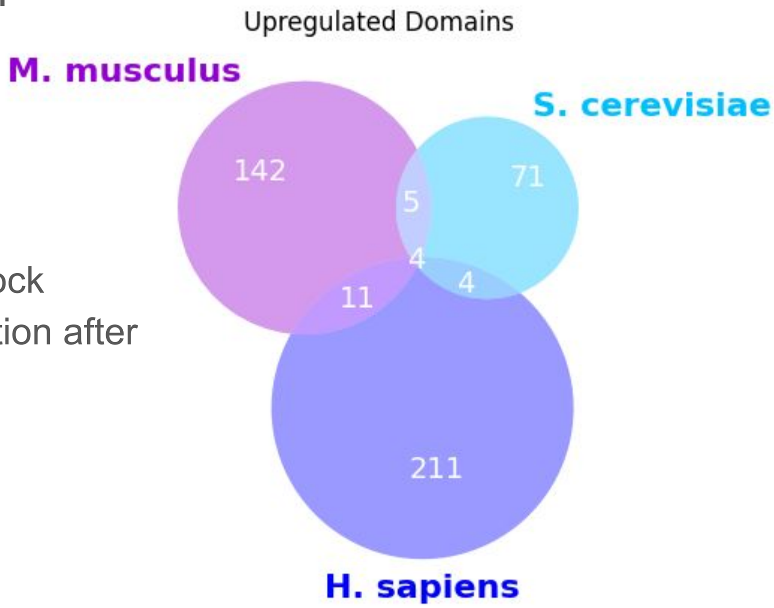
- domain_comp_3.py
  - Inputs from command line: *3 files output from enriched_domains.py (pfam_id \t probability)*
  - Steps:
    - Makes (1) enriched domain list for each species and (2) a pfam dictionary {pfam_ID : domain_description}
    - Finds which domains are unique to each list or shared between all three lists
    - For the shared domains, look up the descriptions for the pfam_ids
    - To make venn diagram: parse file names to get species, diff_direction
  - Outputs:
    - Prints to standard out: (1) stats on number of unique, shared domains, (2) list of domain descriptions for shared domains
    - Venn diagram .png

script on github

Domain results

All species share **4 enriched domains** in their heat-stress induced upregulated genes:

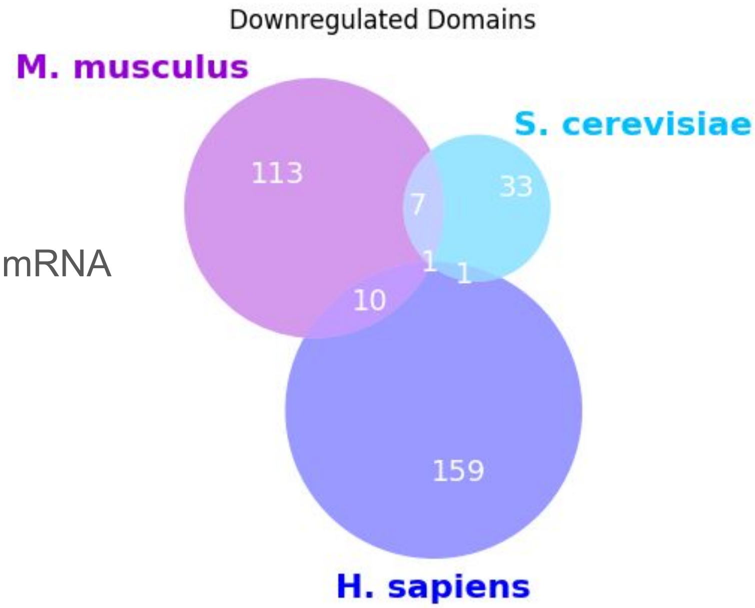
- Activator of Hsp90 ATPase, N-terminal
- DnaJ C terminal domain
  - Chaperone associated with the Hsp70 heat-shock system involved in protein folding and renaturation after stress
- Hsp70 protein
- bZIP transcription factor
  - Found in many eukaryotic transcription factors



Domain results

All species share **1 enriched domain** in their heat-stress induced downregulated genes:

- WD domain, G-beta repeat
  - Highly conserved (present in all eukaryotes)
  - Regulate cellular functions: gene transcription, mRNA modification, etc.



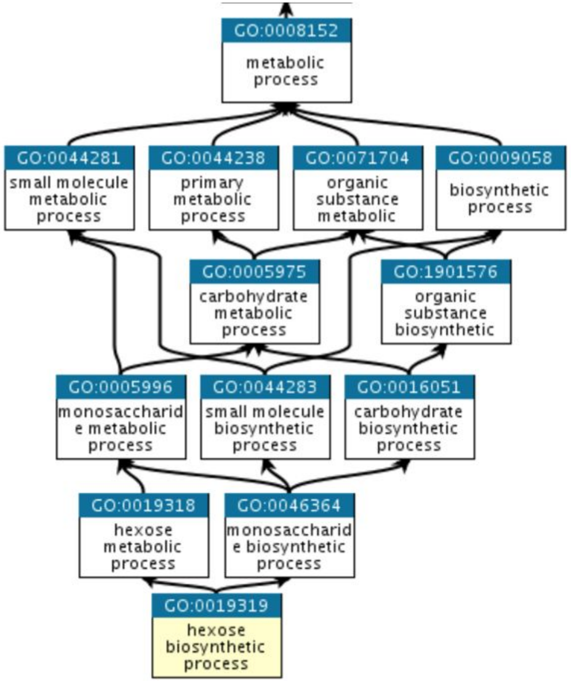
Gene Ontology (GO) enrichment analysis

GO enrichment analysis allows you to retrieve a ***functional profile*** of a gene set in order to better understand underlying biological processes.

The Gene Ontology (GO) provides a system for **hierarchically classifying genes** or gene products into ***terms*** organized in a graph structure (or an ontology) =>

GO terms are grouped into 3 categories: **biological processes**, **cellular locations** and **molecular functions**.

Each gene can be described (annotated) with multiple terms.

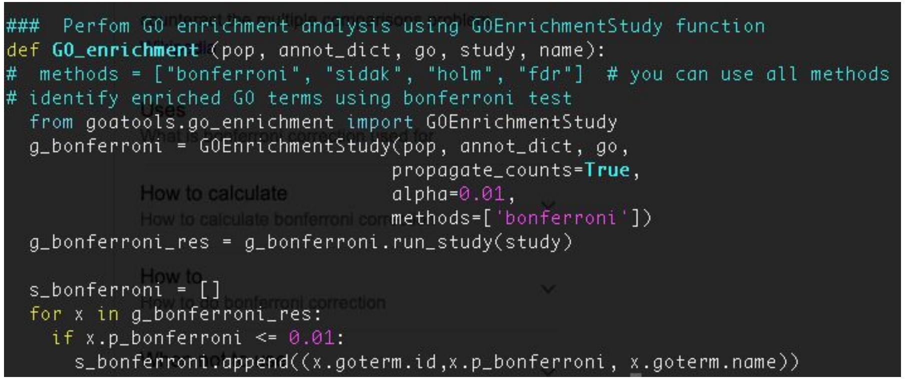

GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

GO enrichment analysis workflow

Input: 1) DEG with Log2FC, p-value and Gene ID’s (txt) >>> list of up- and down-regulated genes (**rnaseqs_to_dict**, **deg_list**)

2) Annotation file with gene ID’s of all genes and associated GO terms (From BioMart) (txt) (**read_annot**)

3) OBO file containing the information about ontology (txt) (**import_OBO**)



```
###  Perfom GO enrichment analysis using GOEnrichmentStudy function
def GO_enrichment (pop, annot_dict, go, study, name):
#  methods = ["bonferroni", "sidak", "holm", "fdr"]  # you can use all methods
# identify enriched GO terms using bonferroni test
  from goatools.go_enrichment import GOEnrichmentStudy
  g_bonferroni = GOEnrichmentStudy(pop, annot_dict, go,
                                  propagate_counts=True,
                                  alpha=0.01,
                                  methods=['bonferroni'])
  g_bonferroni_res = g_bonferroni.run_study(study)

  s_bonferroni = []
  for x in g_bonferroni_res:
    if x.p_bonferroni <= 0.01:
      s_bonferroni.append((x.goterm.id,x.p_bonferroni, x.goterm.name))
```

pop install **goatools** package

Output: 1) A dataframe containing enriched GO terms, p-values, description of processes.

2) A Venn diagram with overlapping GO terms enriched in all 3 species (**make_venn_diagram**).

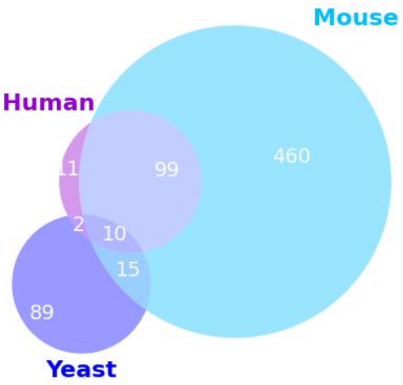
GO results

Goal: To identify commonly enriched GO terms in up-regulated and down-regulated genes in human cell lines, mouse and yeast under heat shock stress.

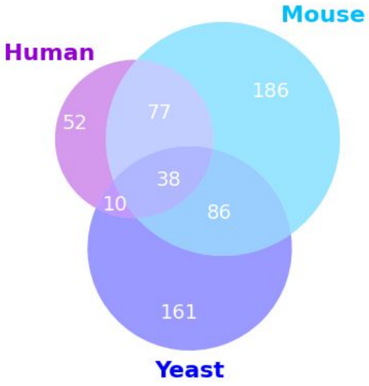
Commonly up-regulated terms: response to chemical cellular response to chemical stimulus, nitrogen compound metabolic process, macromolecule biosynthetic process, macromolecule metabolic process, nucleoplasm, cellular anatomical entity, ribonucleoprotein complex assembly

Commonly down-regulated terms: regulation of cellular metabolic process, regulation of primary metabolic process, regulation of nitrogen compound metabolic process, metabolic process, primary metabolic process, positive regulation of cellular process, intracellular membrane-bounded organelle, membrane-bounded organelle

Commonly up-regulated GO terms



Commonly down-regulated GO terms

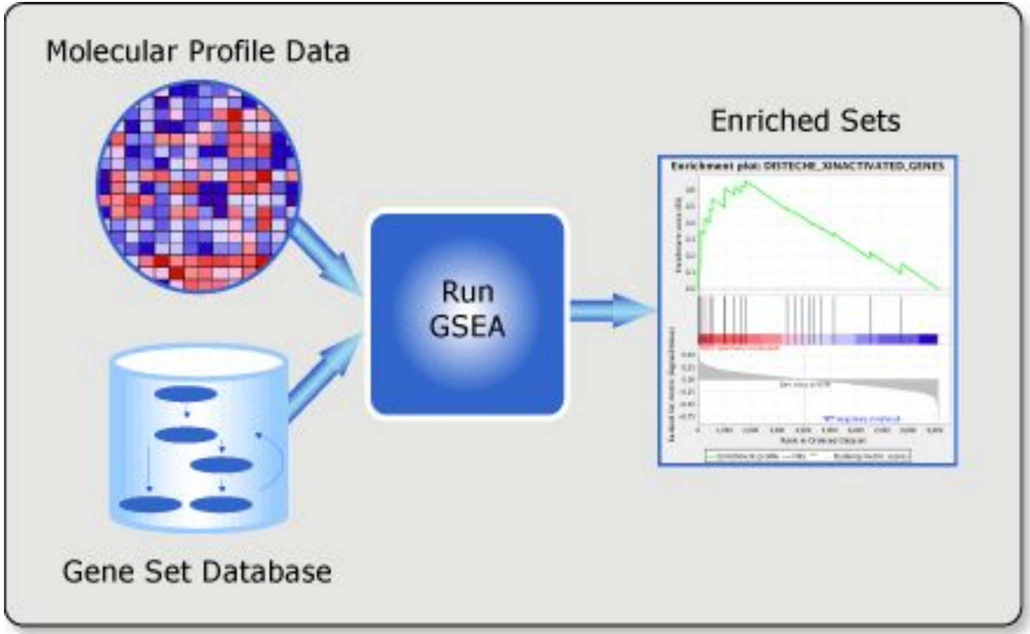


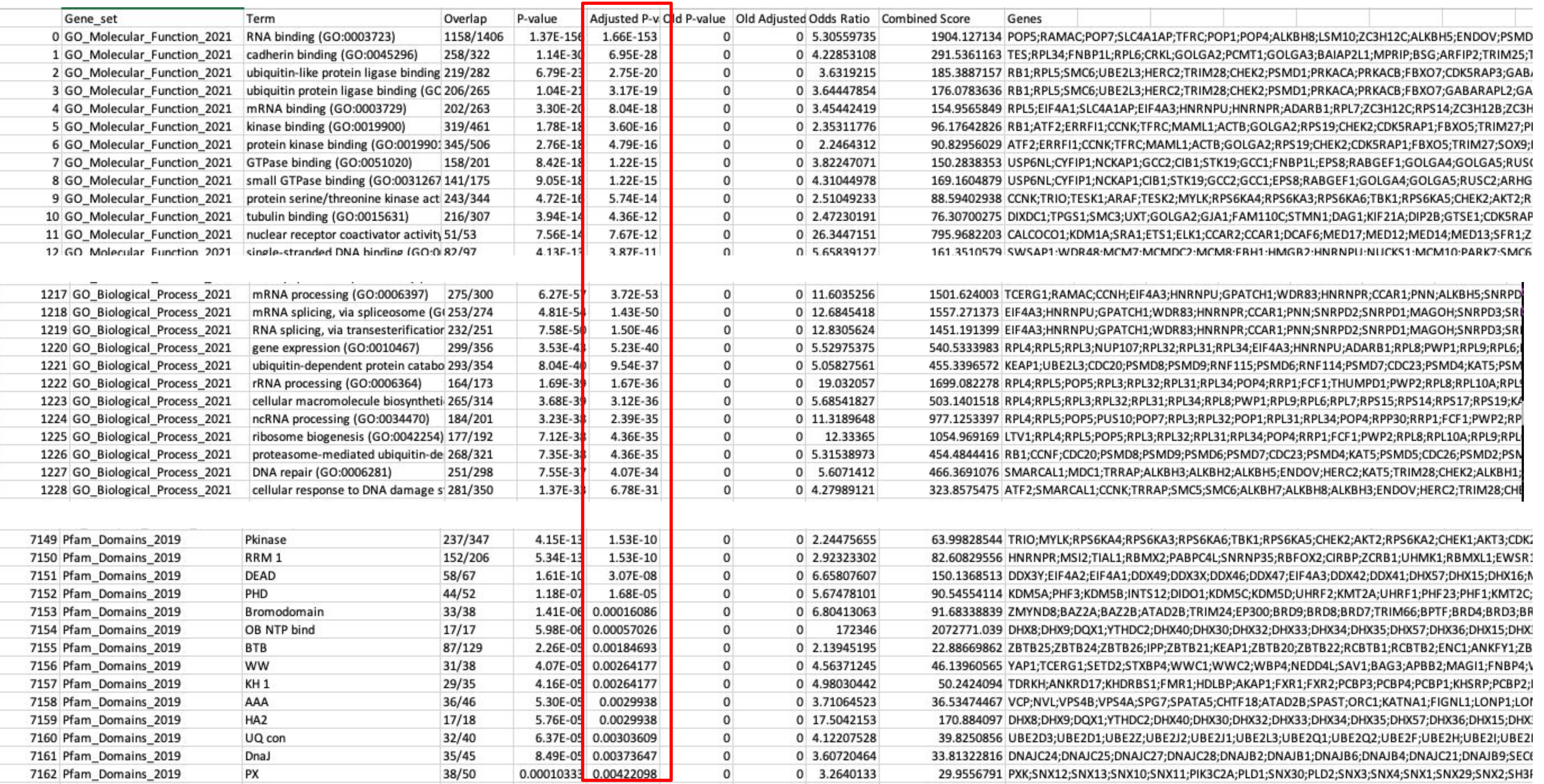
Gene Set Enrichment Analysis (GSEA)

GSEA is a computational method that determines whether a ranked set of genes shows statistically significant, concordant differences between two biological states (heat shock)

Why is it used?

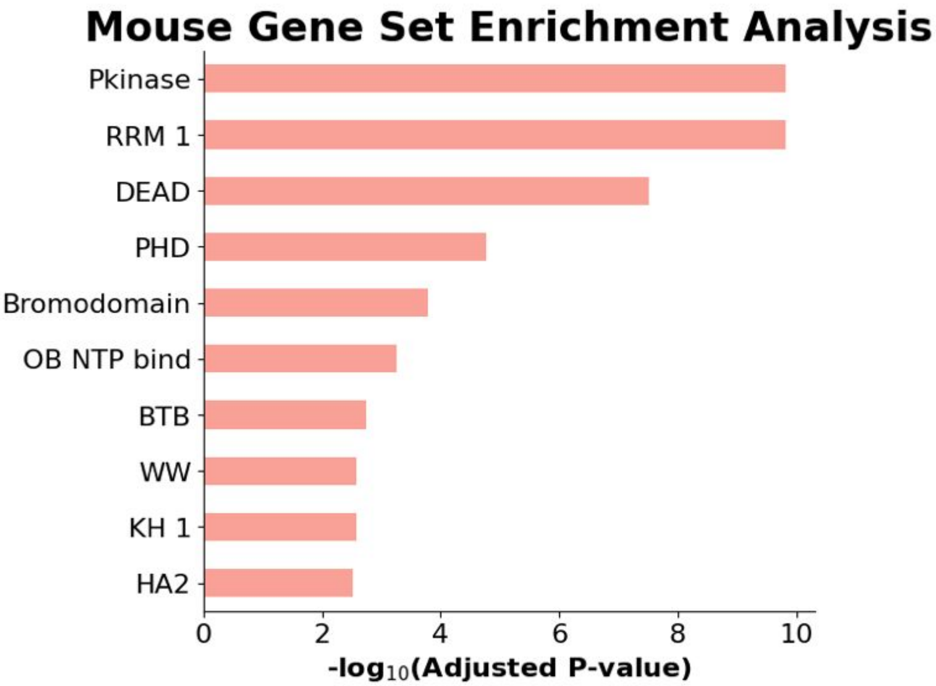
GSEA does not require a pvalue or log2 FC cutoff - GSEA uses all genes and ranks them between between groups based on fold change



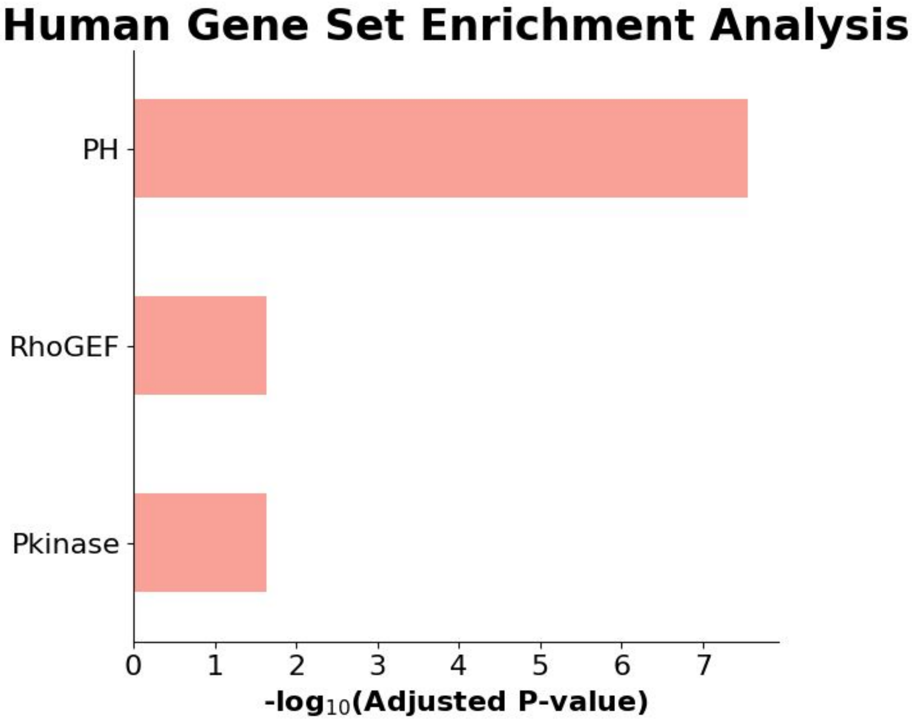


	Gene_set	Term	Overlap	P-value	Adjusted P-v	Old P-value	Old Adjusted	Odds Ratio	Combined Score	Genes
0	GO_Molecular_Function_2021	RNA binding (GO:0003723)	1158/1406	1.37E-156	1.66E-153	0	0	5.30559735	1904.127134	POP5;RAMAC;POP7;SLC4A1AP;TFRC;POP1;POP4;ALKBH8;LSM10;ZC3H12C;ALKBH5;ENDOV;PSMD
1	GO_Molecular_Function_2021	cadherin binding (GO:0045296)	258/322	1.14E-30	6.95E-28	0	0	4.22853108	291.5361163	TES;RPL34;FNBP1L;RPL6;CRKL;GOLGA2;PCMT1;GOLGA3;BAIAP2L1;MPRIP;BSG;ARFIP2;TRIM25;T
2	GO_Molecular_Function_2021	ubiquitin-like protein ligase binding	219/282	6.79E-23	2.75E-20	0	0	3.6319215	185.3887157	RB1;RPL5;SMC6;UBE2L3;HERC2;TRIM28;CHEK2;PSMD1;PRKACA;PRKACB;FBXO7;CDK5RAP3;GABA
3	GO_Molecular_Function_2021	ubiquitin protein ligase binding (GC	206/265	1.04E-21	3.17E-19	0	0	3.64447854	176.0783636	RB1;RPL5;SMC6;UBE2L3;HERC2;TRIM28;CHEK2;PSMD1;PRKACA;PRKACB;FBXO7;GABARAPL2;GA
4	GO_Molecular_Function_2021	mRNA binding (GO:0003729)	202/263	3.30E-20	8.04E-18	0	0	3.45442419	154.9565849	RPL5;EIF4A1;SLC4A1AP;EIF4A3;HNRNPU;HNRNPR;ADARB1;RPL7;ZC3H12C;RPS14;ZC3H12B;ZC3H
5	GO_Molecular_Function_2021	kinase binding (GO:0019900)	319/461	1.78E-18	3.60E-16	0	0	2.35311776	96.17642826	RB1;ATF2;ERRFI1;CCNK;TFRC;MAML1;ACTB;GOLGA2;RPS19;CHEK2;CDK5RAP1;FBXO5;TRIM27;PI
6	GO_Molecular_Function_2021	protein kinase binding (GO:001990	345/506	2.76E-18	4.79E-16	0	0	2.2464312	90.82956029	ATF2;ERRFI1;CCNK;TFRC;MAML1;ACTB;GOLGA2;RPS19;CHEK2;CDK5RAP1;FBXO5;TRIM27;SOX9;
7	GO_Molecular_Function_2021	GTPase binding (GO:0051020)	158/201	8.42E-18	1.22E-15	0	0	3.82247071	150.2838353	USP6NL;CYFIP1;NCKAP1;GCC2;CIB1;STK19;GCC1;FNBP1L;EPS8;RABGEF1;GOLGA4;GOLGA5;RUSC
8	GO_Molecular_Function_2021	small GTPase binding (GO:0031267	141/175	9.05E-18	1.22E-15	0	0	4.31044978	169.1604879	USP6NL;CYFIP1;NCKAP1;CIB1;STK19;GCC2;GCC1;EPS8;RABGEF1;GOLGA4;GOLGA5;RUSC2;ARHG
9	GO_Molecular_Function_2021	protein serine/threonine kinase act	243/344	4.72E-16	5.74E-14	0	0	2.51049233	88.59402938	CCNK;TRIO;TESK1;ARAF;TESK2;MYLK;RPS6KA4;RPS6KA3;RPS6KA6;TBK1;RPS6KA5;CHEK2;AKT2;R
10	GO_Molecular_Function_2021	tubulin binding (GO:0015631)	216/307	3.94E-14	4.36E-12	0	0	2.47230191	76.30700275	DIXDC1;TPGS1;SMC3;UXT;GOLGA2;GJA1;FAM110C;STMN1;DAG1;KIF21A;DIP2B;GTSE1;CDK5RAP
11	GO_Molecular_Function_2021	nuclear receptor coactivator activity	51/53	7.56E-14	7.67E-12	0	0	26.3447151	795.9682203	CALCOCO1;KDM1A;SRA1;ETS1;ELK1;CCAR2;CCAR1;DCAF6;MED17;MED12;MED14;MED13;SFR1;Z
12	GO_Molecular_Function_2021	single-stranded DNA binding (GO:0	82/97	4.13E-13	3.87E-11	0	0	5.65839127	161.3510579	SWSAP1;WDR48;MCM7;MCMDC2;MCM8;FBH1;HMGB2;HNRNPU;NUCKS1;MCM10;PARK7;SMC6
1217	GO_Biological_Process_2021	mRNA processing (GO:0006397)	275/300	6.27E-57	3.72E-53	0	0	11.6035256	1501.624003	TCERG1;RAMAC;CCNH;EIF4A3;HNRNPU;GPATCH1;WDR83;HNRNPR;CCAR1;PNN;ALKBH5;SNRPD
1218	GO_Biological_Process_2021	mRNA splicing, via spliceosome (G	253/274	4.81E-54	1.43E-50	0	0	12.6845418	1557.271373	EIF4A3;HNRNPU;GPATCH1;WDR83;HNRNPR;CCAR1;PNN;SNRPD2;SNRPD1;MAGOH;SNRPD3;SR
1219	GO_Biological_Process_2021	RNA splicing, via transesterificatior	232/251	7.58E-50	1.50E-46	0	0	12.8305624	1451.191399	EIF4A3;HNRNPU;GPATCH1;WDR83;HNRNPR;CCAR1;PNN;SNRPD2;SNRPD1;MAGOH;SNRPD3;SR
1220	GO_Biological_Process_2021	gene expression (GO:0010467)	299/356	3.53E-43	5.23E-40	0	0	5.52975375	540.5333983	RPL4;RPL5;RPL3;NUP107;RPL32;RPL31;RPL34;EIF4A3;HNRNPU;ADARB1;RPL8;PWP1;RPL9;RPL6;
1221	GO_Biological_Process_2021	ubiquitin-dependent protein catabo	293/354	8.04E-40	9.54E-37	0	0	5.05827561	455.3396572	KEAP1;UBE2L3;CDC20;PSMD8;PSMD9;RNF115;PSMD6;RNF114;PSMD7;CDC23;PSMD4;KAT5;PSM
1222	GO_Biological_Process_2021	rRNA processing (GO:0006364)	164/173	1.69E-39	1.67E-36	0	0	19.032057	1699.082278	RPL4;RPL5;POP5;RPL3;RPL32;RPL31;RPL34;POP4;RRP1;FCF1;THUMPD1;PWP2;RPL8;RPL10A;RPL
1223	GO_Biological_Process_2021	cellular macromolecule biosyntheti	265/314	3.68E-39	3.12E-36	0	0	5.68541827	503.1401518	RPL4;RPL5;RPL3;RPL32;RPL31;RPL34;RPL8;PWP1;RPL9;RPL6;RPL7;RPS15;RPS14;RPS17;RPS19;KA
1224	GO_Biological_Process_2021	ncRNA processing (GO:0034470)	184/201	3.23E-38	2.39E-35	0	0	11.3189648	977.1253397	RPL4;RPL5;POP5;PUS10;POP7;RPL3;RPL32;POP1;RPL31;RPL34;POP4;RPP30;RRP1;FCF1;PWP2;RP
1225	GO_Biological_Process_2021	ribosome biogenesis (GO:0042254)	177/192	7.12E-38	4.36E-35	0	0	12.33365	1054.969169	LTV1;RPL4;RPL5;POP5;RPL3;RPL32;RPL31;RPL34;POP4;RRP1;FCF1;PWP2;RPL8;RPL10A;RPL9;RPL
1226	GO_Biological_Process_2021	proteasome-mediated ubiquitin-de	268/321	7.35E-38	4.36E-35	0	0	5.31538973	454.4844416	RB1;CCNF;CDC20;PSMD8;PSMD9;PSMD6;PSMD7;CDC23;PSMD4;KAT5;PSMD5;CDC26;PSMD2;PSM
1227	GO_Biological_Process_2021	DNA repair (GO:0006281)	251/298	7.55E-37	4.07E-34	0	0	5.6071412	466.3691076	SMARCAL1;MDC1;TRRAP;ALKBH3;ALKBH2;ALKBH5;ENDOV;HERC2;KAT5;TRIM28;CHEK2;ALKBH1;
1228	GO_Biological_Process_2021	cellular response to DNA damage s	281/350	1.37E-33	6.78E-31	0	0	4.27989121	323.8575475	ATF2;SMARCAL1;CCNK;TRRAP;SMC5;SMC6;ALKBH7;ALKBH8;ALKBH3;ENDOV;HERC2;TRIM28;CH
7149	Pfam_Domains_2019	Pkinase	237/347	4.15E-13	1.53E-10	0	0	2.24475655	63.99828544	TRIO;MYLK;RPS6KA4;RPS6KA3;RPS6KA6;TBK1;RPS6KA5;CHEK2;AKT2;RPS6KA2;CHEK1;AKT3;CDK
7150	Pfam_Domains_2019	RRM 1	152/206	5.34E-13	1.53E-10	0	0	2.92323302	82.60829556	HNRNPR;MSI2;TIAL1;RBMX2;PABPC4L;SNRNP35;RBFOX2;CIRBP;ZCRB1;UHMK1;RBMXL1;EWSR
7151	Pfam_Domains_2019	DEAD	58/67	1.61E-10	3.07E-08	0	0	6.65807607	150.1368513	DDX3Y;EIF4A2;EIF4A1;DDX49;DDX3X;DDX46;DDX47;EIF4A3;DDX42;DDX41;DHX57;DHX15;DHX16;
7152	Pfam_Domains_2019	PHD	44/52	1.18E-07	1.68E-05	0	0	5.67478101	90.54554114	KDM5A;PHF3;KDM5B;INTS12;DIDO1;KDM5C;KDM5D;UHRF2;KMT2A;UHRF1;PHF23;PHF1;KMT2C;
7153	Pfam_Domains_2019	Bromodomain	33/38	1.41E-06	0.00016086	0	0	6.80413063	91.68338839	ZMYND8;BAZ2A;BAZ2B;ATAD2B;TRIM24;EP300;BRD9;BRD8;BRD7;TRIM66;BPTF;BRD4;BRD3;BR
7154	Pfam_Domains_2019	OB NTP bind	17/17	5.98E-06	0.00057026	0	0	172346	2072771.039	DHX8;DHX9;DQX1;YTHDC2;DHX40;DHX30;DHX32;DHX33;DHX34;DHX35;DHX57;DHX36;DHX15;DHX
7155	Pfam_Domains_2019	BTB	87/129	2.26E-05	0.00184693	0	0	2.13945195	22.88669862	ZBTB25;ZBTB24;ZBTB26;IPP;ZBTB21;KEAP1;ZBTB20;ZBTB22;RCBTB1;RCBTB2;ENC1;ANKFY1;ZB
7156	Pfam_Domains_2019	WW	31/38	4.07E-05	0.00264177	0	0	4.56371245	46.13960565	YAP1;TCERG1;SETD2;STXBP4;WWC1;WWC2;WBP4;NEDD4L;SAV1;BAG3;APBB2;MAGI1;FNBP4;
7157	Pfam_Domains_2019	KH 1	29/35	4.16E-05	0.00264177	0	0	4.98030442	50.2424094	TDRKH;ANKRD17;KHDRBS1;FMR1;HDLBP;AKAP1;FXR1;FXR2;PCBP3;PCBP4;PCBP1;KHSRP;PCBP2;
7158	Pfam_Domains_2019	AAA	36/46	5.30E-05	0.0029938	0	0	3.71064523	36.53474467	VCP;NVL;VPS4B;VPS4A;SPG7;SPATA5;CHTF18;ATAD2B;SPAST;ORC1;KATNA1;FIGNL1;LONP1;LO
7159	Pfam_Domains_2019	HA2	17/18	5.76E-05	0.0029938	0	0	17.5042153	170.884097	DHX8;DHX9;DQX1;YTHDC2;DHX40;DHX30;DHX32;DHX33;DHX34;DHX35;DHX57;DHX36;DHX15;DHX
7160	Pfam_Domains_2019	UQ con	32/40	6.37E-05	0.00303609	0	0	4.12207528	39.8250856	UBE2D3;UBE2D1;UBE2Z;UBE2J2;UBE2J1;UBE2L3;UBE2Q1;UBE2Q2;UBE2F;UBE2H;UBE2I;UBE2
7161	Pfam_Domains_2019	DnaJ	35/45	8.49E-05	0.00373647	0	0	3.60720464	33.81322816	DNAJC24;DNAJC25;DNAJC27;DNAJC28;DNAJB2;DNAJB1;DNAJB6;DNAJB4;DNAJC21;DNAJB9;SEC
7162	Pfam_Domains_2019	PX	38/50	0.00010333	0.00422098	0	0	3.2640133	29.9556791	PXK;SNX12;SNX13;SNX10;SNX11;PIK3C2A;PLD1;SNX30;PLD2;SNX3;SNX4;SNX1;SNX29;SNX2;SH3

Mouse vs Human Enriched Heat Shock Pfam Domains

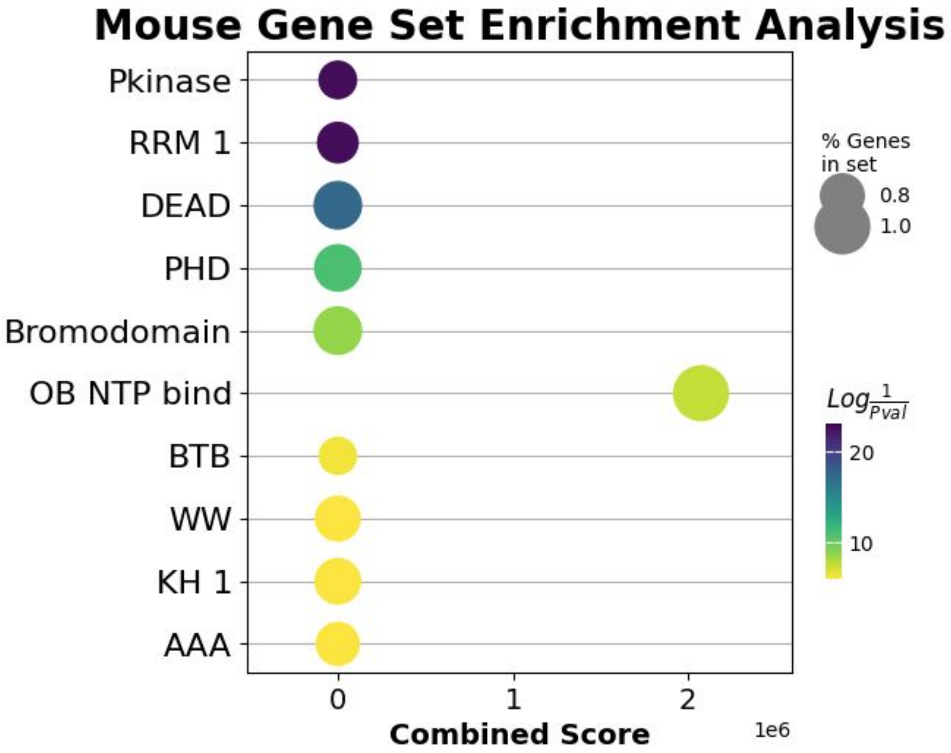


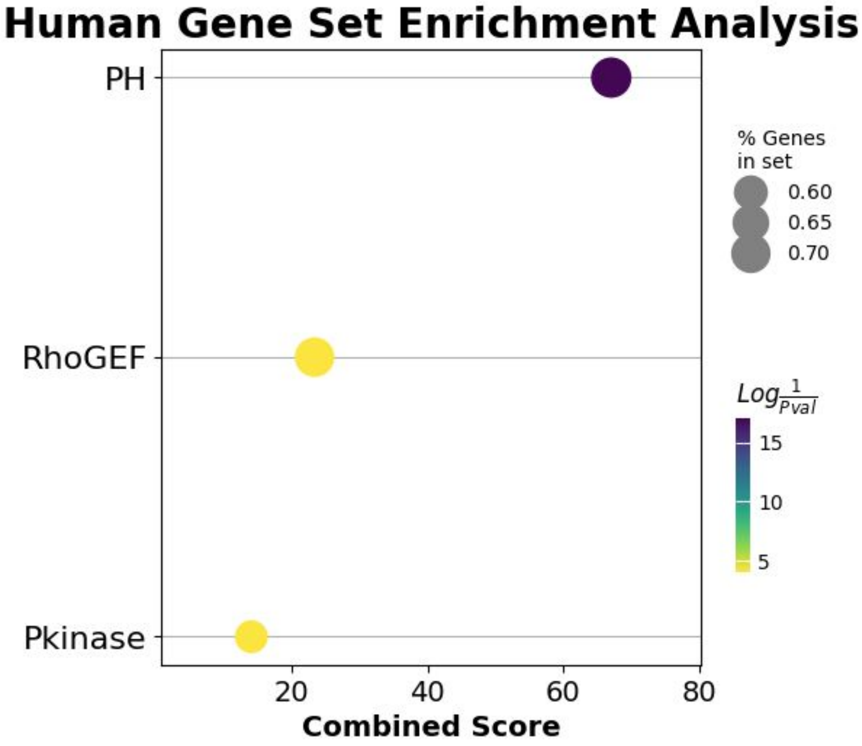
Mouse Gene Set Enrichment Analysis



Human Gene Set Enrichment Analysis

Mouse vs Human Enriched Heat Shock Pfam Domains





🔥 Hot Summary ☀

Take home: Analyzing the same expression data using four different python pipelines provided more evidence for shared heat shock responses among evolutionarily distant organisms!

What would we do differently?

- Generic functions standardized downstream input but it was difficult to anticipate the most useful format for all analyses
- One file one dictionary method would speed up code
  - BUT: combined dict allows for easy mutability of # of species comparing

What worked well?

- Using standardized identifiers for genes (ie. Ensembl IDs) and annotations (ie. pfam IDs) made it easy to use the same code across species

Sometimes it is hard to understand how package works - google a lot!