Pathway Enrichment for Hinds Kinome Manuscript

Robert M Flight

There is an [Executive Summary](#executive-summary) at the end of this report.

## Purpose

To investigate pathway enrichment for Hinds kinome manuscript. Pathway sources include Gene Ontology biological process, KEGG, and Reactome. The primary result here is the **comparison** between Human and Mouse enrichments. Further investigation probably requires a discussion.

## Data

Ideally all of the genes (peptides) measured on the arrays as well as the kinases that target them should constitute the background set. However, we don’t have access to them at the moment, so we will start with the kinases in KEA3 and their targets. The foreground set is the differential things spit out by UKA.

## Methods

Hypergeometric enrichment using categoryCompare2. Kegg pathway annotations were pulled using the Bioconductor KEGGREST package on 2021-10-27.

## Get Data

knitr::opts\_chunk$set(warning = FALSE, message = FALSE)

library(tidygraph)  
library(dplyr)  
library(ggraph)  
library(categoryCompare2)  
library(KEGGREST)  
library(KEGGgraph)  
library(ICIKendallTau)  
theme\_set(cowplot::theme\_cowplot())  
knitr::opts\_chunk$set(fig.width = 8, fig.height = 8)  
source(here::here("rmflight\_reports/kea3\_functions.R"))  
save\_loc = here::here("rmflight\_reports/pathway\_enrichment\_2021-09-29")  
  
gene\_list\_files = dir(here::here("input"), pattern = "^z", full.names = TRUE)  
  
gene\_lists = purrr::map(gene\_list\_files, ~ scan(.x, character(), sep = "\n", quiet = TRUE, skip = 1))  
names(gene\_lists) = c("human\_ptk", "mouse\_ptk", "human\_stk", "mouse\_stk")  
  
human\_list = unique(unlist(gene\_lists[c("human\_ptk", "human\_stk")]))  
mouse\_list = unique(unlist(gene\_lists[c("mouse\_ptk", "mouse\_stk")]))  
  
specific\_list = scan(here::here("input", "very\_specific\_genelist.txt"), what = character(), sep = "\n", quiet = TRUE)

## Map to Entrez ID

library(org.Hs.eg.db)  
all\_entrez = keys(org.Hs.eg.db)  
org\_hs\_df = select(org.Hs.eg.db, keys = all\_entrez, columns = "SYMBOL")  
  
mouse\_entrez = dplyr::left\_join(data.frame(SYMBOL = mouse\_list), org\_hs\_df, by = "SYMBOL") %>%  
 dplyr::pull(ENTREZID)  
human\_entrez = dplyr::left\_join(data.frame(SYMBOL = human\_list), org\_hs\_df, by = "SYMBOL") %>%  
 dplyr::pull(ENTREZID)

We need to filter the background down to the kinases and their targets first here, because I think we are enriched for signaling things, we are going to get signaling things in general.

kinase\_files = dir(here::here("kea3\_datasets"), full.names = TRUE)  
kinase\_files = grep("Cheng.KSIN|PTMsigDB|PhosD.All|STRING.bind", kinase\_files, value = TRUE)  
kinase\_data = purrr::map(kinase\_files, parse\_gmt)  
kinase\_genes = unique(unlist(kinase\_data))  
kinase\_entrez = dplyr::left\_join(data.frame(SYMBOL = kinase\_genes), org\_hs\_df, by = "SYMBOL") %>%  
 dplyr::filter(!is.na(ENTREZID)) %>%  
 dplyr::pull(ENTREZID)  
  
specific\_entrez = dplyr::left\_join(data.frame(SYMBOL = specific\_list), org\_hs\_df, by = "SYMBOL") %>%  
 dplyr::filter(!is.na(ENTREZID)) %>%  
 dplyr::pull(ENTREZID)

## GO Enrichment

go\_all = readRDS(here::here("rmflight\_reports/go\_all.rds"))  
go\_bp = readRDS(here::here("rmflight\_reports/go\_bp.rds"))  
human\_go\_all = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = human\_entrez,  
 universe = kinase\_entrez,  
 annotation = go\_all),  
 p\_adjust = "BH"  
)  
human\_go\_bp = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = human\_entrez,  
 universe = kinase\_entrez,  
 annotation = go\_bp),  
 p\_adjust = "BH"  
)  
  
mouse\_go\_all = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = mouse\_entrez,  
 universe = kinase\_entrez,  
 annotation = go\_all),  
 p\_adjust = "BH"  
)  
mouse\_go\_bp = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = mouse\_entrez,  
 universe = kinase\_entrez,  
 annotation = go\_bp),  
 p\_adjust = "BH"  
)  
  
all\_go = combine\_enrichments(human = human\_go\_all,  
 mouse = mouse\_go\_all)  
  
gocut = 0.001  
  
all\_go\_sig = get\_significant\_annotations(all\_go, padjust <= gocut, counts >= 2)  
  
goall\_table = extract\_stats\_table(all\_go\_sig, specific\_entrez)  
  
  
bp\_go = combine\_enrichments(human = human\_go\_bp,  
 mouse = mouse\_go\_bp)  
bp\_sig = get\_significant\_annotations(bp\_go, padjust <= gocut, counts >= 2)  
  
gobp\_table = extract\_stats\_table(bp\_sig, specific\_entrez)  
dim(gobp\_table)

## [1] 546 14

## KEGG

kegg\_annotation = readRDS(here::here("rmflight\_reports/kegg\_annotation.rds"))  
  
all\_kegg = unique(unlist(kegg\_annotation@annotation\_features))  
  
human\_kegg = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = human\_entrez,  
 universe = kinase\_entrez,  
 annotation = kegg\_annotation),  
 p\_adjust = "BH"  
)  
  
mouse\_kegg = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = mouse\_entrez,  
 universe = kinase\_entrez,  
 annotation = kegg\_annotation),  
 p\_adjust = "BH"  
)  
  
kegg\_enrich = combine\_enrichments(human = human\_kegg,  
 mouse = mouse\_kegg)  
kegg\_sig = get\_significant\_annotations(kegg\_enrich, padjust <= gocut, counts >= 2)  
  
kegg\_table = extract\_stats\_table(kegg\_sig, specific\_entrez)

## Reactome

So, one issue with the KEGG pathways, is they don’t seem to contain an annotation for one particular gene of interest, DDR1. In contrast, reactome seems to have it.

reactome\_annotation = readRDS(here::here("rmflight\_reports/reactome\_annotation.rds"))  
reactome\_path\_ddr1 = "R-HSA-1474244"

human\_reactome = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = human\_entrez,  
 universe = kinase\_entrez,  
 annotation = reactome\_annotation),  
 p\_adjust = "BH"  
)  
  
mouse\_reactome = hypergeometric\_feature\_enrichment(  
 new("hypergeom\_features", significant = mouse\_entrez,  
 universe = kinase\_entrez,  
 annotation = reactome\_annotation),  
 p\_adjust = "BH"  
)  
reactome\_cut = 0.005  
reactome\_enrich = combine\_enrichments(human = human\_reactome,  
 mouse = mouse\_reactome)  
reactome\_sig = get\_significant\_annotations(reactome\_enrich, padjust <= reactome\_cut, counts >= 2)  
reactome\_table = extract\_stats\_table(reactome\_sig, specific\_entrez)

As I’ve looked at these results, I’ve noticed some interesting themes, namely processes and pathways that involve a large amount of extracellular matrix reorganization / adhesion, and various signaling pathways.

The various signaling pathways make sense, because we are looking at the targets of kinase signaling.

The extracellular matrix stuff is interesting, because fibrosis necessarily involves a good amount of extracellular matrix and cellular adhesion processes / pathways.

We will put those into a separate table in the Excel file.

reactome\_table$source = "reactome"  
gobp\_table$source = "GO.BP"  
kegg\_table$source = "KEGG"  
  
all\_results = rbind(gobp\_table,  
 kegg\_table,  
 reactome\_table)  
interesting\_table = all\_results %>%  
 dplyr::filter(grepl("matrix|adhes.\*", pathway, ignore.case = TRUE)) %>%  
 dplyr::arrange(human.padjust)  
dim(interesting\_table)

## [1] 17 15

## Comparisons

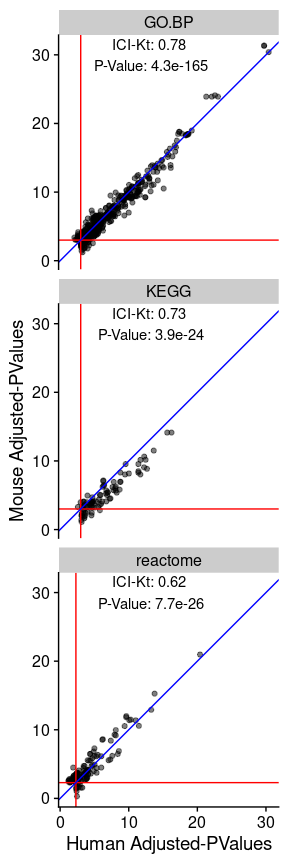
kegg\_table = add\_genes(kegg\_table, kegg\_sig, org\_hs\_df) %>%  
 dplyr::filter(n\_annotated <= 500) %>%  
 dplyr::arrange(human.padjust)  
gobp\_table = add\_genes(gobp\_table, bp\_sig, org\_hs\_df) %>%  
 dplyr::filter(n\_annotated <= 500) %>%  
 dplyr::arrange(human.padjust)  
reactome\_table = add\_genes(reactome\_table, reactome\_sig, org\_hs\_df) %>%  
 dplyr::filter(n\_annotated <= 500) %>%  
 dplyr::arrange(human.padjust)  
  
combined\_annot = c(kegg\_annotation@annotation\_features[kegg\_table$path\_id], go\_bp@annotation\_features[gobp\_table$path\_id], reactome\_annotation@annotation\_features[reactome\_table$path\_id])  
  
export\_pathway\_tables(list(KEGG = kegg\_table,  
 REACTOME = reactome\_table,  
 GOBP = gobp\_table,  
 ADHESION = interesting\_table),  
 pathway\_file = here::here("rmflight\_reports/pathway\_tables.xlsx"))

Finally, we can use Kendall-tau correlation to ask how similar the p-values are within each table between mouse and human. The list includes anything that was **significant** (below the p-value cutoff) for either mouse or human.

split\_source = split(all\_results, all\_results$source)  
ici\_source = purrr::map\_df(split\_source, function(.x){  
 ici\_res = ici\_kt(.x$mouse.padjust, .x$human.padjust, "global")  
 text\_res = paste0("ICI-Kt: ", format(ici\_res["tau"], digits = 2), "\n",  
 " P-Value: ", format(ici\_res["pvalue"], digits = 2))  
 data.frame(source = .x$source[1],  
 tau = ici\_res["tau"],  
 pvalue = ici\_res["pvalue"],  
 text = text\_res)  
})  
  
ici\_values = all\_results %>%  
 dplyr::group\_by(source) %>%  
 dplyr::summarise(ici\_kt = ici\_kt(human.padjust, mouse.padjust, "global")[1])  
  
knitr::kable(ici\_values, digits = 2)

|  |  |
| --- | --- |
| source | ici\_kt |
| GO.BP | 0.78 |
| KEGG | 0.73 |
| reactome | 0.62 |

sig\_cuts = data.frame(cut = c(gocut, gocut, reactome\_cut),  
 source = c("GO.BP",  
 "KEGG",  
 "reactome"))  
ici\_source$x = 13  
ici\_source$y = 30  
  
sig\_plots = ggplot(all\_results, aes(x = -1 \* log10(human.padjust), y = -1 \* log10(mouse.padjust))) + geom\_point(alpha = 0.5) +  
 geom\_hline(aes(yintercept = -1 \* log10(cut)), data = sig\_cuts, color = "red") +  
 geom\_vline(aes(xintercept = -1 \* log10(cut)), data = sig\_cuts, color = "red") +  
 geom\_abline(slope = 1, color = "blue") +  
 geom\_text(data = ici\_source, aes(x = x, y = y, label = text)) +  
 facet\_wrap(~ source, ncol = 1) +  
 coord\_equal() +  
 labs(x = "Human Adjusted-PValues",  
 y = "Mouse Adjusted-PValues")  
sig\_plots



Log10 Adjusted-PValues with P-Value cutoffs indicated by red lines and perfect agreement by blue lines. Points in a specific quadrant indicate pathways significant in only that organism.

## Executive Summary

* Data:
  + Gene Ontology biological process (GO.BP), KEGG pathways, and Reactome pathways.
  + Significant kinase targets from UKA (significant genes) for both Mouse and Human. We merged STK and PTK results together.
  + Kinases and their targets from KEA3 (background genes).
* Process:
  + Map symbols to Entrez IDs using org.Hs.eg.db.
  + Hypergeometric enrichment using categoryCompare2.
  + Significant enrichment are those with at least 2 genes annotated, and then adjusted p-values <=
    - GO.BP: 0.001
    - KEGG: 0.001
    - reactome: 0.005
* Results:
  + See **pathway\_tables\_XXX.xlsx** for the enriched pathways. Each pathway database is a separate worksheet, as well as those pathways related to extracellular matrix or adhesion. Each worksheet is sorted by human adjusted-p-value.
  + Signaling, in all three sources. This isn’t that surprising given we have the kinases. I had hoped modifying the background would suppress these enrichments a bit, but no luck.
  + In particular, however is PI3K signaling at the top p-values in KEGG and Reactome, and also peptidyl-tyrosine phosphorylation in GO.BP.
    - These match a paper that Hunter identified: <https://pubmed.ncbi.nlm.nih.gov/32681257/>
  + KEGG, returned the pathway “Non-alcoholic fatty liver disease” as significant in both mouse and human, which is one pathway that leads to liver fibrosis.
  + Also the pathways “Axon guidance,” “Osteoclast differentiation,” “Focal adhesion” and “Adherens junction” all involve extracellular matrix regulation.
  + Similar story in Reactome.
  + GO.BP also has things related to extracellular matrix and adhesion.
  + We compared the adjusted p-value results of enrichment between human and mouse using Kendall correlation and plotted them against each other. The results show that both organisms pathway enrichment are rather similar, with just a few things coming up in either organism.

### Table Columns

* path\_id: the database identifier.
* human.p: raw enrichment p-value for human.
* human.odds: odds ratio that is roughly observed / expected.
* human.expected: number of expected genes annotated to pathway.
* human.counts: actual number of genes annotated to pathway.
* human.padjust: Benjamini-Hochberg adjusted p-value.
* mouse….: same as above just for mouse.
* pathway: the title of the pathway.
* n\_specific: number of genes in that pathway that are present in the very specific gene list previously sent by Justin.
* source: the pathway source.
* genes: which genes from the UKA enriched gene list are annotated to the pathway.

## Methods Write Up

Gene symbols were mapped to Entrez gene identifiers using the Homo sapiens Bioconductor database package (org.Hs.eg.db) (Carlson 2021b). Gene annotations for Gene Ontology (GO) biological process (BP) (Ashburner et al. 2000) also came from the GOALL column of the org.Hs.eg.db package. Descriptions for each GO Term were extracted from the GO.db package. (Carlson 2021a) Kyoto Encyclopedia Genes and Genomes (KEGG) [M. Kanehisa and Goto (2000); Minoru Kanehisa (2019); kanehisaKEGGIntegratingViruses2021] pathway annotations were queried using KEGGREST package (Tenenbaum and Maintainer 2021) on 2021-11-01. Reactome pathway annotations (Jassal et al. 2020) came from the reactome.db package (Ligtenberg 2021). For each source, a categoryCompare2 (version 0.99.158) (Flight [2015] 2021; Flight et al. 2014) annotation object was created. KEA3 (Kuleshov et al. 2021) annotation files for kinase binding were used to define the background set of genes, including all the kinases and their targets. Hypergeometric enrichment was run for each of human and mouse differential genes after combining the gene lists for both STK and PTK results, for each pathway source. P-values were adjusted using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995). After combining human and mouse results, significant pathways were those with an adjusted p-value <= 0.001 (GO and KEGG) or 0.005 (Reactome) in either human or mouse, with at least 2 genes from the significant list, and no more than 500 genes annotated to the pathway. The similarity of human and mouse enrichments was evaluated using Kendall’s tau on the adjusted p-values, calculated using the ICIKendallTau R package (**ICIKendallTau2021?**).

## References

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology.” *Nature Genetics* 25 (1): 25–29. <https://doi.org/10.1038/75556>.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. <https://www.jstor.org/stable/2346101>.

Carlson, Marc. 2021a. “GO.db: A Set of Annotation Maps Describing the Entire Gene Ontology.” Manual.

———. 2021b. *Org.Hs.eg.db: Genome Wide Annotation for Human* (version 3.14.0).

Flight, Robert M. (2015) 2021. *categoryCompare2* (version 0.99.158). R. <https://github.com/rmflight/categoryCompare2>.

Flight, Robert M., Benjamin J. Harrison, Fahim Mohammad, Mary B. Bunge, Lawrence D. F. Moon, Jeffrey C. Petruska, and Eric C. Rouchka. 2014. “categoryCompare, an analytical tool based on feature annotations.” *Frontiers in Genetics* 5: 98. <https://doi.org/10.3389/fgene.2014.00098>.

Jassal, Bijay, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, et al. 2020. “The reactome pathway knowledgebase.” *Nucleic Acids Research* 48 (D1): D498–503. <https://doi.org/10.1093/nar/gkz1031>.

Kanehisa, M., and S. Goto. 2000. “KEGG: kyoto encyclopedia of genes and genomes.” *Nucleic Acids Research* 28 (1): 27–30. <https://doi.org/10.1093/nar/28.1.27>.

Kanehisa, Minoru. 2019. “Toward understanding the origin and evolution of cellular organisms.” *Protein Science: A Publication of the Protein Society* 28 (11): 1947–51. <https://doi.org/10.1002/pro.3715>.

Kuleshov, Maxim V, Zhuorui Xie, Alexandra B K London, Janice Yang, John Erol Evangelista, Alexander Lachmann, Ingrid Shu, Denis Torre, and Avi Ma’ayan. 2021. “Kea3: Improved Kinase Enrichment Analysis via Data Integration.” *Nucleic Acids Research* 49 (W1): W304–16. <https://doi.org/10.1093/nar/gkab359>.

Ligtenberg, Willem. 2021. *Reactome.db: A Set of Annotation Maps for Reactome* (version 1.77.0).

Tenenbaum, Dan, and Bioconductor Package Maintainer. 2021. *KEGGREST: Client-Side REST Access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)* (version 1.34.0).