KEA3 Enrichment for Hinds Kinome Manuscript

Robert M Flight

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

## Purpose

To summarize the enrichment results from KEA3 on input lists of kinase targets via network view.

## Background on KEA3

KEA3 ([web](https://maayanlab.cloud/kea3/), [publication](https://academic.oup.com/nar/article/49/W1/W304/6279841), [my hypothes.is annotations](https://hyp.is/go?url=https%3A%2F%2Facademic.oup.com%2Fnar%2Farticle%2F49%2FW1%2FW304%2F6279841&group=__world__)) curated a list of kinase - substrate interactions from multiple sources. Each of these sources generates a set of kinase - substrate annotations or gene-lists. The idea is that from your experiment, you upload a list of genes that *might* interact with the kinases in KEA3, then for each source:

\* input list of genes is tested for over-representation in each list of kinase - substrate annotations (hypergeometric test or Fishers test)

\* this test generates a p-value

\* all p-values **ranked** for that source

An overall ranking **across** annotation sources is generated by taking the **average** rank (MeanRank), or reporting the **top rank** across the sources (TopRank).

When reporting results from the web application programming interface (API), we get the *MeanRank* for each kinase, as well as which genes in the input list **overlapped** with the annotated list in KEA3. In addition, they make the kinase - substrate gene-list annotations available as gene sets. Finally, for each annotation source, they do have the underlying raw p-values and false-discovery corrected p-values and ranks available.

## Data

source(here::here("kea3\_functions.R"))

##   
## Attaching package: 'tidygraph'

## The following object is masked from 'package:stats':  
##   
## filter

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Loading required package: ggplot2

save\_loc = here::here("kea3\_enrichment\_2021-09-29")

### KEA3 Enriched Results

Four lists of input genes were passed to KEA3.

* PTK Human Fib vs Ctrl
* PTK Mouse Fib vs Ctrl
* STK Human Fib vs Ctrl
* STK Mouse Fib vs Ctrl

For each list of genes, they were queried on KEA3 using the [web API](https://maayanlab.cloud/kea3/templates/api.jsp), and results transformed from JSON to a list. The code for the API queries is in “kea3\_runs.R” and should be run separately from generating this document.

### Definitions

* **Enriched**: this kinase was in the **top 10** of MeanRank’ed enrichment results from KEA3 (see background on KEA3 above).
* **Overlapped**: this kinase / gene was in the input list of genes and the genes annotated to that kinase in KEA3.

### Human Protein Atlas Data

The consensus expression data from Human Protein Atlas was downloaded. This data will be used to limit results to genes that are known to be expressed in Liver.

## Methods

### Enrichment

For each gene list, we queried KEA3 using the publicly accessible API and stored results as simple lists in R. **Enriched** kinases are those in the **top 10** kinases by MeanRank score. For each of the enriched kinases, we also extracted the genes from the input list that overlap with the kinase annotation.

### Network Generation

From the KEA3 downloads page, various GMT files were downloaded. Using the downloaded datasets, we created weighted networks of kinase-substrate pairs and protein-protein interactions. The weight of the edge in the network between any two genes represents how many datasets that edge was observed in. Therefore, multiple sources of evidence for the pair of proteins to be interacting adds to the weight of the network edge.

### Tissue Expression

Tissue specific gene / protein expression was defined from the Human Protein Atlas (HPA) mRNA expression consensus data, which is defined as maximum observed expression in that tissue for HPA, GTEx, Fantom5. Expression values are first transformed using . The mode for all log(expression) > 1 across all tissues was determined, and a standard deviation (SD) from values greater than the mode calculated, using the mode as the mean for the SD calculation. A lower limit cutoff was calculated as:

All genes with a log(expression) > than in the tissue of interest are said to be expression in the tissue.

## Results

kea3\_results = readRDS(here::here("kea3\_enrichment\_2021-09-29", "kea3\_results.rds"))  
hpa\_consensus = read.table(here::here("hpa\_data", "rna\_consensus\_2021-09-27.tsv"), header = TRUE, sep = "\t") %>%  
 dplyr::mutate(logX = log1p(NX))  
kea3\_network = readRDS(here::here("kea3\_datasets", "kea3\_networks\_2021-09-30.rds"))

First, we need to figure out what is expressed in the liver. We previously examined how to do this in another report. Initially, we will use what is expressed in liver, not necessarily what is “specific” to liver.

mode\_sd\_all = hpa\_consensus %>%  
 dplyr::filter(logX > 1) %>%  
 dplyr::pull(logX) %>%  
 calculate\_mode\_sd()  
  
min\_val = mode\_sd\_all["mode"] - 2 \* mode\_sd\_all["sd"]  
liver\_expressed = hpa\_consensus %>%  
 dplyr::filter(Tissue %in% "liver", logX >= min\_val) %>%  
 dplyr::pull(Gene.name)  
min\_val

## mode   
## 1.157528

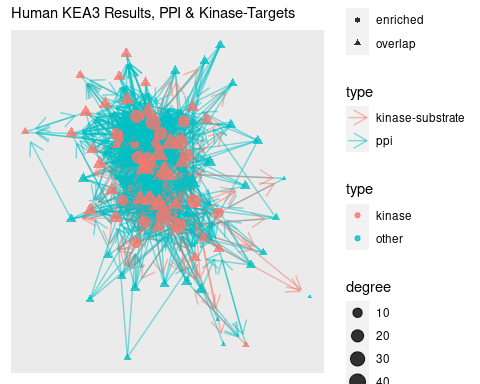
Using this cutoff, we have 12238 genes with robust expression in the liver. This number may be too large, and we may need to filter to what is *specific* to liver instead.

Second, lets look at the KEA3 results and the networks, filtered by what is expressed in liver. In this one, we will use the **top ten** enriched kinases (see definitions above) using *MeanRank*, as well as their overlapping genes.

# changing names to make them easier to filter to.  
names(kea3\_results) = c("PTK\_Human", "PTK\_Rodent", "STK\_Human", "STK\_Roden")  
  
# getting enriched kinases  
human\_enrich = purrr::map\_dfr(kea3\_results[c("PTK\_Human", "STK\_Human")], function(.x){  
 data.frame(gene = .x$`Integrated--meanRank`$TF[1:10],  
 source = "human",  
 type = "kinase")  
})  
human\_enrich = human\_enrich %>%  
 dplyr::filter(gene %in% liver\_expressed)  
  
# and overlapping genes, essentially the kinase targets  
human\_overlaps = purrr::map\_dfr(kea3\_results[c("PTK\_Human", "STK\_Human")],  
 function(.x){  
 overlap\_genes = .x$`Integrated--meanRank`$Overlapping\_Genes[1:10]  
 all\_over = unlist(strsplit(overlap\_genes, ","))  
 data.frame(gene = all\_over,  
 source = "human",  
 type = "target")  
 }) %>%  
 unique()  
rownames(human\_overlaps) = NULL  
human\_overlaps = human\_overlaps %>%  
 dplyr::filter(gene %in% liver\_expressed)  
  
human\_overlaps = human\_overlaps %>%  
 dplyr::filter(!(gene %in% human\_enrich))  
  
human\_results = rbind(human\_enrich, human\_overlaps) %>%  
 dplyr::filter(gene %in% liver\_expressed)  
  
human\_network = kea3\_network %>%  
 activate(nodes) %>%  
 filter(name %in% human\_results$gene) %>%  
 mutate(enriched = case\_when(  
 name %in% human\_enrich$gene ~ "enriched",  
 TRUE ~ "overlap"  
 )) %>%  
 activate(edges) %>%  
 filter(weight > 1) %>% # do we have multiple evidence?  
 activate(nodes) %>%  
 mutate(degree = local\_size(order = 1, mindist = 1)) %>% # how many neighbors?  
 filter(degree > 1)

OK, let’s see what the network looks like?

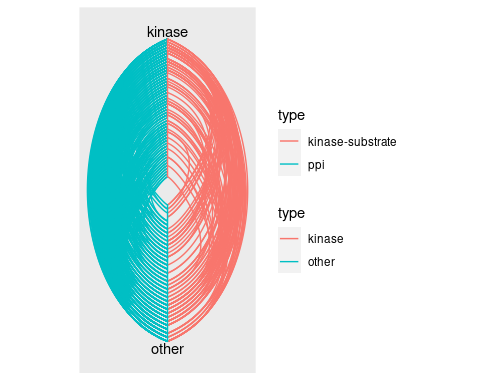
ggraph(human\_network, layout = "fr") +  
 geom\_edge\_link(aes(color = type), arrow = arrow(length = unit(4, 'mm')),  
 end\_cap = circle(3, 'mm'),  
 alpha = 0.5) +  
 geom\_node\_point(aes(color = type, shape = enriched, size = degree),  
 alpha = 0.8) +  
 labs(subtitle = "Human KEA3 Results, PPI & Kinase-Targets")



As we can see from this plot, we have a hairball that is difficult to interpret. Even with labels, this is not likely to make much sense.

As another solution, we will try another type of layout, and see if it helps.

ggraph(human\_network, layout = "hive", axis = type, sort.by = degree) +  
 geom\_edge\_hive(aes(color = type)) +  
 geom\_axis\_hive(aes(color = type)) +  
 coord\_fixed()

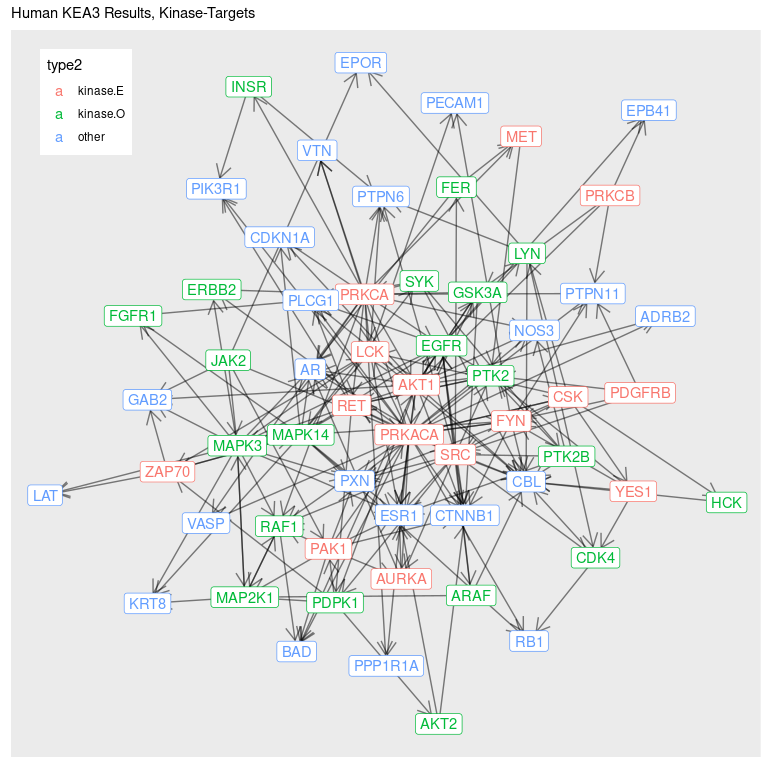


Yeah, that doesn’t help us at all, unfortunately.

OK, we will try the hairball again, but remove any edges that aren’t between Kinase-substrate, and then remove anything that isn’t attached to something else.

human\_network\_kt = human\_network %>%  
 activate(edges) %>%  
 filter(type %in% "kinase-substrate") %>%  
 activate(nodes) %>%  
 mutate(degree = local\_size(order = 1, mindist = 1)) %>%  
 filter(degree > 1) %>%  
 mutate(type2 = case\_when(  
 (type %in% "kinase") & (enriched %in% "enriched") ~ "kinase.E",  
 (type %in% "kinase") & (enriched %in% "overlap") ~ "kinase.O",  
 TRUE ~ "other"  
 ))

# this is in it's own chunk so we can set.seed and reproduce it.  
set.seed(123)  
ggraph(human\_network\_kt, layout = "graphopt") +  
 geom\_edge\_link(arrow = arrow(length = unit(4, 'mm')),  
 end\_cap = circle(3, 'mm'),  
 alpha = 0.5) +  
 geom\_node\_label(aes(label = name, color = type2)) +  
 labs(subtitle = "Human KEA3 Results, Kinase-Targets") +  
 theme(legend.position = c(0.1, 0.9))



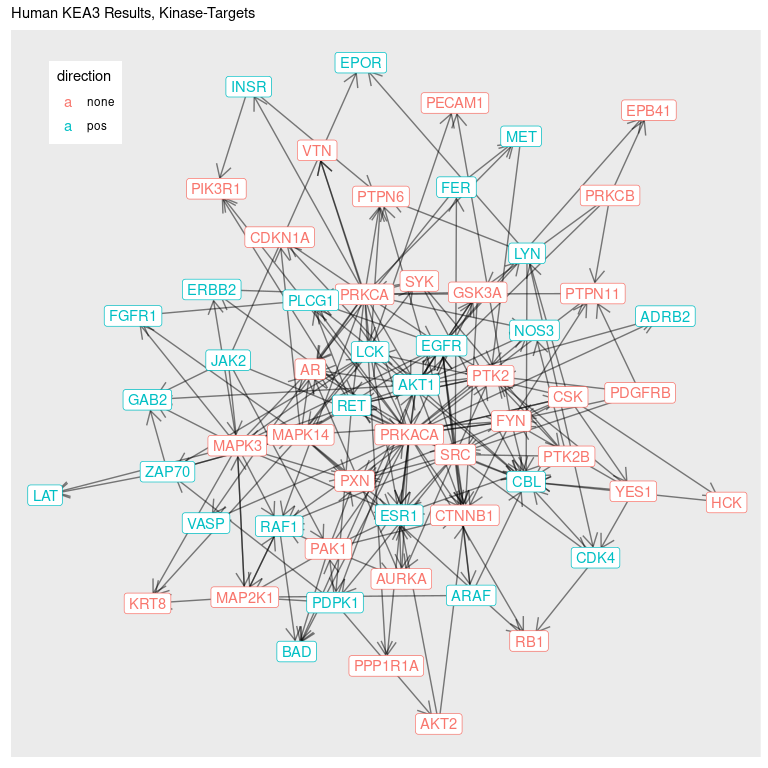
* kinase.E: A kinase gene, and enriched from KEA3
* kinase.O: A kinase gene, and overlapping in the input gene list
* other: A gene that was in the input list, and was a *target* of the kinase in the KEA3 data.

This is **starting** to look interpretable, I think. We can see the enriched kinases, any overlapping kinases that come up from the enrichment overlapping genes, and then the overlapping genes. I’m not sure what else should be here, so I’m going to stop.

What if we add some fold-changes from the substrates on here?

human\_lfc\_files = dir(here::here("input"), pattern = "human.\*LFC", full.names = TRUE)  
human\_lfc = get\_lfc(human\_lfc\_files)  
human\_lfc\_list = split(human\_lfc$peptide2, human\_lfc$direction)  
human\_network\_kt2 = human\_network\_kt %>%  
 activate(nodes) %>%  
 dplyr::mutate(direction = dplyr::case\_when(  
 name %in%human\_lfc\_list$neg ~ "neg",  
 name %in% human\_lfc\_list$pos ~ "pos",  
 TRUE ~ "none"  
 ))

set.seed(123)  
ggraph(human\_network\_kt2, layout = "graphopt") +  
 geom\_edge\_link(arrow = arrow(length = unit(4, 'mm')),  
 end\_cap = circle(3, 'mm'),  
 alpha = 0.5) +  
 geom\_node\_label(aes(label = name, color = direction)) +  
 labs(subtitle = "Human KEA3 Results, Kinase-Targets") +  
 theme(legend.position = c(0.1, 0.9))



## Executive Summary

* Used the differential **human** peptide lists as input to KEA3 enrichment.
* Used the KEA3 kinase-substrate lists to build a gene network.
* Used the HPA consensus data to limit things to what is expressed in Liver.
* Visualized the kinase-substrate network based on the enriched kinases and the differential genes / substrates (see above), with the direction of change for the differential genes applied.

**This is a first pass only**. We need to talk about whether there is anything to add or do differently. For example, we could think about directly taking the UKA results and putting them on the kinase-substrate network. We also have ideas on how to improve on the KEA3 analysis, but we don’t know if that will improve anything on this or it’s necessary for this manuscript.

Finally, after reading the KEA3 manuscript, we are not sure about whether the method they use for combining across datasets is appropriate, and their emphasis on selecting only the **top XXX** results.

TODO

Mapping of gene ids to protein names

UKA data so we can do LFC of targets, and LFC of kinases of fibrous to normal

Final figure network that is just the things that come up in both mouse and human

Targets measured in mouse and human, and calculate correlation using all conditions