

KEGG annotation analysis in R

There are multiple ways to do KEGG annotation in R and the method of choice depend on your starting material. i.e. the annotation you have available.

For Affymetrix GeneChips the easiest approach would in most cases be to use the annotation data from BioConductor. They can be found here:

<http://www.bioconductor.org/packages/release/data/annotation/>

look for “Array annotation data” and the package will be named “my_chip_cdf_name.db”. e.g. “ath1121501.db” for the Affymetrix Ath1 arabidopsis array.

```
require(ath1121501.db)

# map KEGG pathways to Affy ids and ship repeats
affy2kegg<-lapply(as.list(ath1121501PATH), unique)

#A semi-random geneset
kegg2affy<-as.list(revmap(ath1121501PATH))
geneset<-c(sample(kegg2affy$'04120', 10), sample(unlist(kegg2affy), 100))
# '04120' should be over-represented in this set

# some counting
pos_counts <- table(unlist(affy2kegg[geneset]))
bg_count <-table(unlist(affy2kegg)[names(pos_counts)])
total<-sum(bg_count)

# statistics
p_kegg <-phyper(pos_counts, bg_count, total-bg_count, length(geneset),
lower.tail=FALSE )

require(KEGG.db)
kegg_pnames<-unlist(mget(names(p_kegg), KEGGPATHID2NAME))

kegg_result<-data.frame(cbind(p_kegg, pos_counts, bg_count), kegg_pnames)
[order(p_kegg),]

kegg_result
#      p_kegg      pos_counts bg_count kegg_pnames
# 04120 5.408197e-09         16       90 Ubiquitin mediated ...
# 00350 1.112202e-06          7       25 Tyrosine metabolism
# 00641 1.217570e-06          5       12 3-Chloroacrylic acid ...
# 00903 2.481686e-06         11       68 Limonene and pinene ...
# ...
# you may not be able to reproduce this as I sampled the 'geneset' randomly.
```