The KEGGgraph Package

Stat 6570

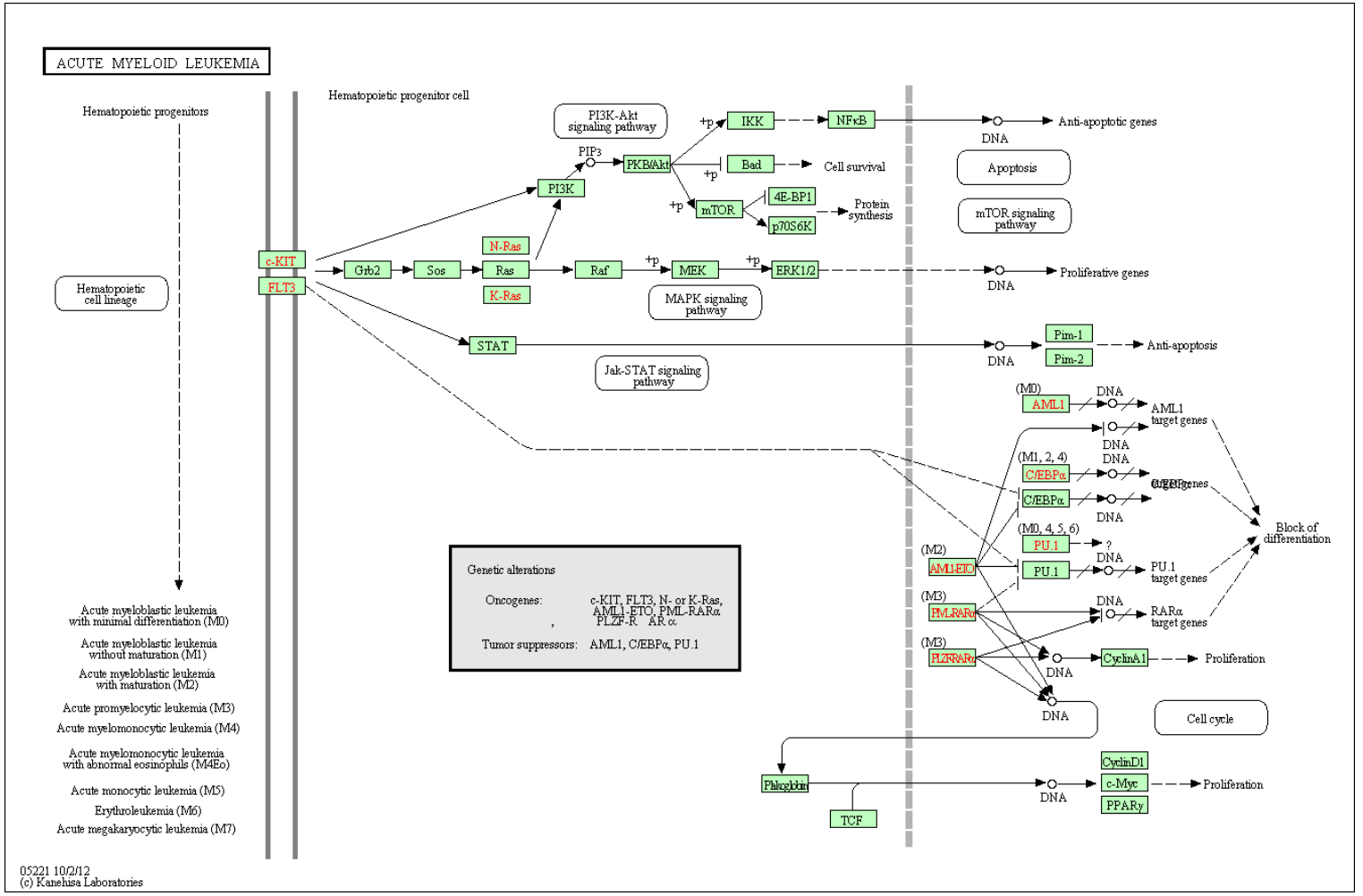
Michelle Carlsen

April 21, 2014

Kyoto Encyclopedia of Genes and Genomes

- An effort to map the whole of the genome
- Created a database of functions and utilities of the biological system
- One of the services of KEGG is KEGG PATHWAY
- They have constructed pathway maps of known biological processes
- They are difficult to understand
- KEGGgraph package was created to use the graph tools in R to create more easily understood mappings

KEGG PATHWAY



Finding Significant KEGG PATHWAY

- Use the global test
- Just how there was a global test, for gene ontology terms, there is one for KEGG PATHWAY

```
#Same data set used in class
#Load and format data
library(affy)
library(ALL)
data(ALL)
eset <- exprs(ALL)
T.cell <- c(rep(0,95),rep(1,33))
Eset <- new("ExpressionSet", exprs=eset)
pData(Eset) <- data.frame(trt=as.character(T.cell))
annotation(Eset) <- "hgu95av2"
```

```
#Perform global test
library(globaltest)
library(KEGG.db)
gt.KEGG <- gtKEGG(trt, Eset)
result <- data.frame(code = names(gt.KEGG), alias = gt.KEGG@extra[, 2],
pvalue = gt.KEGG@result[,1])
head(result)
```

	code	alias	pvalue
05142	05142	Chagas disease (American trypanosomiasis)	4.067826e-62
05340	05340	Primary immunodeficiency	5.951121e-61
04650	04650	Natural killer cell mediated cytotoxicity	1.382162e-59
04640	04640	Hematopoietic cell lineage	2.445855e-59
04380	04380	Osteoclast differentiation	7.889190e-57
04660	04660	T cell receptor signaling pathway	3.541398e-54

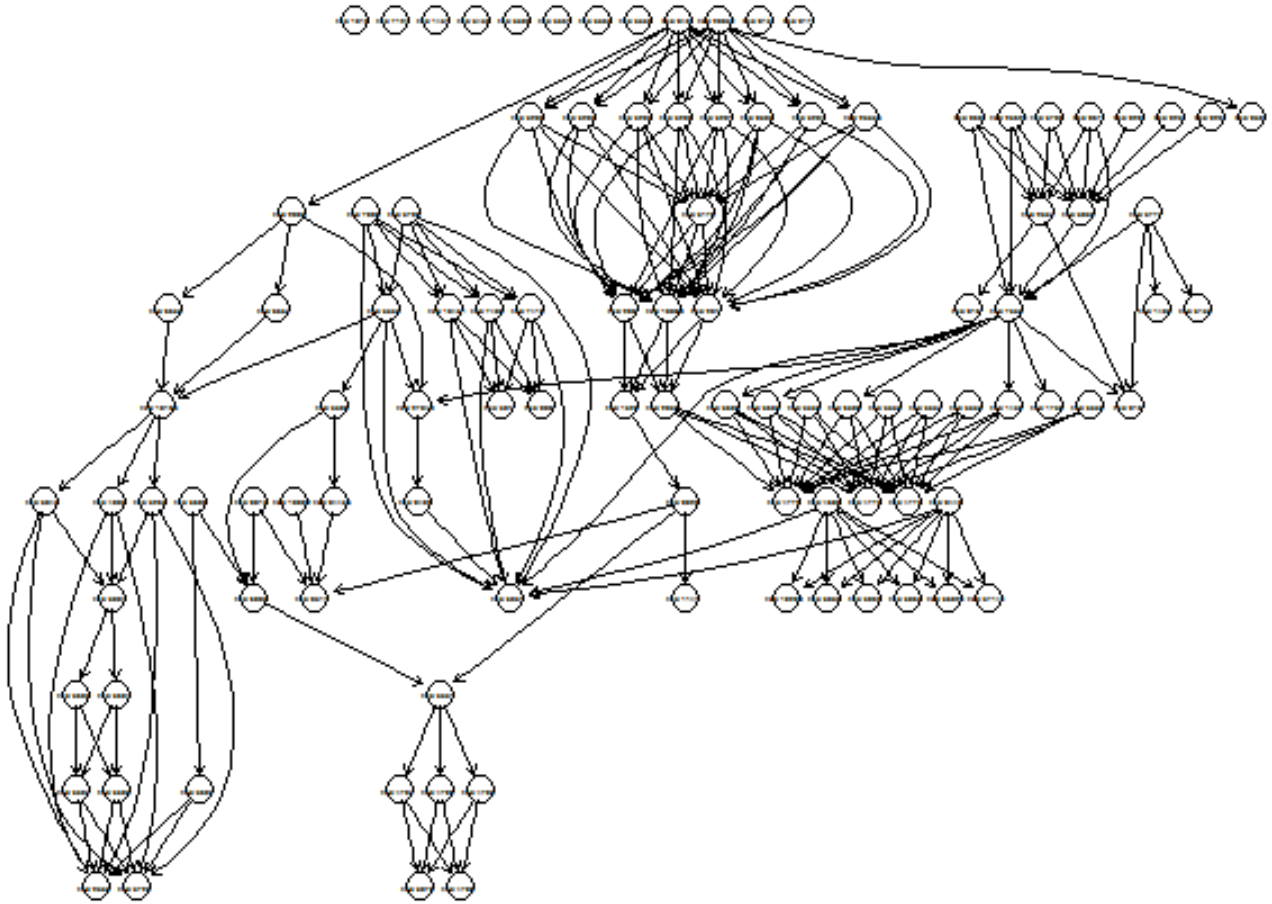
Loading KEGG PATH

- Each pathway is assigned an ID consisting of a 3 letter organism code and 5 number pathway code
- Stored as KGML files (XML) on the KEGG website
- Guaranteed way to load KGML into R

```
# Find KGML
library(KEGGgraph)
getKGMLurl("hsa04660")
# "http://www.genome.jp/kegg-bin/download?entry=hsa04660&format=kgml"
```

- Copy and paste into web browser
- Select Save as → R → library → KEGGgraph → extdata

```
#Load KGML
tcell <- system.file("extdata/hsa04660.xml", package="KEGGgraph")
map.t <- parseKGML2Graph(tcell, expandGenes=TRUE)
plot(map.t)
```



Exploring the KEGG Path

```
#Number of genes
mapkNodes <- nodes(map.t)
length(mapkNodes)
# [1] 104

#Look at some genes
mapkNodes[c(6, 9, 27, 63)]
#[1] "hsa:3567" "hsa:4792" "hsa:2885" "hsa:387"

#Degree of nodes
#Informative of how active a gene is
mapkGoutdegrees <- sapply(edges(map.t), length)
mapkGindegrees <- sapply(inEdges(map.t), length)
topouts <- sort(mapkGoutdegrees, decreasing=T)
topins <- sort(mapkGindegrees, decreasing=T)
```

```
topouts[1:3]
#hsa:940 hsa:29851  hsa:7535
#      9          9          8
topins[1:3]
#hsa:4772 hsa:4773 hsa:4775
#10       10       10

#Look at some edges
mapkEdges <- edges(map.t)
mapkEdges[c(6, 9, 27, 63)]
#$`hsa:3567`
#character(0)

#$`hsa:4792`
#[1] "hsa:4790" "hsa:5970"

#$`hsa:2885`
#[1] "hsa:27040" "hsa:6654"  "hsa:6655"

#$`hsa:387`
#character(0)
```

Finding Active Genes

```
#Filter
library(genefilter)
e.mat <- 2^exprs(ALL)
ffun <- filterfun(pOverA(0.20,100))
t.fil <- genefilter(e.mat,ffun)
small.eset <- log2(e.mat[t.fil,])
T.cell <- c(rep(0,95),rep(1,33))

#Test for signficant genes
library(limma)
Cell <- as.factor(c(rep('B',95),rep('T',33)))
design <- model.matrix(~0+Cell)
colnames(design) <- c('B','T')
fit <- lmFit(small.eset, design)
contrast.Cell <- makeContrasts(B-T, levels=design)
fit.Cell <- contrasts.fit(fit, contrast.Cell)
final.fit.Cell <- eBayes(fit.Cell)
top.Cell <- topTableF(final.fit.Cell, n=nrow(small.eset))
```

Relate Probe IDs to KEGG IDs

```
#Get ENTREZ ID
library(hgu95av2.db)
ids <- top.Cell$ID
ENTREZ <- select(hgu95av2.db, ids, "ENTREZID", "PROBEID")
ENTREZ <- ENTREZ[!duplicated(ENTREZ[1]),]
top.Cell <- cbind(top.Cell, ENTREZ = ENTREZ$ENTREZID)
head(top.Cell)
```

	B...T	AveExpr	F	P.Value	adj.P.Val	ENTREZ
38319_at	-4.655042	6.041217	1242.0961	4.901575e-68	2.110128e-64	915
33238_at	-3.102294	7.292159	514.1754	8.067444e-47	1.736517e-43	3932
35016_at	3.214222	10.337892	497.9320	4.209382e-46	6.040463e-43	972
2059_s_at	-2.668482	7.232735	489.0459	1.058404e-45	1.139107e-42	3932
37039_at	3.265990	11.072596	454.2931	4.448245e-44	3.829939e-41	3122
38095_i_at	3.762299	10.228156	416.3136	3.446245e-42	2.472681e-39	3115

What Genes from the KEGG PATH are Signficant?

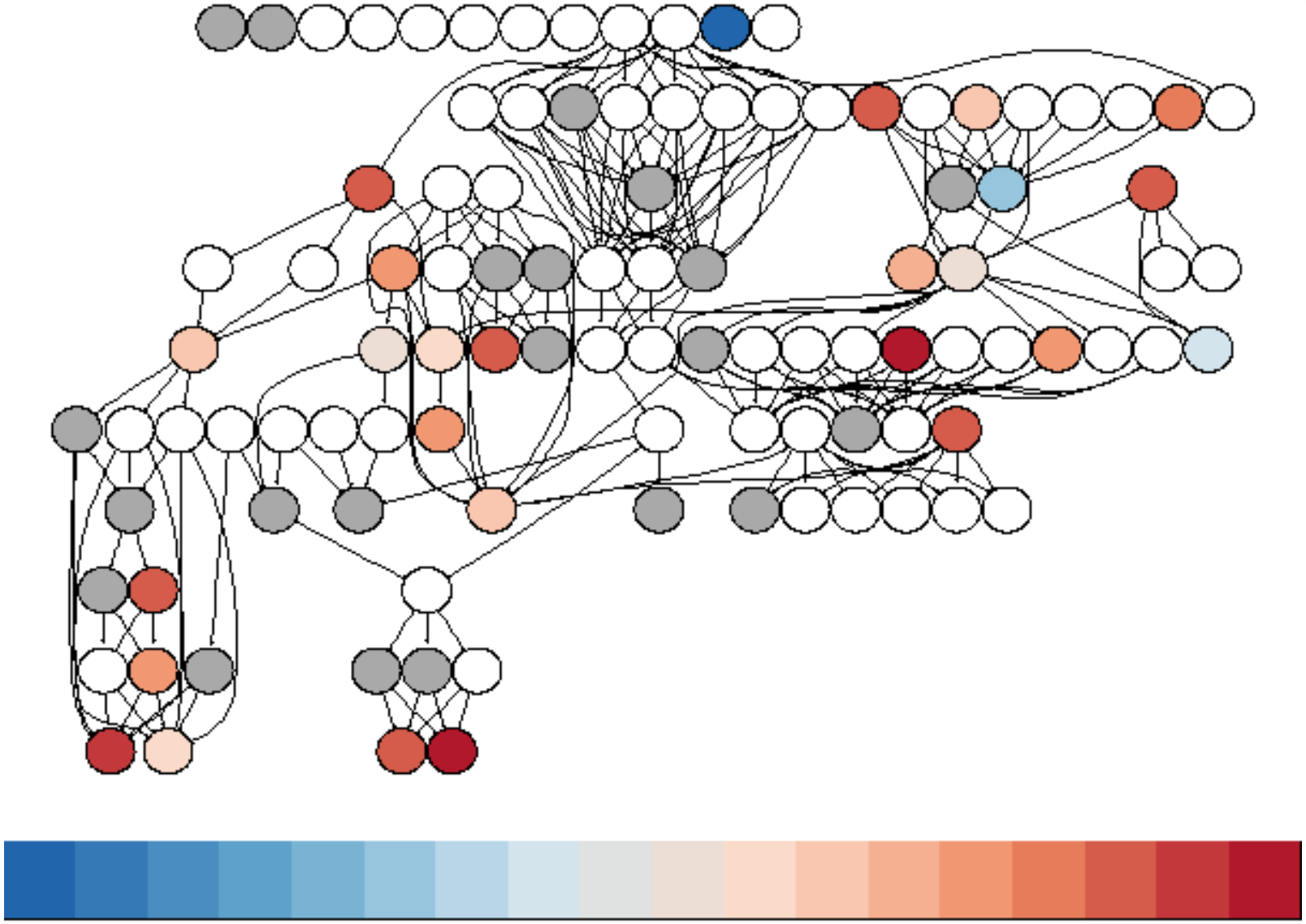
```
#Sort significant genes
tg <- top.Cell[top.Cell$adj.P.Val < 0.05, ]
BT_sig <- tg$B...T                          #Create vector of significant genes
names(BT_sig) <- as.vector(tg$ENTREZ)# and names
BT_all <- as.character(top.Cell$ENTREZ)#Create vector of all gene names

#Relate genes from KEGGgraph to differentially expressed genes
deKID <- translateGeneID2KEGGID(names(BT_sig))
allKID <- translateGeneID2KEGGID(BT_all)
isDiffExp <- nodes(map.t) %in% deKID
sprintf("%2.2f%% genes differentially-expressed", mean(isDiffExp)*100)
#[1] "25.00% genes differentially-expressed"
```

Show Significant Genes on KEGGgraph

```
#Make the map
library(RColorBrewer)
library(org.Hs.eg.db)
library(RBGL)
library(grid)
library(Rgraphviz)
ar <- 18
cols <- rev(colorRampPalette(brewer.pal(6, "RdBu"))(ar))
logfcs <- BT_sig[match(nodes(map.t), deKID)] #log-fold-change of
                                              #significant nodes
names(logfcs) <- nodes(map.t)
logfcs[is.na(logfcs)] <- 0
incol <- round((logfcs+5)*3)
incol[incol>ar] <- ar
```

```
undetected <- !nodes(map.t) %in% allKID #nodes not in our affymetrix data
logcol <- cols[incol] #color by log fold change
logcol[logfcs==0] <- "darkgrey" #genes that are not differentially expressed
logcol[undetected] <- "white" #genes not in our data set
names(logcol) <- names(logfcs)
nA <- makeNodeAttrs(map.t, fillcolor=logcol, label="", width=10, height=1.2)
par(mar=c(3,5,0,5), mgp=c(0,0,0))
layout(mat=matrix(c(rep(1,8),2), ncol=1, byrow=TRUE))
plot(map.t, "dot", nodeAttrs=nA)
image(as.matrix(seq(1,ar)), col=cols, yaxt="n", xaxt="n")
mtext("down-regulation", side=1,  at=0, line=1)
mtext("up-regulation", side=1,  at=1, line=1)
```

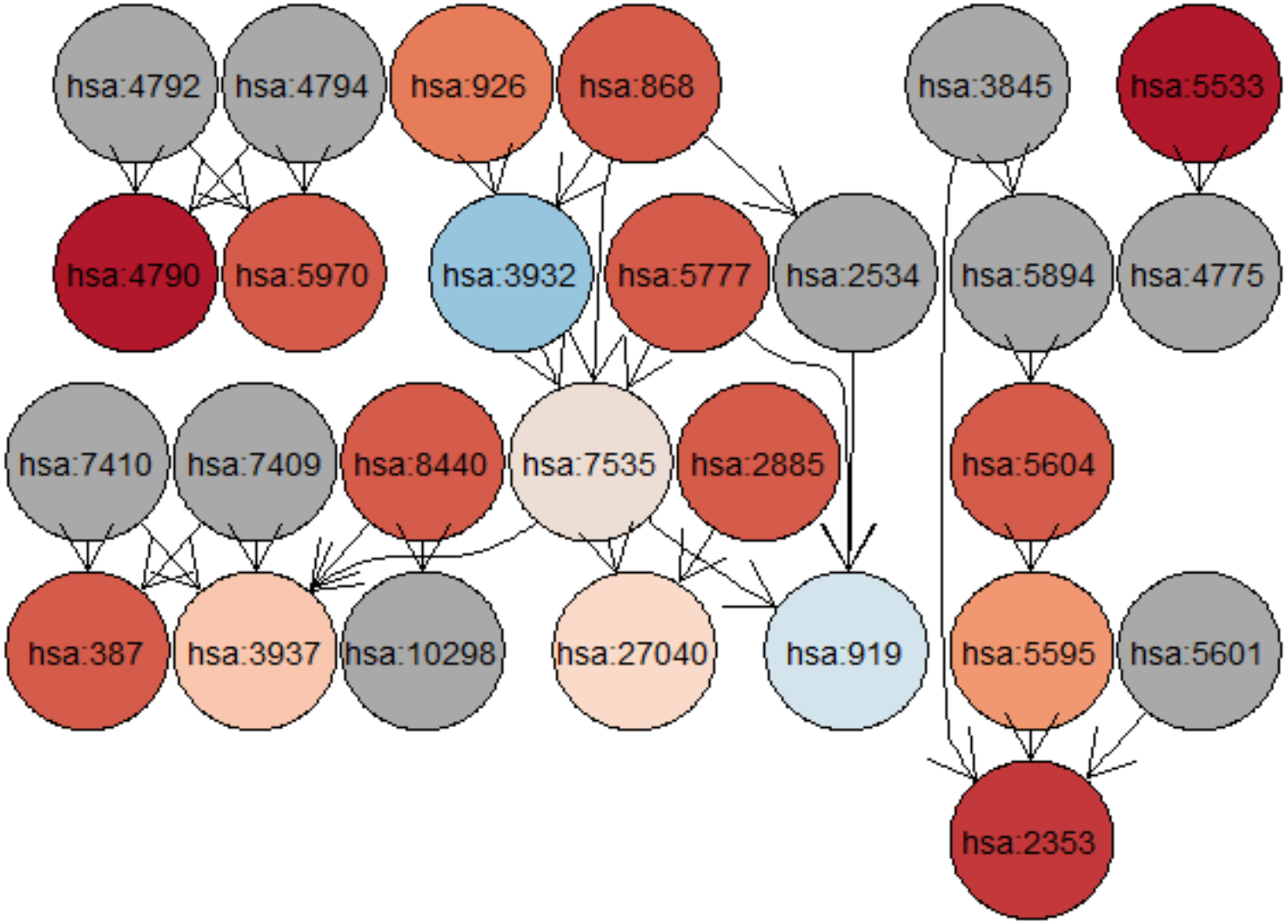


Zoom in on Up-regulated Genes only

- Hard to get a lot of information from that Graph

```
#Mark Genes with degree of zero
gDeg <- degree(map.t)
gIsSingle <- gDeg[[1]] + gDeg[[2]] == 0
options(digits=3)
gGeneID <- translateKEGGID2GeneID(nodes(map.t))
gSymbol <-  sapply(gGeneID, function(x) mget(x, org.Hs.egSYMBOL,
ifnotfound=NA)[[1]])
isUp <- logfcs > 0
isDown <- logfcs < 0
singleUp <- isUp & gIsSingle
singleDown <- isDown & gIsSingle
```

```
#Map of Up regulated genes and neighbors only
ups <- nodes(map.t)[logfcs > 0]
upNs <- unique(unlist(neighborhood(map.t, ups, return.self=TRUE)))
upSub <- subKEGGgraph(upNs, map.t)
upNeighbor <- nodes(upSub)[sapply(neighborhood(upSub,
nodes(upSub)), length)>0]
upNeighbor <- setdiff(upNeighbor, nodes(map.t)[undetected])
upSub <- subKEGGgraph(upNeighbor, upSub)
upSubGID <- translateKEGGID2GeneID(nodes(upSub))
upSymbol <- gSymbol[upSubGID]
upn <- makeNodeAttrs(upSub, fillcolor=logcol[nodes(upSub)],
fixedsize=TRUE, width=10, height=10, font=20)
dev.off()
plot(upSub, "dot", nodeAttrs = upn)
```


hsa:4792
hsa:4794
hsa:926
hsa:868
hsa:3845
hsa:5533
hsa:4790
hsa:5970
hsa:3932
hsa:5777
hsa:2534
hsa:5894
hsa:4775
hsa:7410
hsa:7409
hsa:8440
hsa:7535
hsa:2885
hsa:5604
hsa:387
hsa:3937
hsa:10298
hsa:27040
hsa:919
hsa:5595
hsa:5601
hsa:2353

Should this Path be Significant?

- Remember that genes with a high degree are considered active.
- If many differentially expressed genes are active, then the path will be significant.
- Does it make sense that this path is significant?
  - All of the up-regulated genes have at least one degree, and many have multiple

What are these genes?

```
#Exploring graph
getKEGGnodeData(map.t, "hsa:5533")
KEGG Node (Entry 'hsa:5533'):
#-------------------------------------------------------
#[ displayName ]: PPP3CA, CALN, CALNA, CALNA1, CCN1, CNA1, PPP2B...
#[ Name ]: hsa:5533
#[ Type ]: gene
#[ Link ]: http://www.kegg.jp/dbget-bin/www_bget?hsa:5530+hsa:5532+hsa:5533+hsa:5534+hsa:5535
#-------------------------------------------------------
```

- Following the link you can find more information than you ever wanted to know

```
getKEGGedgeData(map.t, 'hsa:5595~hsa:2353')

#  KEGG Edge (Type: PPrel):
#-------------------------------------------------
---------
#[ Entry 1 ID ]: hsa:5595
#[ Entry 2 ID ]: hsa:2353
#[ Subtype ]:
#   [ Subtype name ]: activation
#   [ Subtype value ]: -->
#   [ Subtype name ]: phosphorylation
#   [ Subtype value ]: +p
#-------------------------------------------------
---------
```

Sources

- Dr. Steven's bioinformatics notes
  - Unit 3.4
  - Unit 4.3
- Bioconductor "KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor" vignette
  - http://master.bioconductor.org/packages/release/bioc/vignettes/KEGGgraph/inst/doc/KEGGgraph.pdf
- Bioconductor "KEGGgraph: Application Examples" vignette
  - http://master.bioconductor.org/packages/release/bioc/vignettes/KEGGgraph/inst/doc/KEGGgraphApp.pdf
- KEGG: Kyoto Encyclopedia of Genes and Genomes
  - www.genome.jp/kegg/
- R – help pages
- R Graphical Manual: Get KGML file (url) with KEGG PATHWAY ID and (optional) organism
  - http://rgm.ogalab.net/RGM/R_rdfile?f=KEGGgraph/man/getKGMLurl.Rd&d=R_BC