STAT588/BIOL588: Genomic Data Science
Lecture 19: Annotation and Enrichment Analysis

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Objectives of Lecture 19

- Annotation
- Presenting Results
- Pathway Databases: GO, KEGG
- Hypergeometric Test
- Gene Set Enrichment Analysis

Biological Databases Related to Microarray

- Gene Ontology (GO)
- Kyoto Encyclopedia of Genes and Genomes (KEGG)
- Biocarta

Night Sky

Databases:GO

The Gene Ontology (GO) is a structured vocabulary of terms describing gene products according to molecular function, biological process, and cellular component.

- all : all [601410 gene products]
  - GO:0008150 : biological_process [448346 gene products]
  - GO:0005575 : cellular_component [425872 gene products]
  - **GO:0003674 : molecular_function [480087 gene products]**
    - GO:0016209 : antioxidant activity [3080 gene products]
    - GO:0005488 : binding [220352 gene products]
    - GO:0003824 : catalytic activity [196343 gene products]
    - GO:0016247 : channel regulator activity [712 gene products]
    - GO:0042056 : chemoattractant activity [136 gene products]
    - GO:0045499 : chemorepellent activity [34 gene products]
    - GO:0036370 : D-alanyl carrier activity [0 gene products]
    - GO:0009055 : electron carrier activity [7307 gene products]
    - GO:0030234 : enzyme regulator activity [11512 gene products]
    - GO:0016530 : metallochaperone activity [118 gene products]
    - GO:0060089 : molecular transducer activity [23510 gene products]
    - GO:0016015 : morphogen activity [41 gene products]
    - GO:0001071 : nucleic acid binding transcription factor activity [17602 gene products]
    - GO:0045735 : nutrient reservoir activity [238 gene products]
    - GO:0000988 : protein binding transcription factor activity [3874 gene products]
    - GO:0031386 : protein tag [88 gene products]
    - GO:0004872 : receptor activity [21080 gene products]
    - GO:0030545 : receptor regulator activity [338 gene products]
    - GO:0005198 : structural molecule activity [21519 gene products]
    - GO:0045182 : translation regulator activity [225 gene products]
    - GO:0005215 : transporter activity [29498 gene products]

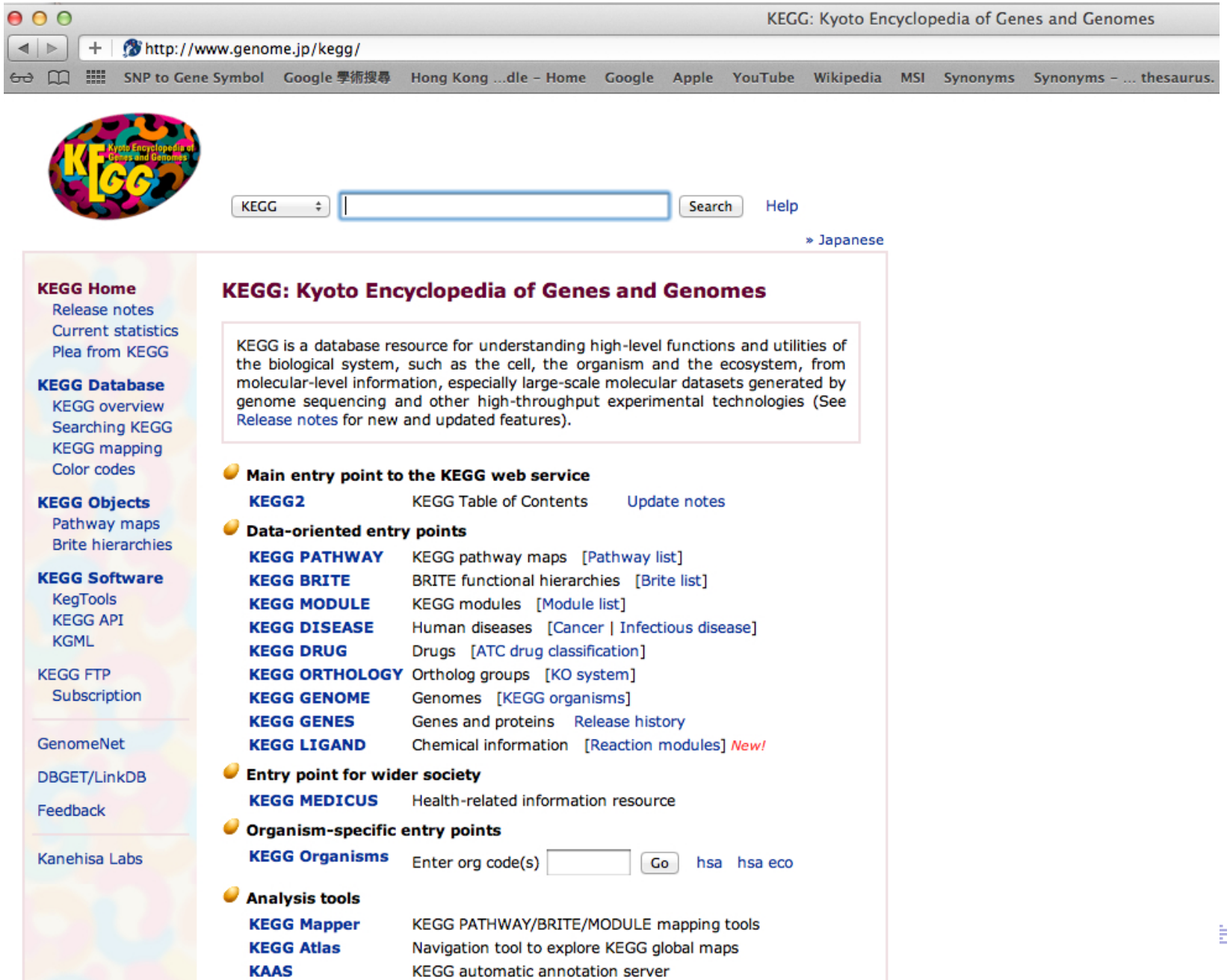
GO: Biological Process

- all : all [601410 gene products]
  - GO:0008150 : biological_process [448346 gene products]
    - GO:0022610 : biological adhesion [8062 gene products]
    - GO:0065007 : biological regulation [107173 gene products]
    - GO:0015976 : carbon utilization [243 gene products]
    - GO:0001906 : cell killing [1177 gene products]
      - GO:0097278 : complement-dependent cytotoxicity [0 gene products]
      - GO:0031640 : killing of cells of other organism [845 gene products]
      - GO:0001909 : leukocyte mediated cytotoxicity [343 gene products]
      - GO:0031342 : negative regulation of cell killing [65 gene products]
      - GO:0031343 : positive regulation of cell killing [252 gene products]
      - GO:0031341 : regulation of cell killing [306 gene products]
    - **GO:0008283 : cell proliferation [9709 gene products]**
      - GO:0003263 : cardioblast proliferation [49 gene products]
      - GO:0071838 : cell proliferation in bone marrow [20 gene products]
      - GO:0003295 : cell proliferation involved in atrial ventricular junction remodeling [0 gene products]
      - GO:0035736 : cell proliferation involved in compound eye morphogenesis [3 gene products]
      - GO:0060722 : cell proliferation involved in embryonic placenta development [4 gene products]
      - GO:0061323 : cell proliferation involved in heart morphogenesis [53 gene products]
      - GO:2000793 : cell proliferation involved in heart valve development [4 gene products]
      - GO:0090255 : cell proliferation involved in imaginal disc-derived wing morphogenesis [0 gene products]
      - GO:0072111 : cell proliferation involved in kidney development [98 gene products]
      - GO:0035988 : chondrocyte proliferation [28 gene products]
      - GO:0097360 : chorionic trophoblast cell proliferation [0 gene products]
      - GO:0035726 : common myeloid progenitor cell proliferation [10 gene products]
      - GO:0050673 : epithelial cell proliferation [1745 gene products]
      - GO:0070341 : fat cell proliferation [28 gene products]
      - GO:0048144 : fibroblast proliferation [426 gene products]
      - GO:0036093 : germ cell proliferation [13 gene products]
      - GO:0048134 : germ-line cyst formation [40 gene products]
      - GO:0014009 : glial cell proliferation [108 gene products]
      - GO:0003419 : growth plate cartilage chondrocyte proliferation [26 gene products]
      - GO:0071335 : hair follicle cell proliferation [27 gene products]
      - GO:0035172 : hemocyte proliferation [47 gene products]
      - GO:0071425 : hemopoietic stem cell proliferation [76 gene products]

GO: Cellular Component

- all : all [601410 gene products]
  - GO:0008150 : biological_process [448346 gene products]
  - **GO:0005575 : cellular_component [425872 gene products]**
    - GO:0005623 : cell [265176 gene products]
    - GO:0044464 : cell part [265127 gene products]
    - GO:0031012 : extracellular matrix [3562 gene products]
    - GO:0044420 : extracellular matrix part [1571 gene products]
    - GO:0005576 : extracellular region [25538 gene products]
    - GO:0044421 : extracellular region part [10293 gene products]
    - GO:0032991 : macromolecular complex [70880 gene products]
    - GO:0016020 : membrane [114123 gene products]
    - GO:0044425 : membrane part [76648 gene products]
    - GO:0031974 : membrane-enclosed lumen [22143 gene products]
    - GO:0009295 : nucleoid [643 gene products]
    - GO:0043226 : organelle [183834 gene products]
    - GO:0044422 : organelle part [67066 gene products]
    - GO:0055044 : symplast [848 gene products]
    - GO:0045202 : synapse [3910 gene products]
    - GO:0044456 : synapse part [2900 gene products]
    - GO:0019012 : virion [4197 gene products]
    - GO:0044423 : virion part [3699 gene products]
  - GO:0003674 : molecular_function [480087 gene products]

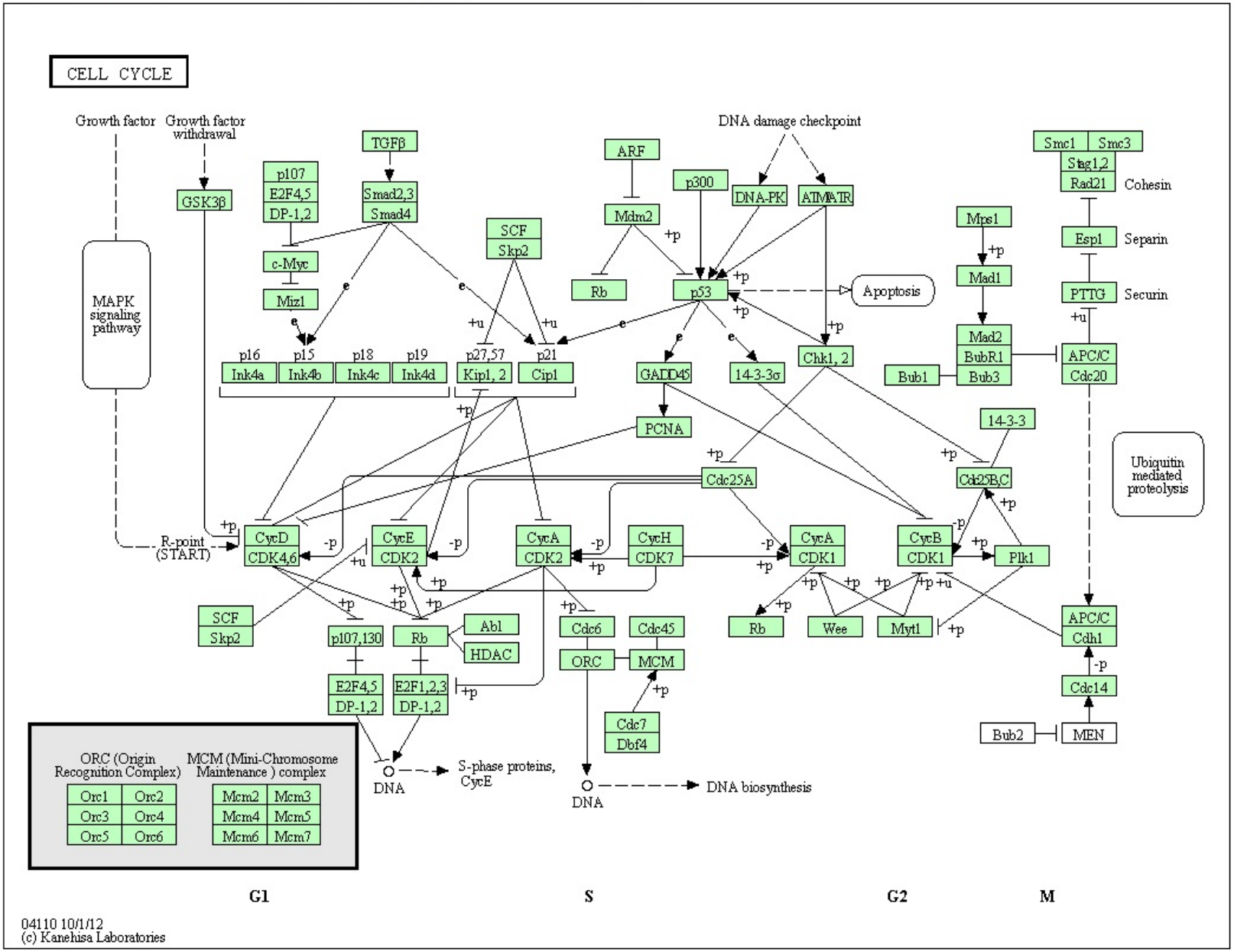
KEGG pathway



KEGG **Cell cycle - Homo sapiens (human)**

[Pathway menu | Organism menu | Pathway entry | Download KGML | Show description | User data mapping]

Homo sapiens (human) Go 100%

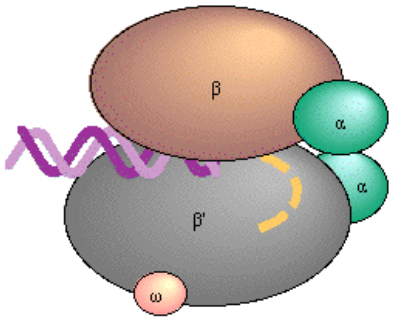


KEGG **RNA polymerase - Reference pathway**

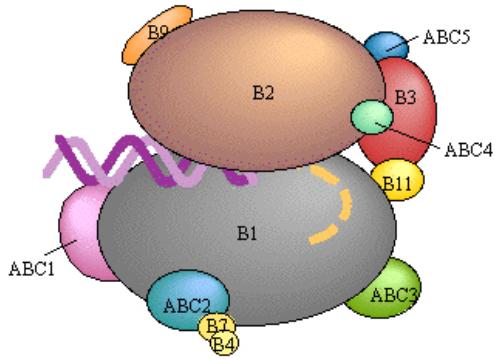
[Pathway menu | Organism menu | Pathway entry | User data mapping]

Reference pathway | Go | 100%

RNA POLYMERASE



RNA polymerase (Thermus aquaticus)



RNA polymerase II (Saccharomyces cerevisiae)

Bacterial

β	α	ω	δ
β'			

Archaeal

B	D	F	H	K	E
A	G		N	L	P

Eukaryotic Pol II

Core subunits

B2	B3
B1	B11

Pol II specific subunits

B4	B7	B9

Pol I, II, and III common subunits

ABC1	ABC2	ABC3
ABC4	ABC5	

Eukaryotic Pol III

Core subunits

C2	AC2
C1	AC1

Pol III specific subunits

C3	C4	C11	
C25	C31	C34	C37

Eukaryotic Pol I

Core subunits

A2	AC2
A1	AC1

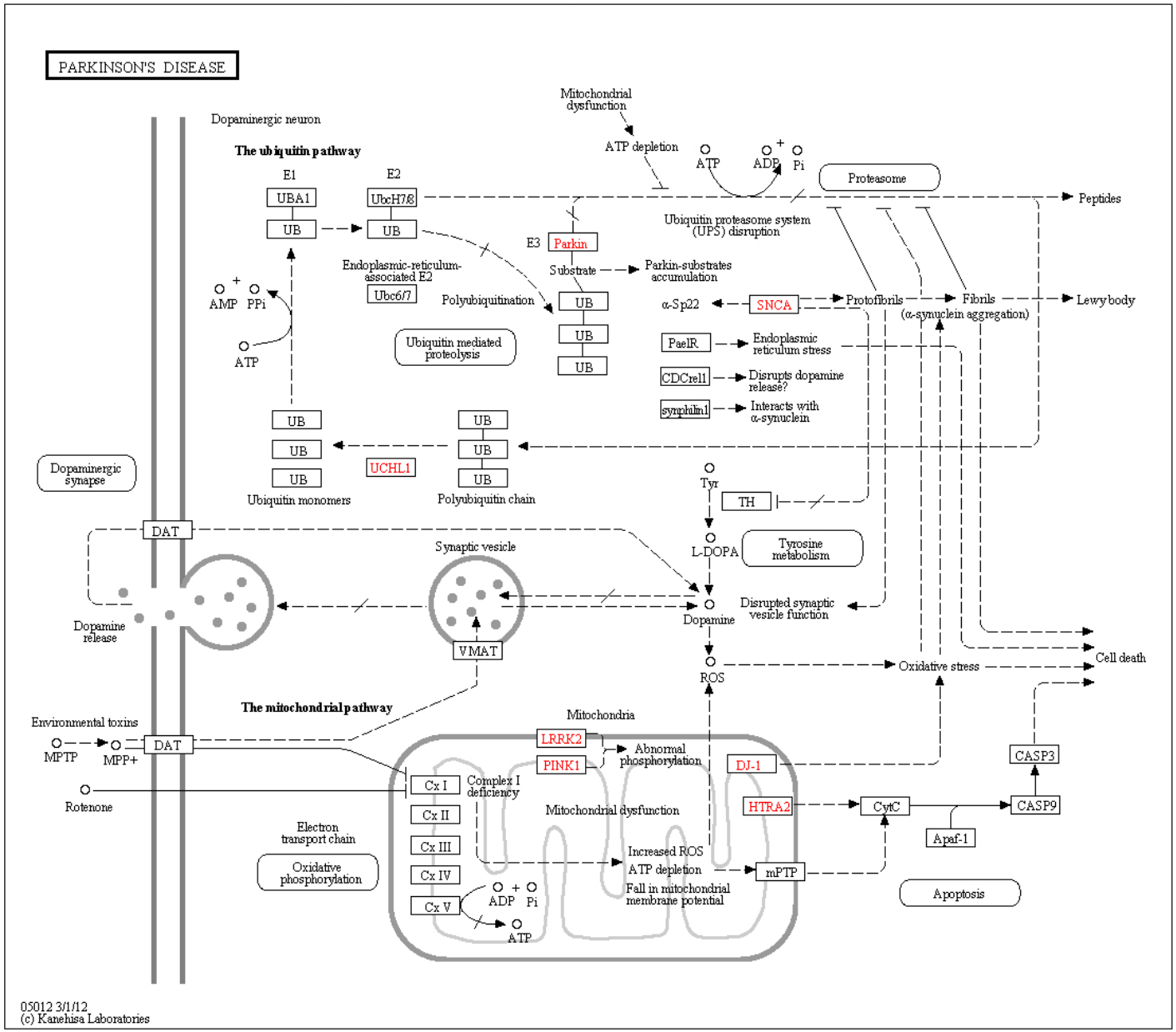
Pol I specific subunits

A12	A14	A34
A49	A43	

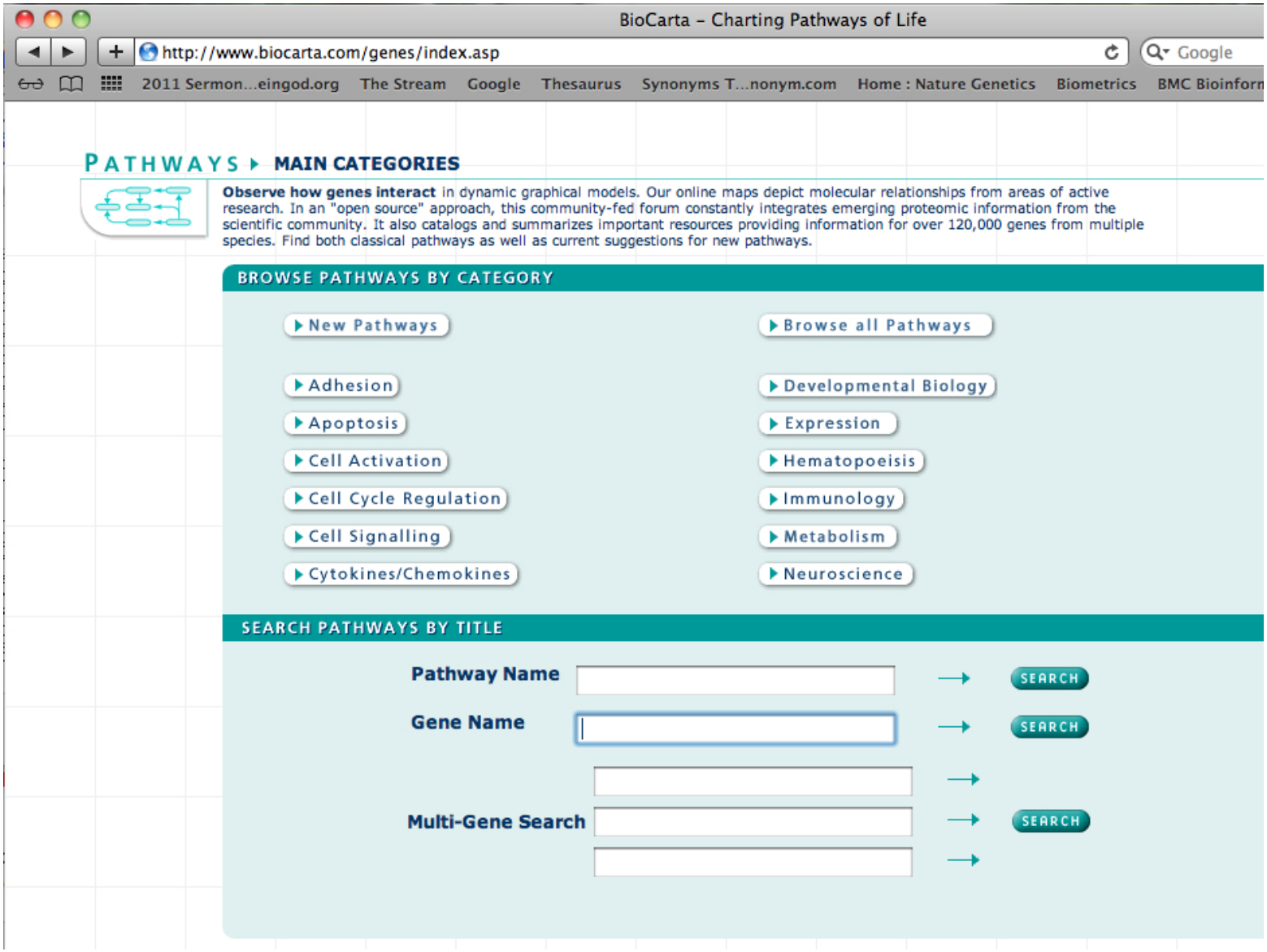
KEGG **Parkinson's disease - Reference pathway**

[Pathway menu | Organism menu | Pathway entry | Show description | User data mapping]

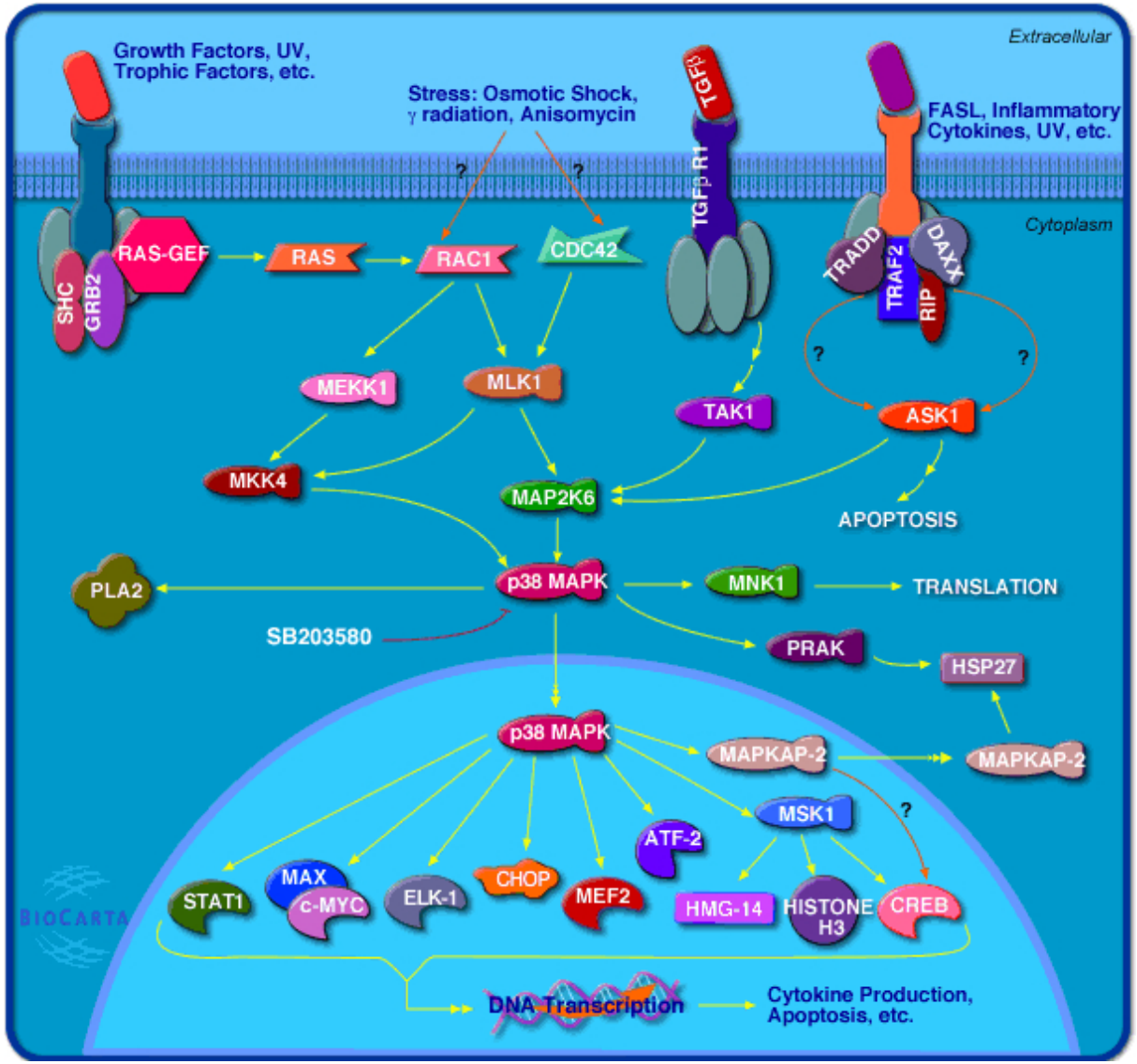
Reference pathway | Go | 100%



Biocarta


BioCarta – Charting Pathways of Life
http://www.biocarta.com/genes/index.asp
Google
2011 Sermon...eingod.org
The Stream
Google
Thesaurus
Synonyms T...nonym.com
Home : Nature Genetics
Biometrics
BMC Bioinform
PATHWAYS ▸ MAIN CATEGORIES
Observe how genes interact in dynamic graphical models. Our online maps depict molecular relationships from areas of active research. In an "open source" approach, this community-fed forum constantly integrates emerging proteomic information from the scientific community. It also catalogs and summarizes important resources providing information for over 120,000 genes from multiple species. Find both classical pathways as well as current suggestions for new pathways.
BROWSE PATHWAYS BY CATEGORY
New Pathways
Browse all Pathways
Adhesion
Apoptosis
Cell Activation
Cell Cycle Regulation
Cell Signalling
Cytokines/Chemokines
Developmental Biology
Expression
Hematopoeisis
Immunology
Metabolism
Neuroscience
SEARCH PATHWAYS BY TITLE
Pathway Name
SEARCH
Gene Name
SEARCH
Multi-Gene Search
SEARCH

P38 MAPK Signaling Pathway



Top Table

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
156	ABL1	1.10	9.20	9.03	4.88e-14	1.23e-10	21.29
1915	ABL1	1.15	9.00	8.59	3.88e-13	4.89e-10	19.34
155	ABL1	1.20	7.90	7.34	1.23e-10	1.03e-07	13.91
163	YES1	1.43	5.00	7.05	4.55e-10	2.87e-07	12.67
2066	PON2	1.18	4.24	6.66	2.57e-09	1.30e-06	11.03
2014	KLF9	1.78	8.62	6.39	8.62e-09	3.63e-06	9.89
1262	ALDH1A1	1.03	4.33	6.24	1.66e-08	6.00e-06	9.27
437	MARCKS	1.68	4.47	5.97	5.38e-08	1.70e-05	8.16
1269	AHNAK	1.35	8.44	5.81	1.10e-07	3.08e-05	7.49
1366	ANXA1	1.12	5.09	5.48	4.27e-07	1.08e-04	6.21

Enrichment Analysis

Is the selected set of genes enriched in the GO term of cell cycle?

	Related to Cell Cycle	Annotated but not Related to Cell Cycle	Not Annotated	Total
DE gene	100	691	9	800
Non DE gene	285	5012	65	5362
All gene	385	5703	74	6162

Enrichment Analysis

	Related to Cell Cycle	Annotated but not Related to Cell Cycle	Total
DE gene	100	691	791
Non DE gene	285	5012	5297
All gene	385	5703	6162

$$\frac{100}{791} = 12.64\% \qquad \frac{285}{5297} = 5.38\%$$

Enrichment Analysis

	Related to Cell Cycle	Annotated but not Related to Cell Cycle	Total
DE gene	100	691	791
Non DE gene	285	5012	5297
All gene	385	5703	6162

$$\chi^2 = \sum_{cell} \frac{(Observed - Expected)^2}{Expected}$$

$$\chi^2 = \frac{(100 - \frac{791\times385}{6088})^2}{\frac{791\times385}{6088}} + \frac{(691 - \frac{791\times5703}{6088})^2}{\frac{791\times5703}{6088}} + \frac{(285 - \frac{5297\times385}{6088})^2}{\frac{5297\times385}{6088}}$$

$$+ \frac{(5012 - \frac{5297\times5703}{6088})^2}{\frac{5297\times5703}{6088}} = 61.26$$

Enrichment Analysis

	Related to Cell Cycle	Annotated but not Related to Cell Cycle	Total
DE gene	100	691	791
Non DE gene	285	5012	5297
All gene	385	5703	6162

```
> chisq.test(matrix(c(285, 5012, 100, 691), 2, 2),
correct=F)
Pearson's Chi-squared test
X-squared = 61.2644, df = 1, p-value = 4.99e-15
> fisher.test(matrix(c(285, 5012, 100, 691), 2, 2))
Fisher's Exact Test for Count Data
p-value = 1.099e-12
95 percent confidence interval:
0.3073581 0.5055634
sample estimates:
odds ratio
0.3929809
```

Enrichment Analysis

Chi-squared test is an approximate test and may not perform well when sample size small. Fisher's exact test is a better alternative.

Enrichment Analysis: Hypergeometric Test

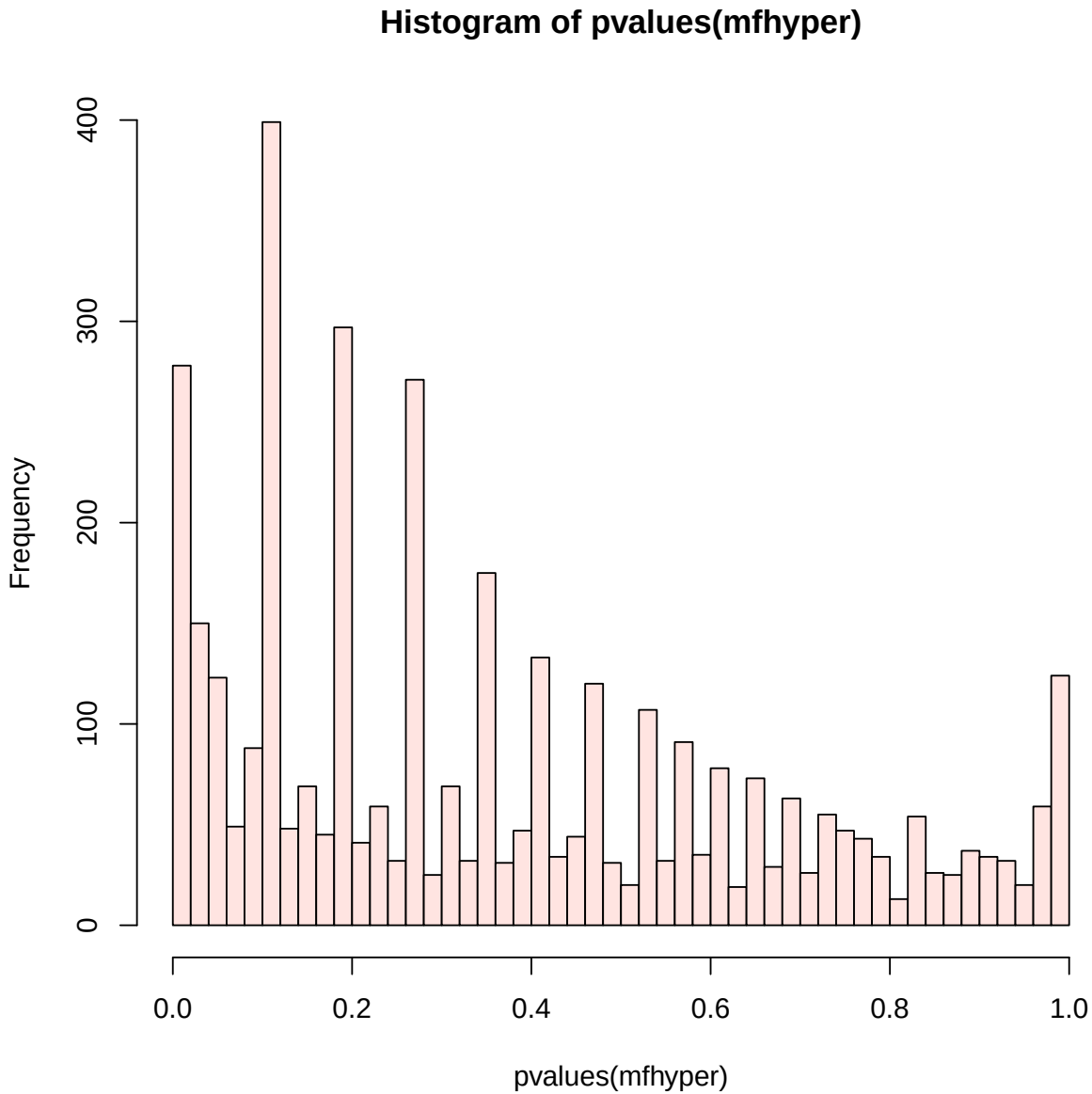
```
>sel = order(rt$p.value)[1:400]
>ALLsub = ALLfilt_af4bcr[sel,]
>EGsub =
as.character(hgu95av2ENTREZID[featureNames(ALLsub)])
>library("GOstats")
>affyUniverse = featureNames(ALLfilt_af4bcr)
>uniId = hgu95av2ENTREZID[affyUniverse]
>entrezUniverse = unique(as.character(uniId))
>params = new("GOHyperGParams", geneIds=EGsub,
universeGeneIds=entrezUniverse,
annotation="hgu95av2", ontology="BP",
pvalueCutoff=0.001, conditional=FALSE,
testDirection="over")
>mfhyper = hyperGTest(params)
>sum = summary(mfhyper, p=0.001)
>head(sum)
```

Hypergeometric Test

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size
1	GO:0007166	0.00	1.99	73.04	116	720
2	GO:0007165	0.00	1.80	134.92	185	1330
3	GO:0023052	0.00	1.77	145.27	195	1432
4	GO:0006955	0.00	2.20	43.01	77	424
5	GO:0007154	0.00	1.73	147.40	195	1453
6	GO:0048583	0.00	1.85	76.79	116	757

	GOBPID	Term
1	GO:0007166	cell surface receptor signaling pathway
2	GO:0007165	signal transduction
3	GO:0023052	signaling
4	GO:0006955	immune response
5	GO:0007154	cell communication
6	GO:0048583	regulation of response to stimulus

Hypergeometric Test



Enrichment Analysis

- In practice, we need search through thousands of GO terms to determine which GO terms is enriched in the selected genes $\rightarrow$ multiple comparisons
- However, tests are highly dependent
  - Hierarchical structure of the GO
  - Each gene can belong to multiple GO terms.
    e.g. human *HOXA7* gene belongs to four GO terms: "Development", "Nucleus", "DNA dependent regulation and transcription", "Transcription factor activity".

Conditional Hypergeometric Test

```
>hgCutoff = 0.001
>params = new("GOHyperGParams",
geneIds=selectedEntrezIds,
universeGeneIds=entrezUniverse,
annotation="hgu95av2.db", ontology="BP",
pvalueCutoff=hgCutoff, conditional=FALSE,
testDirection="over")
>paramsCond = params
>conditional(paramsCond) = TRUE
>hgCond = hyperGTest(paramsCond)
>sum = summary(mfhyper, p=0.001)
>head(sum)
```

Hypergeometric Test Using KEGG

```
>frame = toTable(org.Hs.egPATH)
>keggframeData = data.frame(frame$path_id,
frame$gene_id)
>head(keggframeData)
>keggFrame = KEGGFrame(keggframeData, organism =
"Homo sapiens")
>gsc <- GeneSetCollection(keggFrame, setType =
KEGGCollection())
> kparams <- GSEAKEGGHyperGParams(name = "My Custom
GSEA based annot Params", geneSetCollection = gsc,
geneIds = EGsub, universeGeneIds = entrezUniverse,
pvalueCutoff = 0.1, testDirection = "over")
> kOver <- hyperGTest(kparams)
> summary(kOver)
```

Hypergeometric Test Using KEGG

	KEGGID	P value	Odds Ratio	Exp Count	Count	Size	Term
1	04514	0.00	8.28	4.43	20	43	Cell adhesion molecules (CAMs)
2	05320	0.00	15.20	1.65	10	16	Autoimmune thyroid disease
3	04940	0.00	9.14	2.27	11	22	Type I diabetes mellitus
4	05332	0.00	9.14	2.27	11	22	Graft-versus-host disease
5	05330	0.00	10.11	1.96	10	19	Allograft rejection
6	05416	0.00	5.61	3.81	14	37	Viral myocarditis

Enrichment Analysis

- Simple and naive way:
  - Get p values from Fisher's exact test for all pathways
  - Correct by Benjamini-Hochberg procedure to control FDR
- Problem:
  - Fisher's test simplify DE statistics into $\{0, 1\} \rightarrow$ loss information
  - Does not consider gene dependence structure and pathway hierarchical dependence structure
- Improved Methods:
  - Use averaged t-statistics as the pathway-specific enrichment score.
  - Apply permutation test to get p values and FDR control

GSEA

The goal of GSEA is to detect modest but coordinated changes in prespecified sets of related genes. Such a set might include all the genes in a specific pathway, for instance.

$$
\begin{aligned}
z_K &= \frac{\sum_K t_k}{\sqrt{K}} \sim N(0,1) \\
\chi^2_g &= \frac{\sum_K (t_i - \bar{t})^2 - (K-1)}{2(K-1)} \\
\tilde{\chi^2_g} &= \sum_K t_i^2
\end{aligned}
$$

K: number of genes in the set

GSEA

```
>library("genefilter")
>ALLfilt_bcrneg = nsFilter(ALL_bcrneg,
var.cutoff=0.5)$eset
>table(ALLfilt_bcrneg$mol.biol)
BCR/ABL NEG
37 42
>library("GSEABase")
>gsc = GeneSetCollection(ALLfilt_bcrneg,
setType=KEGGCollection())
>Am = incidence(gsc)
>dim(Am)
>nsF = ALLfilt_bcrneg[colnames(Am),]
>rtt = rowttests(nsF, "mol.biol")
>rttStat = rtt$statistic
```

GSEA

```
>selectedRows = (rowSums(Am)>10)
>Am2 = Am[selectedRows, ]
>tA = as.vector(Am2 %*% rttStat)
>tAadj = tA/sqrt(rowSums(Am2))
>names(tA) = names(tAadj) = rownames(Am2)
>library(Category)
>set.seed(123)
>NPERM = 1000
>pvals = gseattperm(nsF, nsF$mol.biol, Am2, NPERM)
>pvalCut = 0.025
>lowC = names(which(pvals[, 1]<=pvalCut))
>highC = names(which(pvals[, 2]<=pvalCut))
```