Introduction to Annotation for Gene Expression Analyses

Utah State University – Spring 2014

STAT 5570: Statistical Bioinformatics

Notes 4.1

References

- Chapters 7 & 14 of Bioconductor Monograph (course text)
- www.geneontology.org
- www.genome.jp/kegg

Annotation

- In many organisms, the genome is “well-annotated”
- We have information about many genes’:
  - Molecular function, biological process, cellular component (GO terms)
  - Chromosomal location
  - KEGG pathways
- We can filter based on this information

  - like only test DE for genes involved in the induction of apoptosis, for example
- We can also look for “over-representation”

  - like maybe apoptosis-related genes are disproportionately abundant in list of DE genes

Recall gene filtering (Notes 3.2)

- Reduce multiple testing severity by focusing on a subset of genes
- Two types of filtering:
  - Non-specific:
    - look only at genes with certain properties in expression values
    - examples: pOverA, cv
  - Specific:
    - look only at genes “known” to satisfy some biological constraint
    - examples: biological processes, pathways

Annotation: GO

- Ontology
  - “a structured vocabulary that characterizes some conceptual domain”
  - a structured way of incorporating knowledge into a hierarchical classification system
  - a structured way to organize subject indices

- Gene Ontology (GO)
  - a system developed to organize “known” information about genes
  - www.geneontology.org – constantly updated information

Current categories of gene ontologies

- Molecular Function: (MF) = elemental task
  - tasks performed by individual gene products
- Biological Process: (BP) = biological objective
  - broad goals accomplished by ordered assemblies of molecular functions
- Cellular Component: (CC) = location
  - subcellular structures, locations, and complexes

How are these determined?

Annotation:

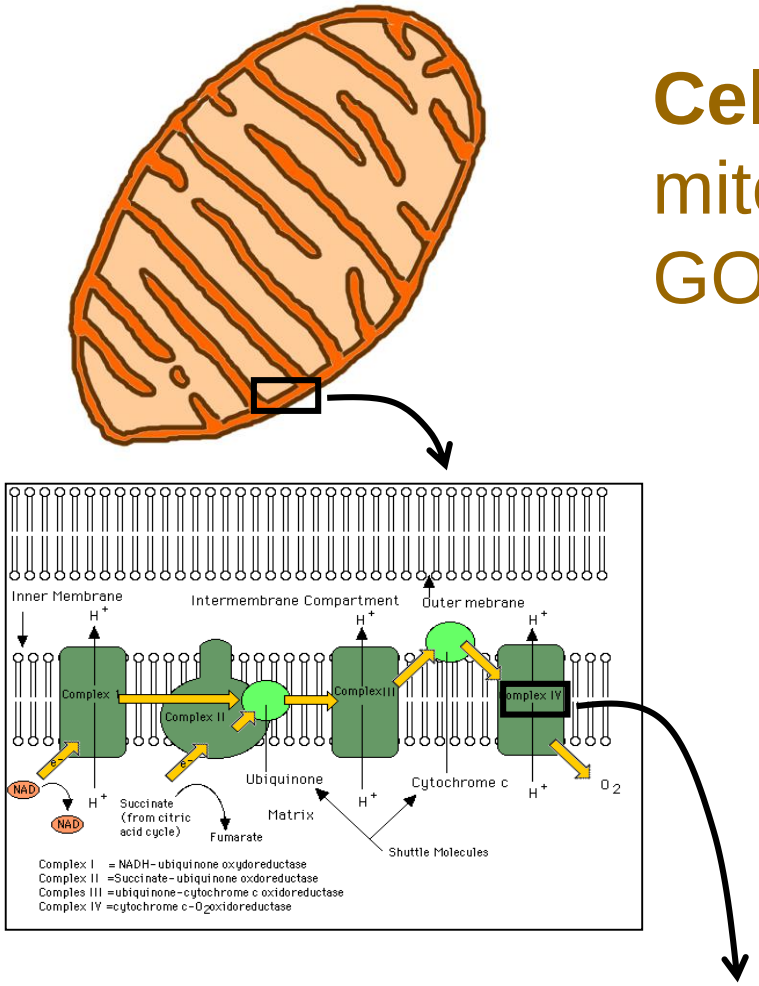
a statement that a gene product

- has a particular molecular function,
- is involved in a particular biological process, or
- is located within a certain cellular component,

} GO Term

as determined by a particular method, } Evidence Code

as described in a particular source } Reference



substrate + O_2 = CO_2 +H_20 product

Cellular component: mitochondrial inner membrane GO:0005743

Biological process: Electron transport GO:0006118

Molecular function: monooxygenase activity GO:0004497

(from a tutorial slide presentation by Suparna Mundodi at geneontology.org)

GO Evidence Codes

Code	Definition	
IEA	**I**nferred from **E**lectronic **A**nnotation	
NAS	**N**on-traceable **A**uthor **S**tatement	
TAS	**T**raceable **A**uthor **S**tatement	
ND	**N**o **D**ata	Use with annotation to unknown
IDA	**I**nferred from **D**irect **A**ssay	
*IPI	**I**nferred from **P**hysical **I**nteraction	
*IGI	**I**nferred from **G**enetic **I**nteraction	
IMP	**I**nferred from **M**utant **P**henotype	
IEP	**I**nferred from **E**xpression **P**attern	
*IC	**I**nferred from **C**urator	
*ISS	**I**nferred from **S**equence **S**imilarity	

Manually annotated

(from a tutorial slide presentation by Harold J. Drabkin at geneontology.org)

Ex: Find apoptosis-related GO terms

	GO.ID	GO.Term
1	GO:0003377	regulation of apoptosis by sphingosine-1-phosphate signaling pathway
2	GO:0006921	cellular component disassembly involved in execution phase of apoptosis
3	GO:0006922	cleavage of lamin involved in execution phase of apoptosis
4	GO:0006923	cleavage of cytoskeletal proteins involved in execution phase of apoptosis
5	GO:0030972	cleavage of cytosolic proteins involved in execution phase of apoptosis
6	GO:0060785	regulation of apoptosis involved in tissue homeostasis
7	**GO:0097194**	**execution phase of apoptosis**
8	GO:0097200	cysteine-type endopeptidase activity involved in execution phase of apoptosis
9	GO:0097297	activation of cysteine-type endopeptidase activity involved in execution phase of apoptosis
10	GO:1900117	regulation of execution phase of apoptosis
11	GO:1900118	negative regulation of execution phase of apoptosis
12	GO:1900119	positive regulation of execution phase of apoptosis
13	GO:2001270	regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis
14	GO:2001271	negative regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis
15	GO:2001272	positive regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis

Ex: Find apoptosis induction-related GO

```
# load necessary libraries; these are big packages
library(GO.db); library(AnnotationDbi)

# Define getGOTerms function; don't worry about syntax
getGOTerms <- function(term)
  { xx <- as.list(GOTERM) # takes about 30 seconds
    all.Terms <- lapply(xx,Term)  # get Term of each GOTERM element
    t <- grep(term,all.Terms) # get index where Term includes term
       # use agrep for approximate string matching
    GO.Term <- unlist(all.Terms[t])
    GO.ID <- names(GO.Term)
    GO.frame <- data.frame(GO.ID=GO.ID, GO.Term=GO.Term)
    rownames(GO.frame) <- 1:length(GO.ID)
    return(GO.frame)
   }

# Get data.frame summarizing all GO terms
# including the string 'apoptosis'
GO.frame <- getGOTerms('apoptosis')   # 15 terms
```

```
# Get information about one of the IDs
GOTERM$"GO:0097194 "
```

```
GOID: GO:0097194
Term: execution phase of apoptosis
Ontology: BP
Definition: A stage of the apoptotic process that starts with
 the controlled breakdown of the cell through the action of
 effector caspases or other effector molecules (e.g.
 cathepsins, calpains etc.). Key steps of the execution phase
 are rounding-up of the cell, retraction of pseudopodes,
 reduction of cellular volume (pyknosis), chromatin
 condensation, nuclear fragmentation (karyorrhexis), plasma
 membrane blebbing and fragmentation of the cell into apoptotic
 bodies. The execution phase ends when the cell has died.
Synonym: apoptosis
Synonym: execution phase of apoptotic process
```

Other GO actions

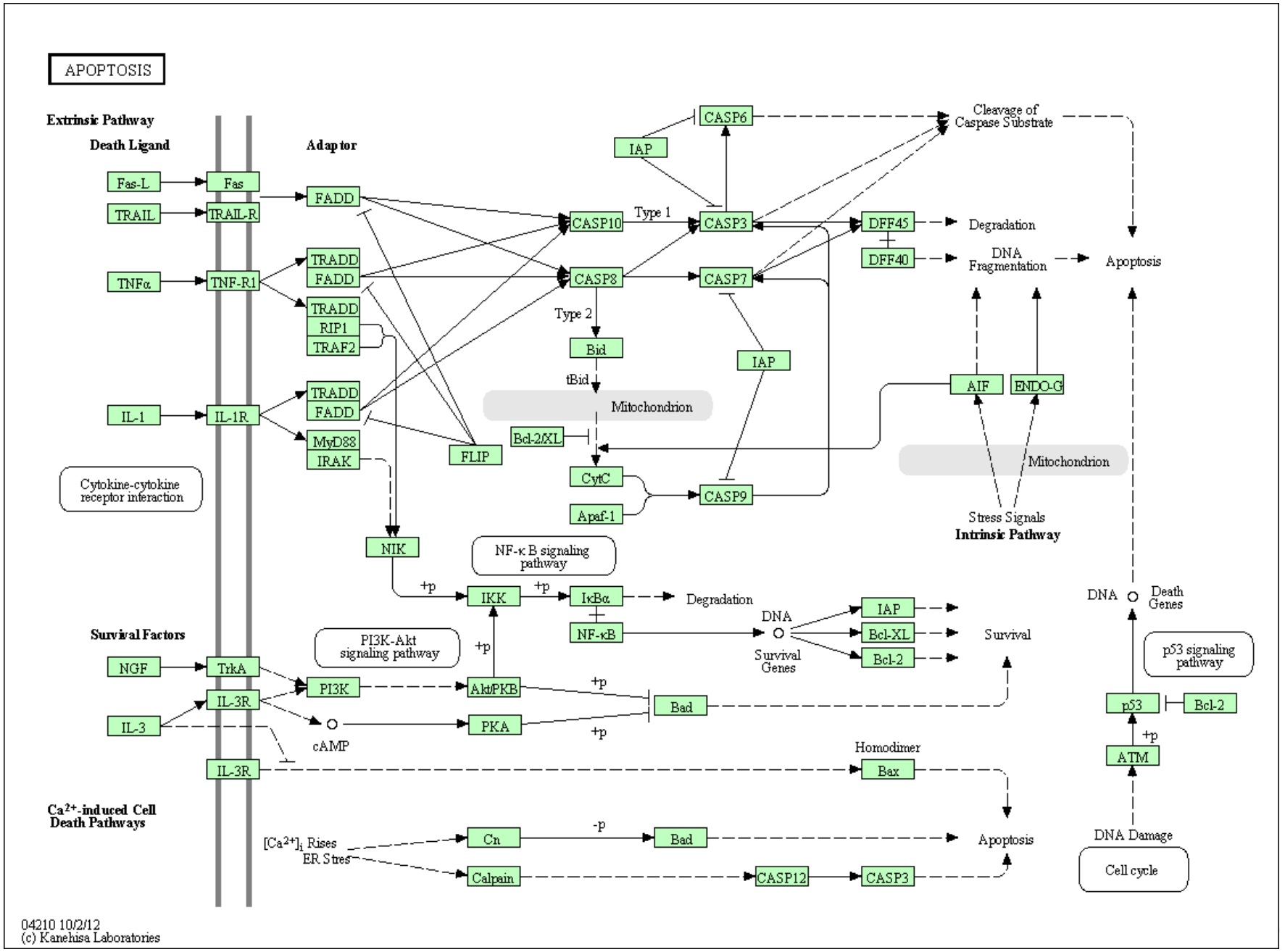
```
# find genes on the hgu95av2 array related to GO:0097194
library(hgu95av2.db) # load annotation Bioconductor meta-data
gn.apop <- hgu95av2GO2ALLPROBES$"GO:0097194" # 145 unique

# get only IDA-evidenced genes related to GO:0097194
apop <- gn.apop
t.non.NA <- !is.na(names(apop)); apop <- apop[t.non.NA]
t.IDA <- names(apop)=="IDA"
gn.apop.IDA <- apop[t.IDA]   # 61 unique

# find what GO terms are associated with a specific gene
GO.gn <-  names(hgu95av2GO$"33563_s_at")  # 23 unique
# "GO:0007155" "GO:0006919" ... "GO:0008191"

# get expr. levels for apoptosis-induction genes
library(affy); library(ALL); data(ALL)
gn <- featureNames(ALL)
t <- is.element(gn,gn.apop)
small.eset <- exprs(ALL)[t,]  # then test DE on these 145 genes
```

Annotation: KEGG

- Kyoto Encyclopedia of Genes and Genomes
  - grand challenge: computer representation of cell and organism, with ability to make predictions
  - based on genomic and molecular information

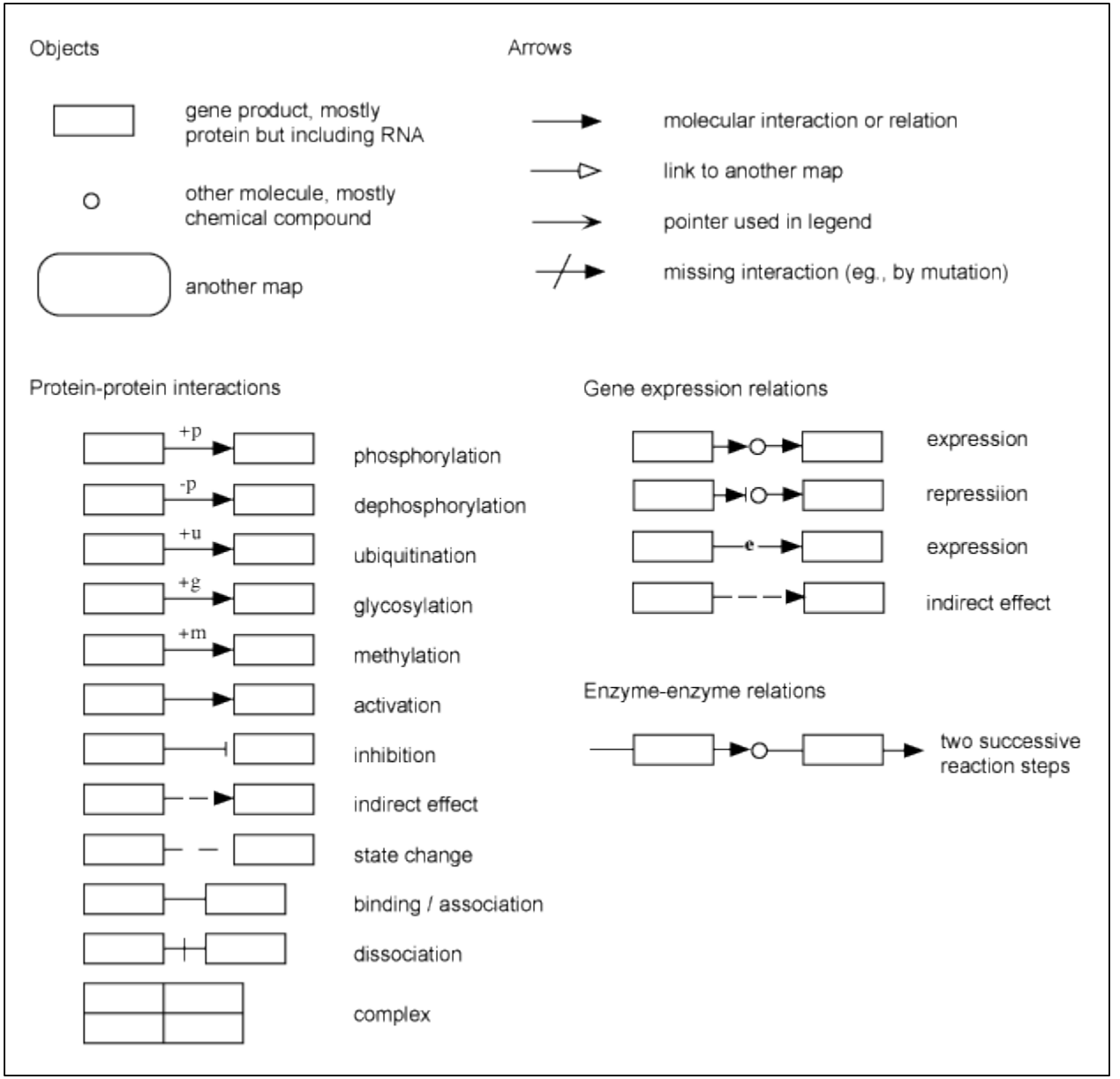
- Most often used for summarizing: "pathway" information

http://www.genome.jp/kegg/

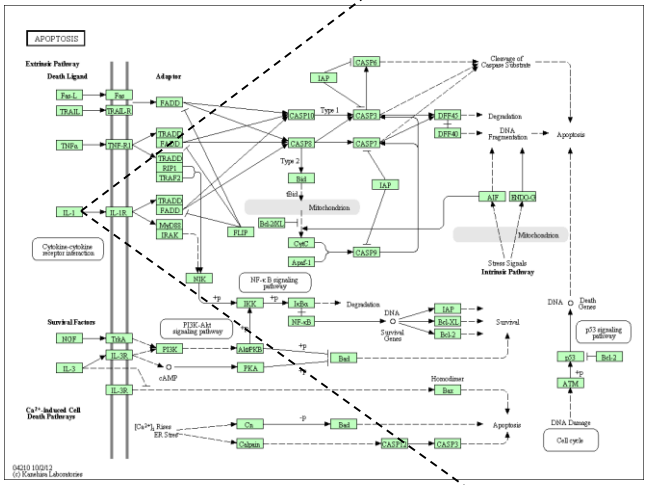


http://www.genome.jp/kegg/pathway/hsa/hsa04210.html

KEGG key



http://www.genome.jp/kegg/document/help_pathway.html



Homo sapiens (human): 3552

Help

Entry	3552 CDS T01001
Gene name	IL1A, IL-1A, IL1, IL1-ALPHA, IL1F1
Definition	interleukin 1, alpha
Orthology	K04383 interleukin 1 alpha
Organism	hsa Homo sapiens (human)
Pathway	hsa04010 MAPK signaling pathway hsa04060 Cytokine-cytokine receptor interaction hsa04210 Apoptosis hsa04380 Osteoclast differentiation hsa04640 Hematopoietic cell lineage hsa04940 Type I diabetes mellitus hsa05020 Prion diseases hsa05132 Salmonella infection hsa05133 Pertussis hsa05140 Leishmaniasis hsa05152 Tuberculosis hsa05162 Measles hsa05164 Influenza A hsa05323 Rheumatoid arthritis hsa05332 Graft-versus-host disease
Disease	H00084 Graft-versus-host disease
Drug target	Rilonacept: D06635

AA seq	271 aa AA seq DB search MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETS KTSKLTFKESMVVVATNGKVLKKRRLSLSQSITDDDLEAIANDSEEEIIKPRSAPFSFLS NVKYNFMRIIKYEFILNDALNQSIIRANDQYLTAAALHNLDEAVKFDMGAYKSSKDDAKI TVILRISKTQLYVTAQDEDQPVLLKEMPEIPKTITGSETNLLFFWETHGTKNYFTSVAHP NLFIATKQDYWVCLAGGPPSITDFQILENQA
NT seq	816 nt NT seq atggccaaagttccagacatgtttgaagacctgaagaactgttacagtgaaaatgaagaa gacagttcctccattgatcatctgtctctgaatcagaaatccttctatcatgtaagctat ggcccactccatgaaggctgcatggatcaatctgtgtctctgagtatctctgaaacctct aaaacatccaagcttaccttcaaggagagcatggtggtagtagcaaccaacgggaaggtt ctgaagaagagacggttgagtttaagccaatccatcactgatgatgacctggaggccatc gccaatgactcagaggaagaaatcatcaagcctaggtcagcacctttagcttcctgagc aatgtgaaatacaactttatgaggatcatcaaatacgaattcatcctgaatgacgccctc aatcaaagtataattcgagccaatgatcagtacctcacggctgctgcattacataatctg gatgaagcagtgaaatttgacatgggtgcttataagtcatcaaaggatgatgctaaaatt accgtgattctaagaatctcaaaaactcaattgtatgtgactgcccaagatgaagaccaa ccagtgctgctgaaggagatgcctgagatacccaaaaccatcacaggtagtgagaccaac ctcctcttcttctgggaaactcacggcactaagaactatttcacatcagttgcccatcca aacttgtttattgccacaaagcaagactactgggtgtgcttggcagggggccacctct atcactgactttcagatactggaaaaccaggcgtag

http://www.genome.jp/kegg/pathway/hsa/hsa04210.html

KEGG pathways and Bioconductor

- Packages to consider:
  - KEGG.db (now deprecated)
  - reactome.db
    - www.reactome.org (includes some nice online tools)
    - “Open-source, open-access, manually curated and peer-reviewed pathway database”

Ex: Find KEGG pathways related to apoptosis

```
# load pathway information
library(reactome.db)

# Define getreactomeTerms function; don't worry about syntax
getreactomeTerms <- function(term)
  { xx <- toTable(reactomePATHID2NAME)
    t <- grep(term,xx$path_name,ignore.case=TRUE)
       # use agrep for approximate string matching
    xx.t <- xx[t,]; rownames(xx.t) <- 1:length(t)
    return(xx.t)  }

# Find all reactome.db terms with a certain string
apop.frame <- getreactomeTerms('apoptosis')
apop.frame # 87 terms
```

```
     DB_ID                                        path_name
1   109581                          Homo sapiens: Apoptosis
2   109607   Homo sapiens: Extrinsic Pathway for Apoptosis
3   109606   Homo sapiens: Intrinsic Pathway for Apoptosis
...
85 5119942                  Arabidopsis thaliana: Apoptosis
86 5149897   Oryza sativa: Intrinsic Pathway for Apoptosis
87 5149898                          Oryza sativa: Apoptosis
```

```
# find human genes (on hgu95av2 array) that map to a pathway
library(hgu95av2.db) # load annotation Bioconductor meta-data

# Define getPathGenes function; don't worry about syntax,
# but note that this is for the hgu95av2 array version.
# Here, reactomeID is a vector of DB_ID's from reactome.db
getPathGenes <- function(reactomeID)
  {
    xx <- toTable(reactomeEXTID2PATHID)
    yy <- toTable(hgu95av2ENTREZID)
    t <- is.element(xx$DB_ID, reactomeID)
    entrez <- xx$gene_id[t]
    t <- is.element(yy$gene_id, entrez)
    return(yy$probe_id[t])
   }

gn.apop <- getPathGenes(109581) # 216 genes
gn.extrinsic.apop <- getPathGenes(109607) # 33 genes

# This is organism specific - no oryza genes on human array
gn.oryza.apop <- getPathGenes(5149898) # 0 genes
```

```
# find pathways for a given gene (on hgu95av2 array)
library(hgu95av2.db) # load annotation Bioconductor meta-data

# Define getGenePaths function; don't worry about syntax,
# but note that this is for the hgu95av2 array version.
# Here, gn is a vector of Affy probeset ID's
getGenePaths <- function(gn)
  { xx <- toTable(reactomeEXTID2PATHID)
    yy <- toTable(hgu95av2ENTREZID)
    t <- is.element(yy$probe_id, gn)
    entrez <- yy$gene_id[t]
    t <- is.element(xx$gene_id, entrez)
    reactomeIDs <- xx$DB_ID[t]
    zz <- toTable(reactomePATHID2NAME)
    t <- is.element(zz$DB_ID,reactomeIDs)
    out.z <- zz[t,]; rownames(out.z) <- 1:sum(t)
    return(out.z) }

gn.paths <- getGenePaths("33903_at") # 3 paths
gn.paths
#   DB_ID                                          path_name
#1 109581                                Homo sapiens: Apoptosis
#2 169911                  Homo sapiens: Regulation of Apoptosis
#3 418889 Homo sapiens: Role of DCC in regulating apoptosis
```

Summary

- GO and KEGG (Reactome)
  - prior biological “knowledge”
  - specific filtering
- Caution – this annotation information is:
  - imperfect
  - changing
- Looking ahead:
  - “over-representation” of GO terms