BiomedicalInformatics
3184 Graves Hall 333 W. 10th Ave. Columbus, OH 43210 614-292-4778 (phone) 614-292-7659 (fax)

Microarray Review, Clustering Ontology, Pathway and Enrichment Analysis

BMI 5730

Kun Huang
Department of Biomedical Informatics
The Ohio State University

Where do I get the gene list?

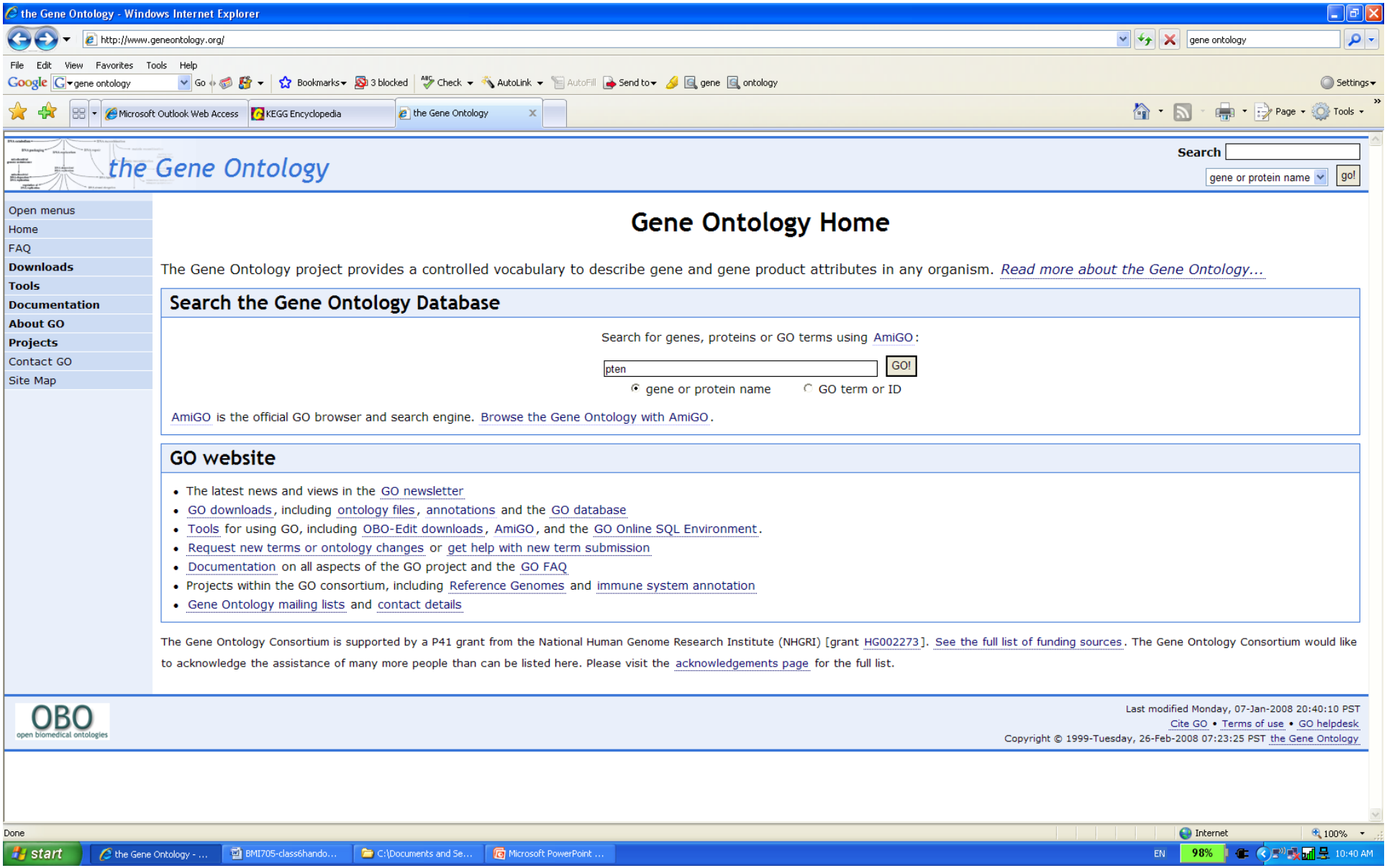
- Comparative study
  - e.g., microarray experiments between two types of samples or two disease states (can also be from RT-PCA, proteomics, …)
- Clustering / classification of genes
  - e.g., co-expressed genes
- Homologue analysis
  - e.g., genes from BLAST
- Other sources

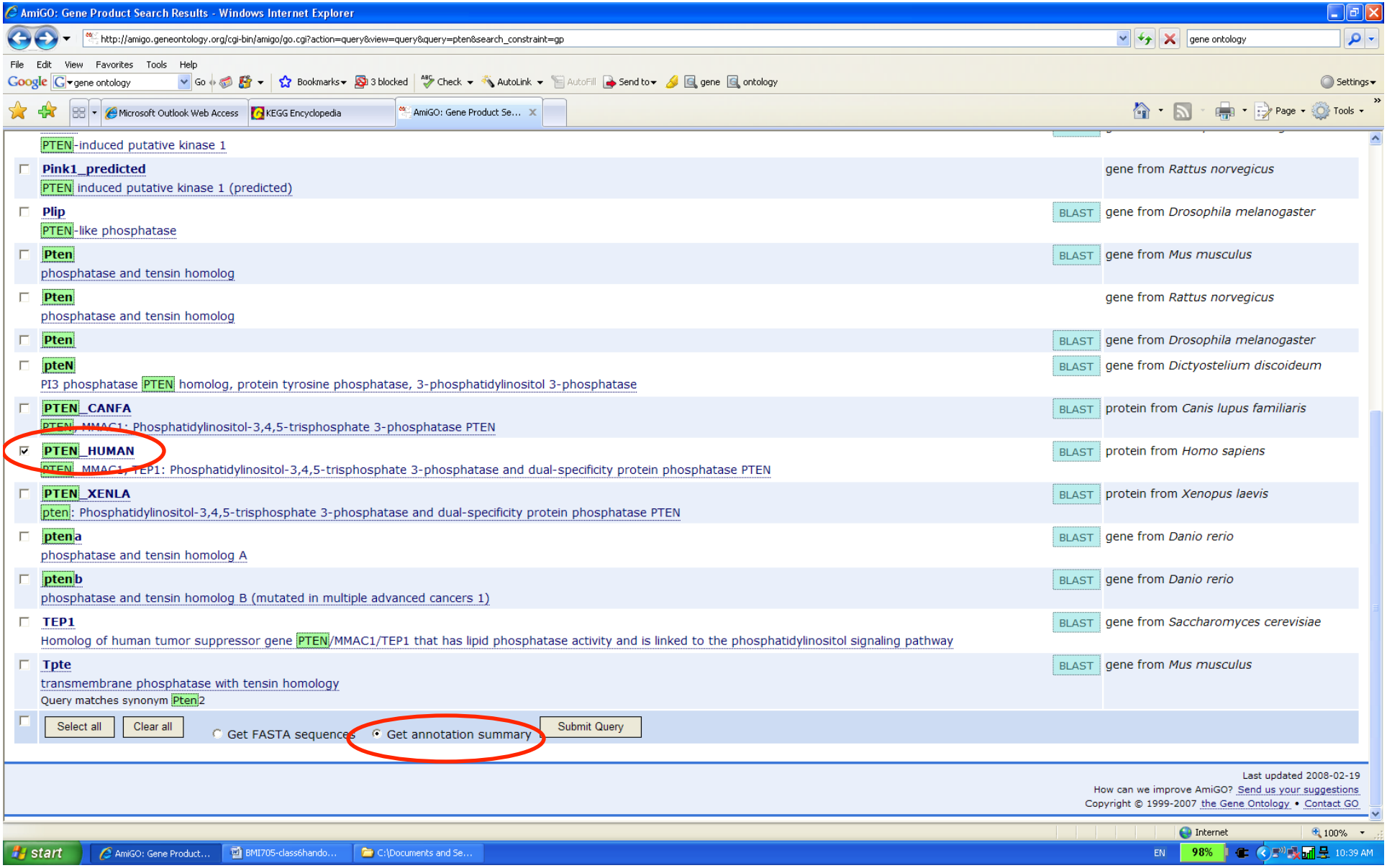
What do I do with the gene list – *enrichment (over-representation) analysis*?

- Find commonality among the gene
  - Common molecular functions (GO)
  - Common **biological processes** (GO)
  - Common cellular components (GO)
  - Common pathways
  - Interact with common genes
  - Common sequences / molecular structures
  - Regulated by common Transcription Factors
  - Targeted by common microRNAs
  - Involved in the same disease
  - …

Gene Ontology

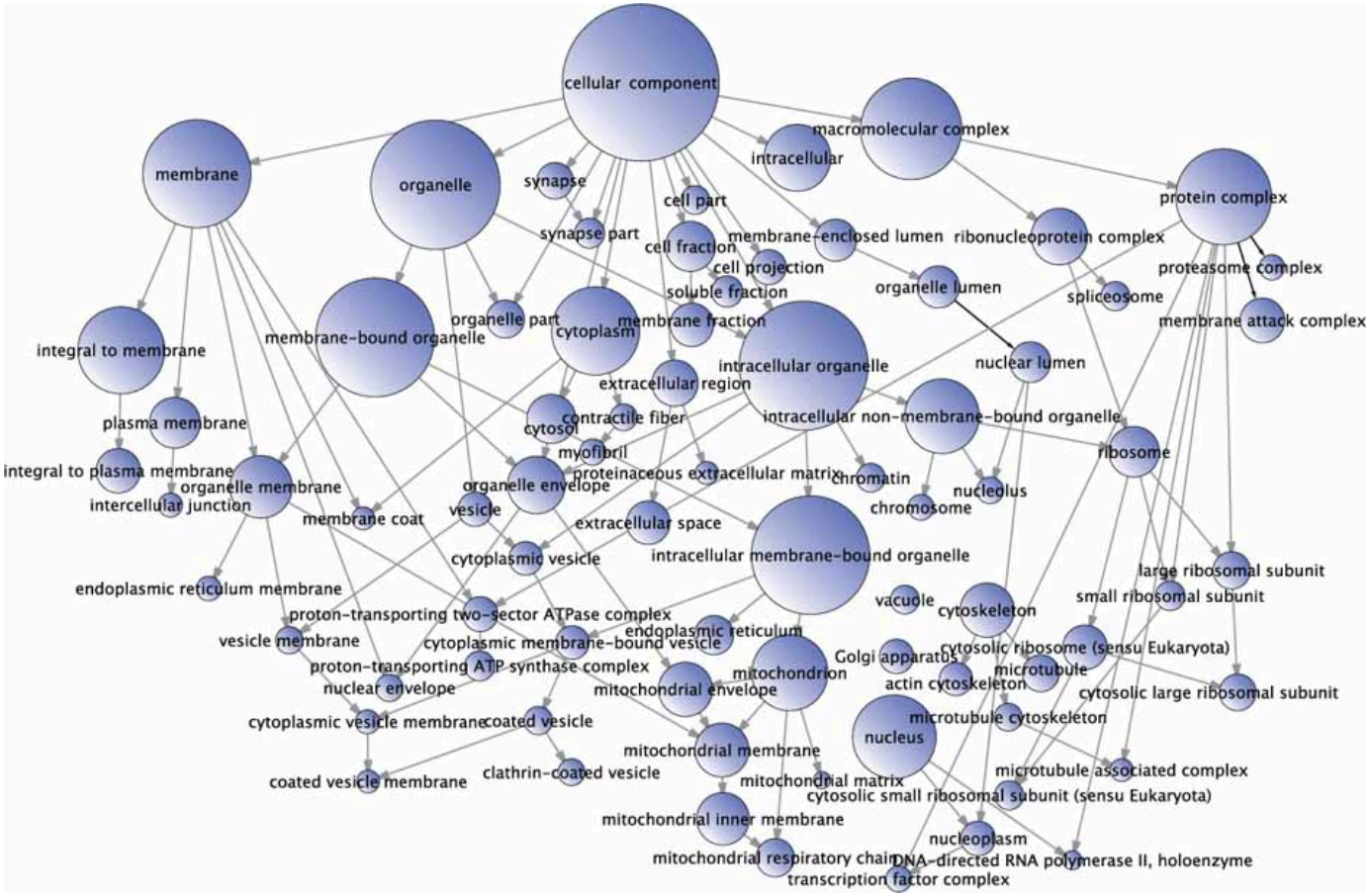
- Generate new hypothesis based on the commonality


the Gene Ontology - Windows Internet Explorer
http://www.geneontology.org/
gene ontology
File Edit View Favorites Tools Help
Microsoft Outlook Web Access
KEGG Encyclopedia
the Gene Ontology
the Gene Ontology
Search
gene or protein name
go!
Open menus
Home
FAQ
Downloads
Tools
Documentation
About GO
Projects
Contact GO
Site Map
Gene Ontology Home
The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Read more about the Gene Ontology...
Search the Gene Ontology Database
Search for genes, proteins or GO terms using AmiGO:
pten
GO!
gene or protein name
GO term or ID
AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.
GO website
The latest news and views in the GO newsletter
GO downloads, including ontology files, annotations and the GO database
Tools for using GO, including OBO-Edit downloads, AmiGO, and the GO Online SQL Environment.
Request new terms or ontology changes or get help with new term submission
Documentation on all aspects of the GO project and the GO FAQ
Projects within the GO consortium, including Reference Genomes and immune system annotation
Gene Ontology mailing lists and contact details
The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant HG002273]. See the full list of funding sources. The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the acknowledgements page for the full list.
OBO
open biomedical ontologies
Last modified Monday, 07-Jan-2008 20:40:10 PST
Cite GO • Terms of use • GO helpdesk
Copyright © 1999-Tuesday, 26-Feb-2008 07:23:25 PST the Gene Ontology
Done
Internet
100%
start
the Gene Ontology - ...
BMI705-class6hando...
C:\Documents and Se...
Microsoft PowerPoint ...
EN
98%
10:40 AM

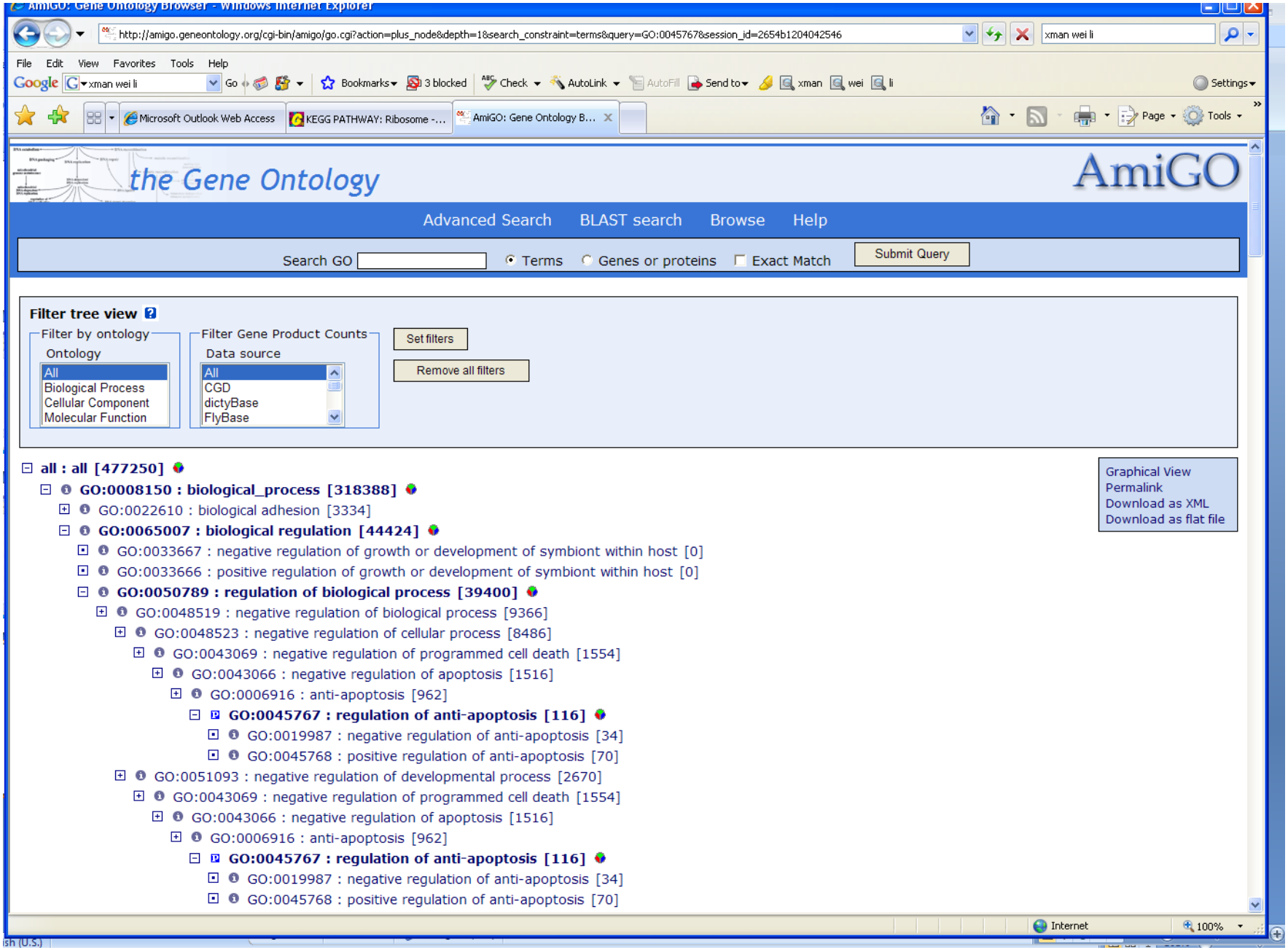

AmiGO: Gene Product Search Results - Windows Internet Explorer
http://amigo.geneontology.org/cgi-bin/amigo/go.cgi?action=query&view=query&query=pten&search_constraint=gp
gene ontology
File Edit View Favorites Tools Help
Microsoft Outlook Web Access
KEGG Encyclopedia
AmiGO: Gene Product Se...
PTEN-induced putative kinase 1
Pink1_predicted
PTEN induced putative kinase 1 (predicted)
gene from Rattus norvegicus
Plip
PTEN-like phosphatase
BLAST
gene from Drosophila melanogaster
Pten
phosphatase and tensin homolog
gene from Mus musculus
gene from Rattus norvegicus
gene from Drosophila melanogaster
pteN
PI3 phosphatase PTEN homolog, protein tyrosine phosphatase, 3-phosphatidylinositol 3-phosphatase
gene from Dictyostelium discoideum
PTEN_CANFA
PTEN_MMAC1: Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase PTEN
protein from Canis lupus familiaris
PTEN_HUMAN
PTEN_MMAC1_TEP1: Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN
protein from Homo sapiens
PTEN_XENLA
pten: Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN
protein from Xenopus laevis
ptena
phosphatase and tensin homolog A
gene from Danio rerio
ptenb
phosphatase and tensin homolog B (mutated in multiple advanced cancers 1)
gene from Danio rerio
TEP1
Homolog of human tumor suppressor gene PTEN/MMAC1/TEP1 that has lipid phosphatase activity and is linked to the phosphatidylinositol signaling pathway
gene from Saccharomyces cerevisiae
Tpte
transmembrane phosphatase with tensin homology
Query matches synonym Pten2
gene from Mus musculus
Select all
Clear all
Get FASTA sequences
Get annotation summary
Submit Query
Last updated 2008-02-19
How can we improve AmiGO? Send us your suggestions
Copyright © 1999-2007 the Gene Ontology • Contact GO
start
AmiGO: Gene Product...
BMI705-class6hando...
C:\Documents and Se...
EN
98%
10:39 AM

Gene Ontology

- Controlled vocabulary
- Defined relationship between terms (e.g., “part of”, “is a”, “regulates”)
- A “DAG”, not a “tree” – directed acyclic graph



http://www.biomedcentral.com/content/supplementary/1471-2164-9-96-s6.jpeg

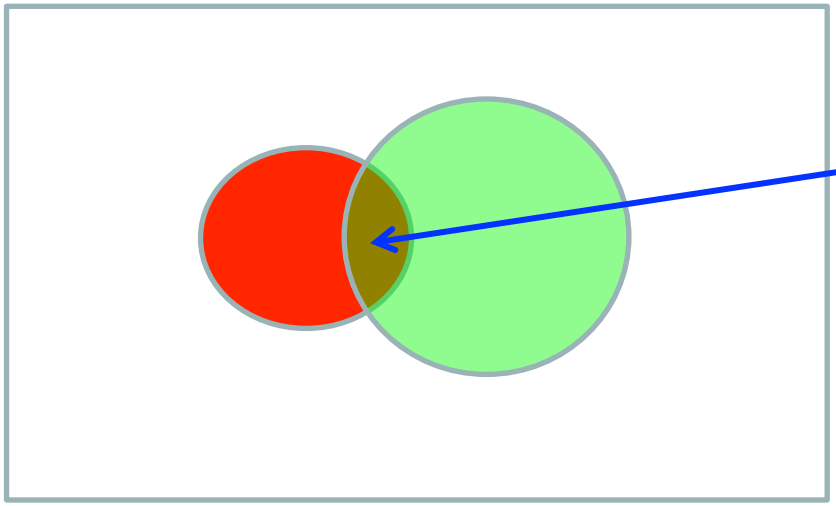

AmiGO: Gene Ontology Browser - Windows Internet Explorer
http://amigo.geneontology.org/cgi-bin/amigo/go.cgi?action=plus_node&depth=1&search_constraint=terms&query=GO:0045767&session_id=2654b1204042546
xman wei li
File Edit View Favorites Tools Help
Microsoft Outlook Web Access
KEGG PATHWAY: Ribosome -...
AmiGO: Gene Ontology B...
the Gene Ontology
AmiGO
Advanced Search BLAST search Browse Help
Search GO
Terms
Genes or proteins
Exact Match
Submit Query
Filter tree view
Filter by ontology
Ontology
All
Biological Process
Cellular Component
Molecular Function
Filter Gene Product Counts
Data source
All
CGD
dictyBase
FlyBase
Set filters
Remove all filters
all : all [477250]
GO:0008150 : biological_process [318388]
GO:0022610 : biological adhesion [3334]
GO:0065007 : biological regulation [44424]
GO:0033667 : negative regulation of growth or development of symbiont within host [0]
GO:0033666 : positive regulation of growth or development of symbiont within host [0]
GO:0050789 : regulation of biological process [39400]
GO:0048519 : negative regulation of biological process [9366]
GO:0048523 : negative regulation of cellular process [8486]
GO:0043069 : negative regulation of programmed cell death [1554]
GO:0043066 : negative regulation of apoptosis [1516]
GO:0006916 : anti-apoptosis [962]
GO:0045767 : regulation of anti-apoptosis [116]
GO:0019987 : negative regulation of anti-apoptosis [34]
GO:0045768 : positive regulation of anti-apoptosis [70]
GO:0051093 : negative regulation of developmental process [2670]
GO:0043069 : negative regulation of programmed cell death [1554]
GO:0043066 : negative regulation of apoptosis [1516]
GO:0006916 : anti-apoptosis [962]
GO:0045767 : regulation of anti-apoptosis [116]
GO:0019987 : negative regulation of anti-apoptosis [34]
GO:0045768 : positive regulation of anti-apoptosis [70]
Graphical View
Permalink
Download as XML
Download as flat file
Internet
100%

Gene Ontology

- Annotation for genes
- Manually curated
- Electronic annotation
- ND – known unknowns (https://www.youtube.com/watch?v=GiPe1OiKQuk)

How do I find commonality from my gene list?

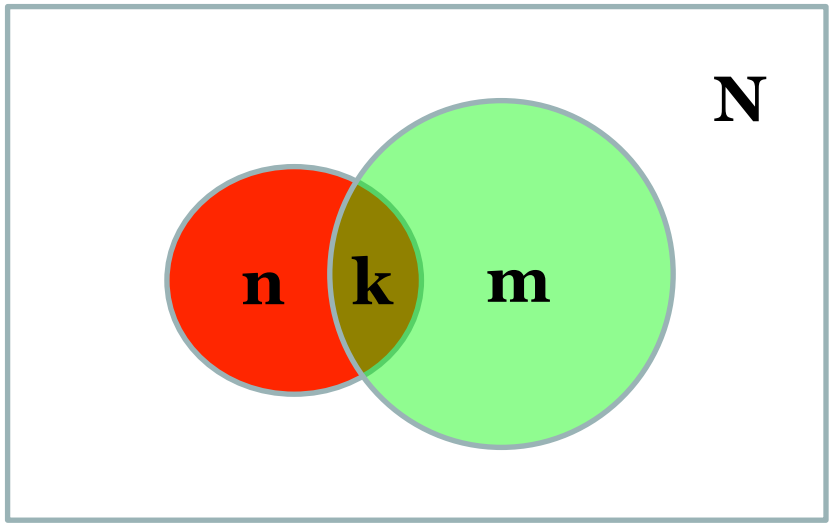
- Using a priori knowledge (e.g., gene ontology, pathway, annotation, etc.)
- Chi-square, Fisher's exact test, hypergeometric test, Bayesian-based methods, etc.



How significant is the intersection?

Fisher exact test

- Related to Chi-square test
- Hypergeometric distribution



Contingency table

	Red	Not red	Total
Green	k	$m-k$	m
Not green	$n-k$	$N-m-n+k$	$N-m$
Total	n	$N-n$	N

Probability of ***k*** successes (or red) in ***m*** draws without replacement, from a population of size ***N*** containing a total of ***n*** successes (or red).

Hypergeometric test

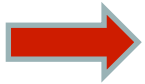
$$P(X=k) = \frac{(n\,!\,k) \times (N-n\,!\,m-k)}{(N\,!\,m)}$$

$$P(X \geq k) = \sum_{X=k}^{} \frac{(n\,!\,k) \times (N-n\,!\,m-k)}{(N\,!\,m)}$$

Sum of all p(k, n, N, m) for more extreme k's for overrepresentation

Are there any GO terms that have a larger than expected subset of our selected genes in their annotation list?

In GO (Green) vs. not in GO (No green)
Selected (Red) vs. not selected (Not red)



Is the GO term (i.e. pathway) significantly enriched?

Gene set enrichment analysis (GSEA)

- Available at Broad Institute (http://www.broadinstitute.org/gsea/index.jsp)
- Biological functions/pathways in common between gene sets
- To determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction
- Need ranking (e.g., based on p-value or fold change)
- Enrichment score: Kolmogorov-Smirnov/ranking statistics

Reflecting the degree to which a set S is overrepresented

at the extremes (top or bottom) of the entire ranked list L

- Permutation test for p-value
  - Permute labels
  - Permute values

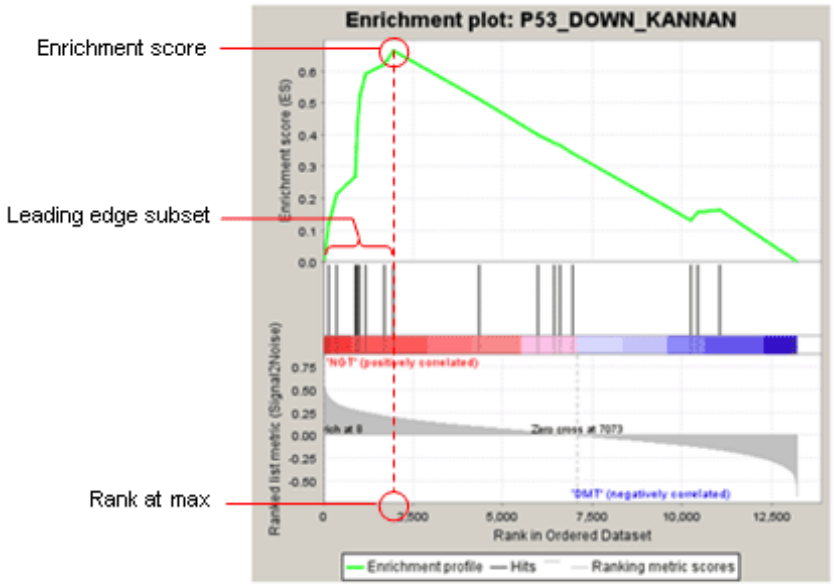
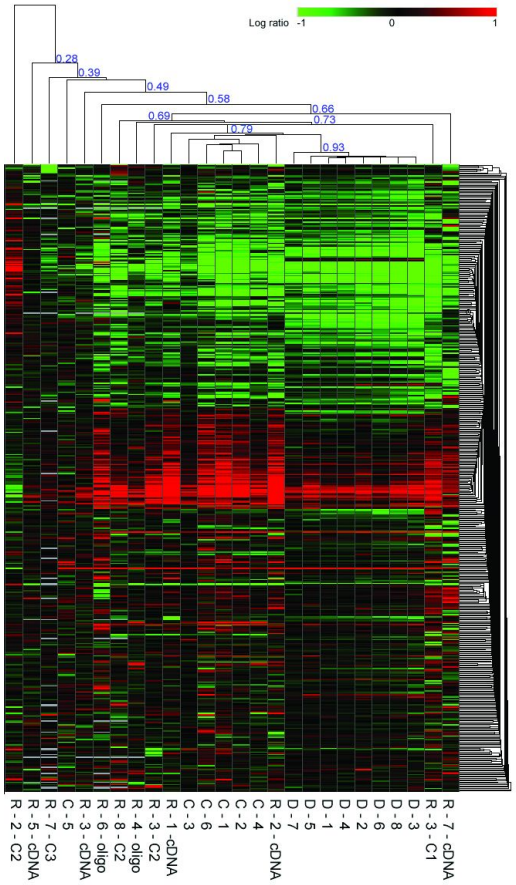


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

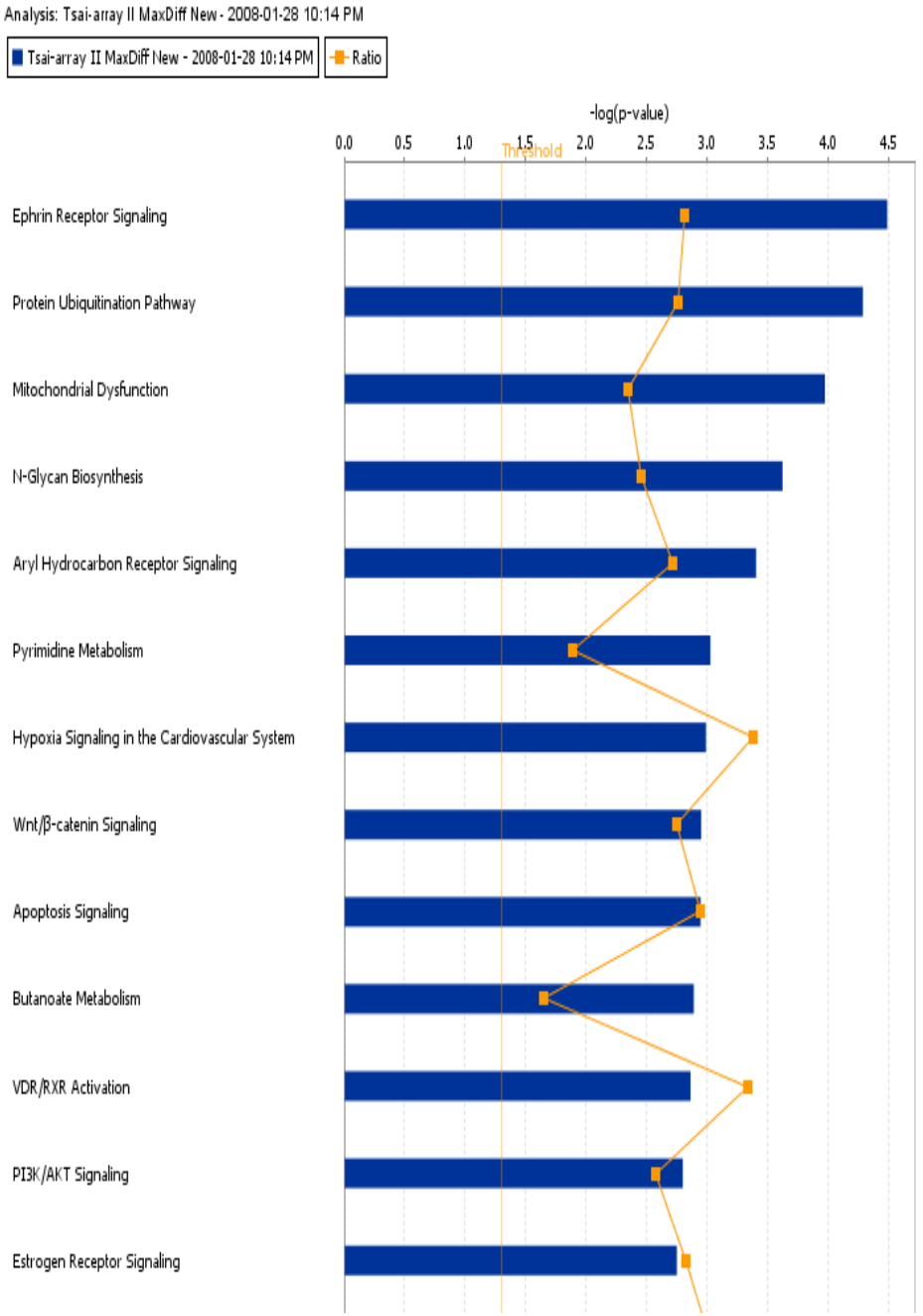
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1239896/pdf/pnas-0506580102.pdf
Subramanian A et al. PNAS 2005;102:15545-15550

What softwares are available?

- DAVID (http://david.abcc.ncifcrf.gov/)
- TOPPGene
- Cytoscape
  - GOTerm
  - BiNGO
- GSEA
- GenMapp (Free)
- Pathway Studio (Commercial)
- MetaCore (Commercial)
- Ingenuity Pathway Analysis (Commercial)
  - Manually curated
  - On-demand computation



Supplementary Figure 1: Clustering of laboratory/platform combinations using log ratio values of commone genes



Genes

Functions, pathways and networks

Demo

- DAVID (http://david.abcc.ncifcrf.gov/)
- TOPPGene (https://toppgene.cchmc.org/)
- Ingenuity Pathway Analysis (http://www.ingenuity.com)

Gene List1: AURKA BIRC5 ASPM BUB1 CCNA2 CCNB2 CDC2
ACOT7 CDC20 CDC45L CDCA8 CENPE CENPF CEP55 CKS2
CHEK1 DKFZp762E1312 DLG7 DNA2L E2F8 EPR1
FANCI HMMR KIF4A LMNB1 MAD2L1 MELK
NCAPG RANBP1 RRM2 SPAG5 STIL TACC3 TPX2
TRIP13 TTK UBE2C UBE2S

Gene List2: ABCF1 ABHD16A AK2 AK3 BAG6 BAIAP2 C6orf136 C6orf15
CCHCR1 DOM3Z DONSON EGLN3 EHMT2 ELOVL5 FBXO9
FBXW11 GTF2H4 LOC154761 MDC1 MOCOS MRPS10 MRPS18B
MRPS18C NFYA PPP1R11 RDBP RNF5 RPL7L1 RPS10-NUDT3///
NUDT3 SLC39A4///SLC39A7 SLC52A2 TOMM6///PRICKLE4
TOMM70A TRIM27 UBR2 VARS ZNF239 ZNF451 ZNRD1 ZNRF1

ID Mapping/Conversion

- Very challenging (and annoying)
- DAVID (http://david.abcc.ncifcrf.gov/)


KEGG PATHWAY Database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.kegg.com/kegg/pathway.html
Go Links
Google Go Bookmarks 4905 blocked Check AutoLink AutoFill Send to Settings
KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions, and relations
http://www.genome.jp/kegg/pathway.html
KEGG2 KID PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET
Pathway Maps
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:
1. Metabolism
Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid
Glycan PK/NRP Cofactor/vitamin Secondary metabolite Xenobiotics
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Human Diseases
and also on the structure relationships (KEGG drug structure maps) in:
6. Drug Development
Search PATHWAY for Go Clear
bfind mode bget mode
1. Metabolism
1.1 Carbohydrate Metabolism
Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Starch and sucrose metabolism
Aminosugars metabolism
Nucleotide sugars metabolism
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
Inositol metabolism
Inositol phosphate metabolism
KEGG Orthology (KO)
KEGG pathway modules
Overview of biosynthetic pathways
Enzymes (+diseases)
Compounds with biological roles
1.2 Energy Metabolism
Oxidative phosphorylation Revised!
Photosynthesis Revised!
Photosynthesis - antenna proteins New!
Photosynthesis proteins
Internet

Tryptophan metabolism - Reference pathway - Microsoft Internet Explorer

File Edit View Favorites Tools Help

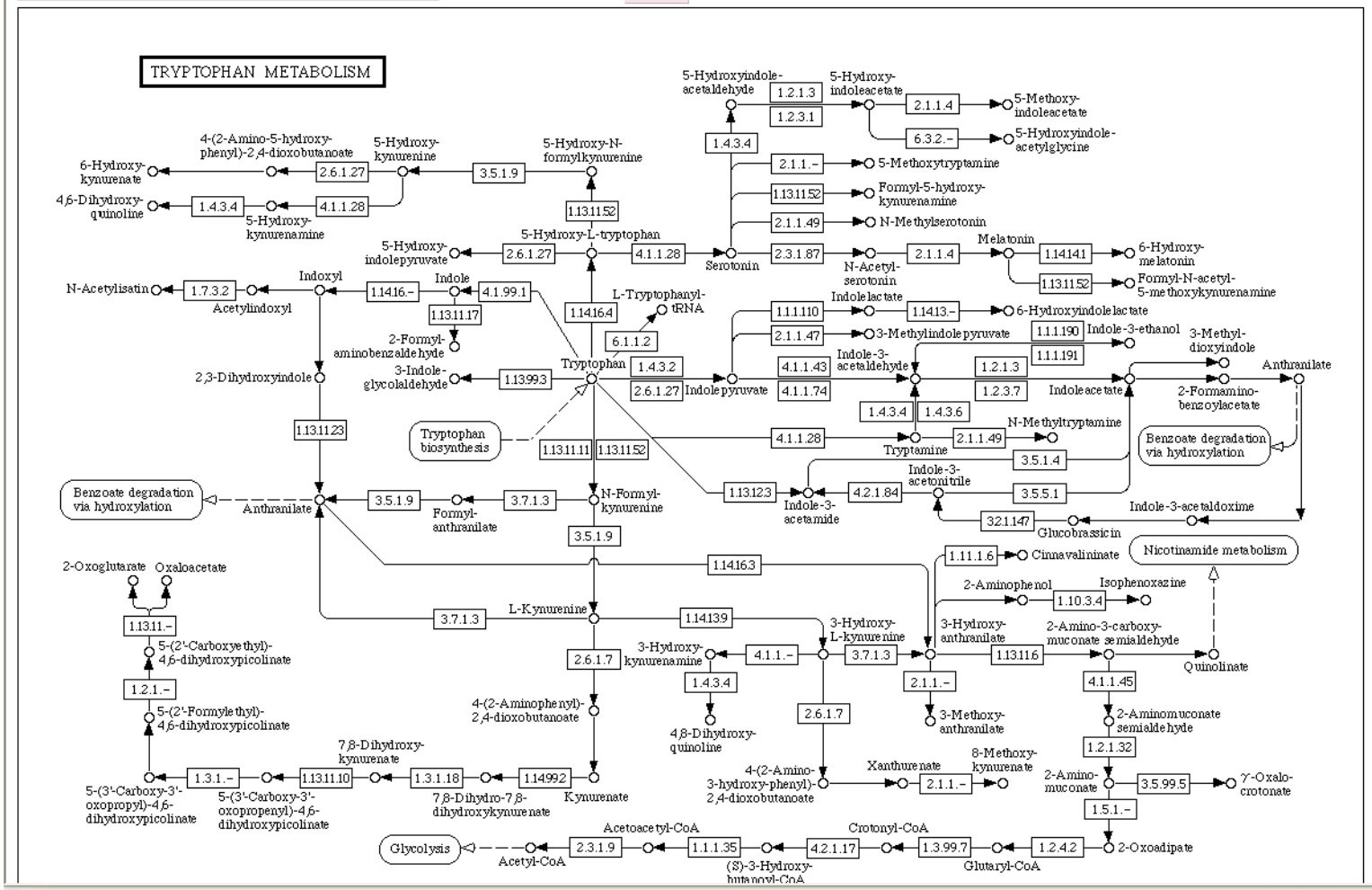
Address http://www.kegg.com/kegg/pathway/map/map00380.html

Tryptophan metabolism - Reference pathway

[Pathway menu]

Reference pathway Go Current selection Select

Help

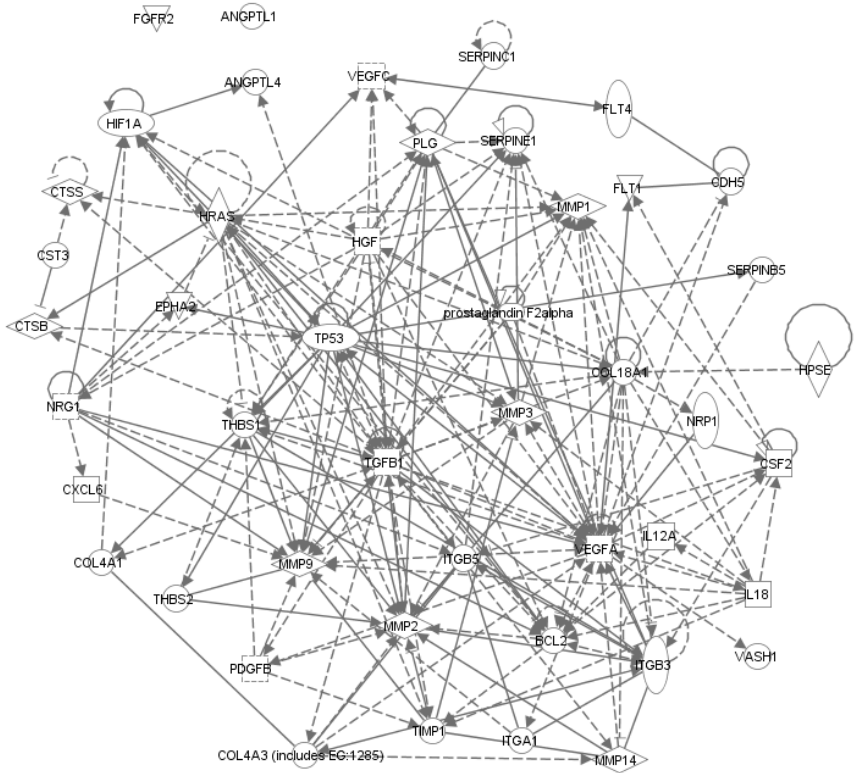


Internet

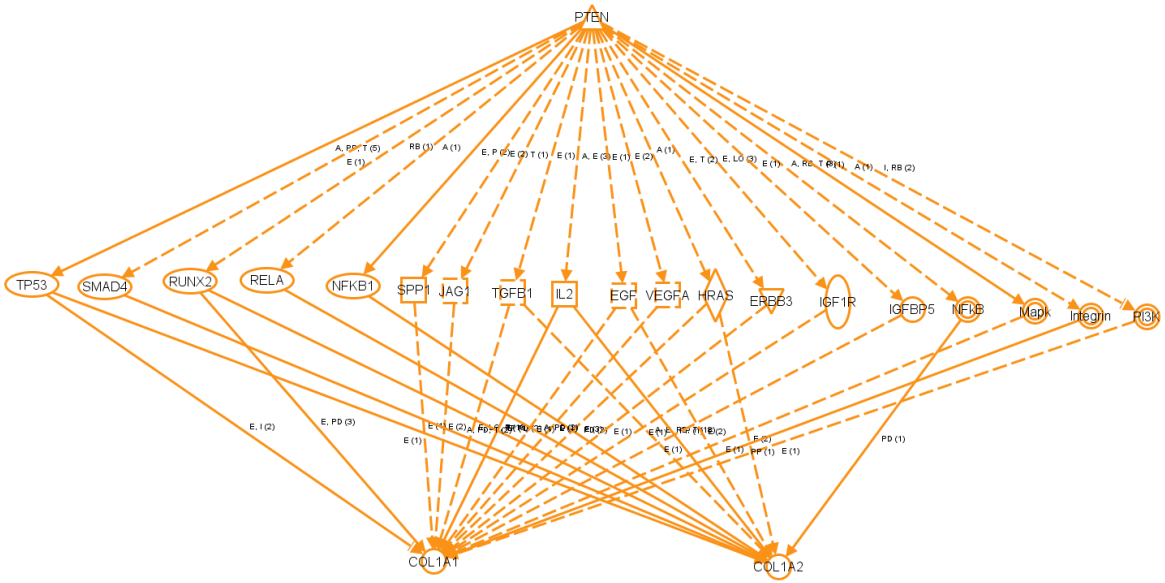
Ingenuity Pathway Analysis (IPA)

- Tumor Morphology — 46
  - angiogenesis — 46
    - angiogenesis [neoangiogenesis] of tumor — 46

ANGPTL1, ANGPTL4, BCL2, CDH5, COL18A1, COL4A1, COL4A3 (includes EG:1285), CSF2, CST3, CTSB, CTSS, CXCL6, EPHA2, FGFR2, FLT1, FLT4, HGF, HIF1A, HPSE, HRAS, IL18, IL12A, ITGA1, ITGB3, ITGB5, MMP1, MMP2, MMP3, MMP9, MMP14, NRG1, NRP1, PDGFB, PLG, prostaglandin F2alpha, SERPINB5, SERPINC1, SERPINE1, TGFB1, THBS1, THBS2, TIMP1, TP53, VASH1, VEGFA, VEGFC



New Pathway 6



http://www.ingenuity.com/

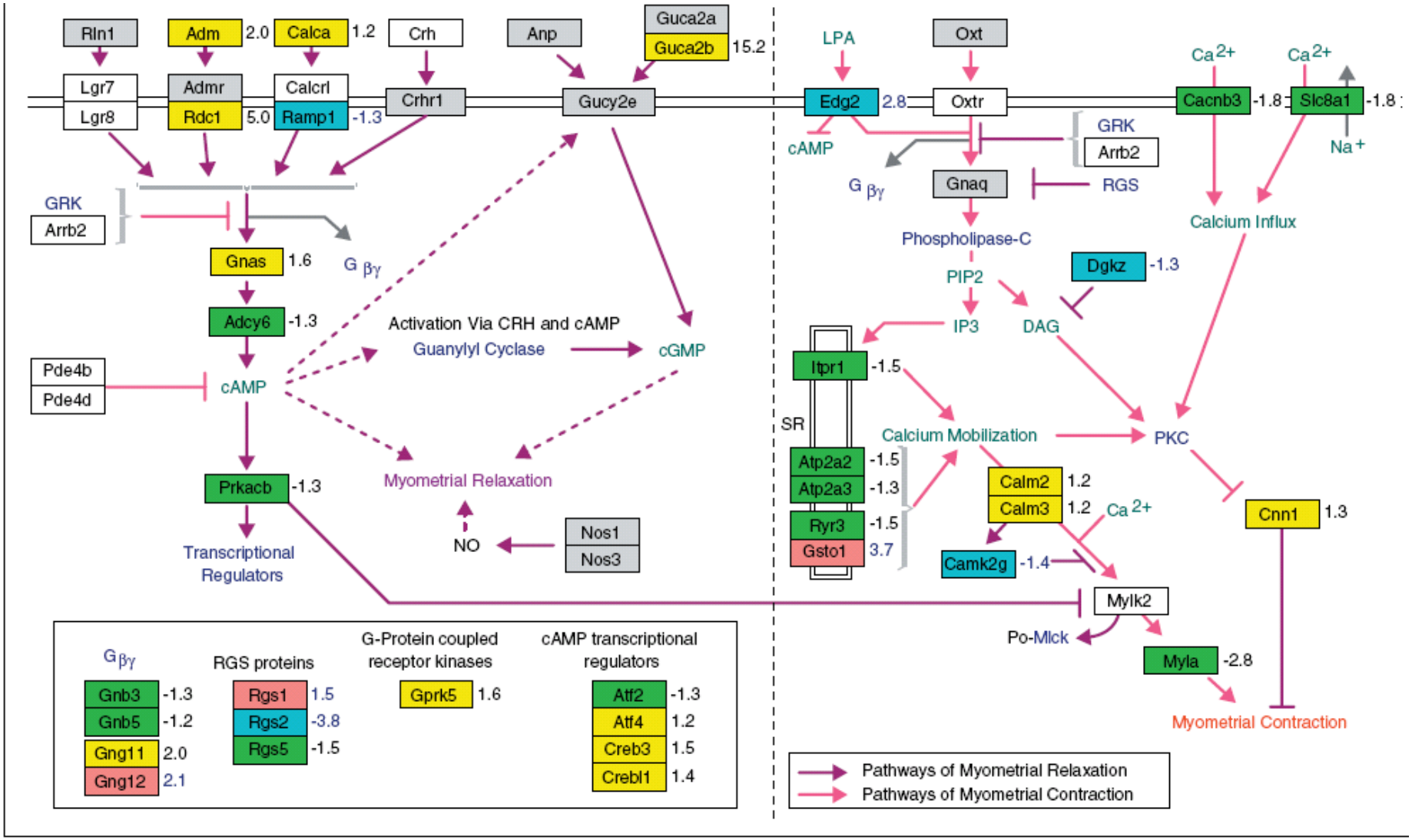


Figure 3

Analysis of pathways of uterine smooth muscle contraction. **(a)** Prostaglandin synthesis and **(b)** G-protein signaling pathways in the myometrium are overlaid with gene-expression color criterion and fold-changes from the program GenMAPP. Interactions suggested by results of this microarray analysis are included in these figures. Detailed gene-expression data, statistics and full gene annotations are available on the GenMAPP interactive version of these pathways online [40].

Pathway – What's out there?