Microarray Lab: GEO, Clustering, Heat Maps, and Functional Groups

Things to do:
1. Locate Data on GEO
2. Find Datasets for your **Gene/Disease** of Interest
3. Find enriched GO terms of genes that cluster with your gene in particular experiment using PANTHER

*NOTE this data is very specific to a type of experiment that has been previously published. If you can find these datasets that come from an actual paper you have found already on your gene/disease then look for that paper in Pubmed and link to GEO to find the dataset. The goal of this lab is to show you that RNAseq data is publicly available and that you can obtain gene expression information easily for each experiment that has been performed.
*_A REMINDER that your model organism homologs will have a LOT more data information here._

Part 1: Locate Interesting Data on GEO (Gene Expression Omnibus)
   *About GEO: http://www.ncbi.nlm.nih.gov/projects/geo/info/overview.html
GEO Profiles: http://www.ncbi.nlm.nih.gov/geoprofiles/
GEO DataSets: https://www.ncbi.nlm.nih.gov/gds/

1. Try searching for your gene of interest in the **Gene Profiles** query. Remember more research will be on your MODEL organisms. Try **ataxin-2** you will find the top hit is Drosophila. There is the control expression and in the Enigma mutant, ataxin-2 levels are decreased. This might be important to you if you know what Enigma is. Now check for the Human homolog experiments for **ATX2.**

A2bp1 - Enigma mitochondrial protein knockdown
1. Annotation: A2bp1, **Ataxin-2** binding protein 1
   Organism: Drosophila melanogaster
   Reporter: GPL72, 153863_at (ID_REF), GDS1662, 39198 (Gene ID), FBgn0036087 (CLONE ID)
   DataSet type: Expression profiling by array, transformed count, 6 samples
   ID: 17415764
   GEO DataSets   Gene   Profile neighbors   Chromosome neighbors

   Links from GEO Profiles

   Enigma mitochondrial protein knockdown
   Analysis of Kc-167 cells following knockdown of Enigma (Egm) using RNAi. Egm is a mitochondrial protein with homology to enzymes of the beta-oxidation of fatty acids. Results provide insight into the mechanisms underlying the increased oxidative stress tolerance conferred by decreased Egm levels.
   Organism: Drosophila melanogaster
   Type: Expression profiling by array, transformed count, 2 protocol sets
   Platform: GPL72   Series: GSE3566 6 Samples
   Download data: CEL
   DataSet   Accession: GDS1662   ID: 1662
   PubMed   Full text in PMC   Similar studies   GEO Profiles   Analyze DataSet
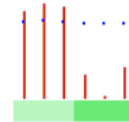
2. Go to GEO DataSets, search for **your condition (**ie breast cancer**)** *How to cluster Heat maps: https://www.ncbi.nlm.nih.gov/geo/info/datasets.html

Part 2: Hierarchically Cluster Data
Directions: Find a **disease dataset (example, breast cancer)** that has a **cluster/heatmap image** next to it, and click on that symbol. This will open the GEO Dataset browser for that dataset. (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser/)

> **Try clustering the results using the "Cluster Heatmap" tab** (lower left hand tab after you click the heatmap). **For Hierarchical clustering, the Distance setting should be set to "Pearson Correlation" and the Linkage setting should be set to "Average/UPGMA". Once these settings are set, press the "Display" button. To see the entire list of genes found in this experiment, select the entire heat map with the box and double click. This will give you the list of genes on the right hand side.

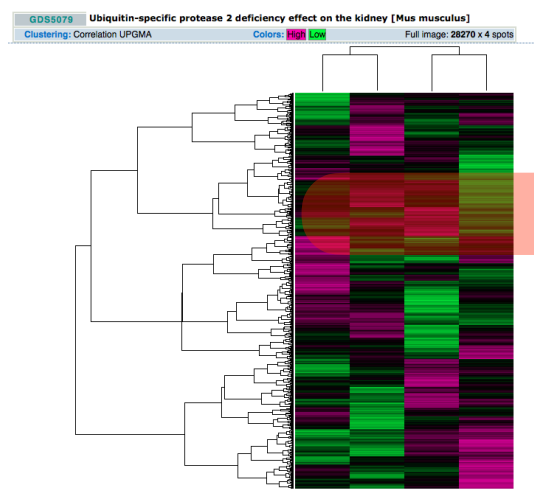80 DataSet records   << | < | Page 2 of 4 | > | >>

DataSet	Title	Organism(s)	Platform	Series	▸ Samples
GDS4296	NCI-60 cancer cell line panel	Homo sapiens	GPL570	GSE32474	174
GDS4214	Simian immunodeficiency virus encephalitis model of HIV-associated dementia: hippocampus	Macaca mulatta	GPL3535	GSE13824	36
GDS4153	Alpha-synuclein deficiency effect on striatum and cerebellum: time course	Mus musculus	GPL1261	GSE19534	26
GDS4152	Multiple sclerosis patients with high serum semaphorin4A levels: peripheral blood cells	Homo sapiens	GPL570	GSE26484	10
GDS4012	Progressive diabetic neuropathy: sural nerve	Homo sapiens	GPL10526	GSE24290	35
GDS4002	High, Normal, and Low Anxiety: multiple brain regions	Mus musculus	GPL6481	GSE29015	75
GDS4001	High and Low Anxiety: multiple brain regions	Mus musculus	GPL7467	GSE29015	70
GDS3993	BKS db/db model of type 2 diabetes: sciatic nerve	Mus musculus	GPL9746	GSE27382	13
GDS3977	Adenoid cystic carcinoma xenograft models	Homo sapiens	GPL570	GSE28996	22
GDS3917	Various central nervous system regions of adult inbred strain	Mus musculus	GPL1261	GSE16496	102

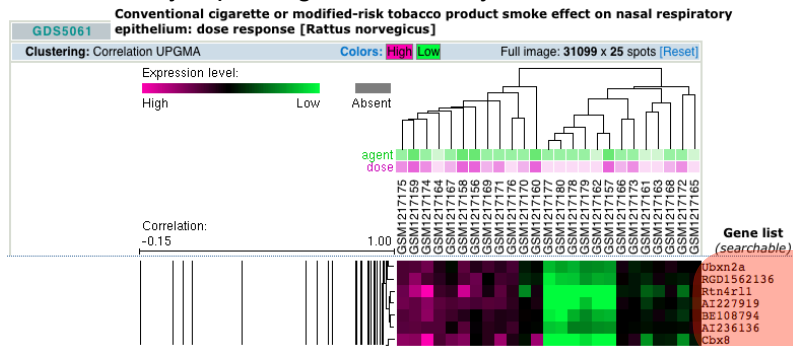DataSet Record GDS4001: (Expression Profiles) (Data Analysis Tools) (Sample Subsets)

Title:	High and Low Anxiety: multiple brain regions
Summary:	Analysis of seven brain regions closely connected to the emotion-regulating parts of the limbic system, isolated by micropunching from CD-1 males selectively bred for high (HAB) or low (LAB) anxiety-related behavior. Results provide insight into the molecular mechanisms underlying anxiety.
Organism:	Mus musculus
Platform:	GPL7467: MPIP24K
Citation:	Czibere L, Baur LA, Wittmann A, Gemmeke K et al. Profiling trait anxiety: transcriptome analysis reveals cathepsin B (Ctsb) as a novel candidate gene for emotionality in mice. PLoS One 2011;6(8):e23604. PMID: 21897848
Reference Series:	GSE29015
Value type:	log2 ratio

Sample count:	70
Series published:	2011/05/04

Cluster Analysis

Download
- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

GDS5079   Ubiquitin-specific protease 2 deficiency effect on the kidney [Mus musculus]
Clustering: Correlation UPGMA   Colors: High Low   Full image: 28270 x 4 spots

2. Find a group of highly correlated genes that is different between the experimental groups.

a. Zoom in on a cluster by dragging the dotted outline box to your cluster of interest, moving the upper and lower boundaries so that only your cluster is in the box, and then double click in the box.

b. Refine your cluster by repeating as necessary; this cluster will be used in part 4

Conventional cigarette or modified-risk tobacco product smoke effect on nasal respiratory
GDS5061   epithelium: dose response [Rattus norvegicus]
Clustering: Correlation UPGMA   Colors: High Low   Full image: 31099 x 25 spots [Reset]

Expression level:
High   Low   Absent

Correlation:
-0.15   1.00

Gene list
(searchable)

Ubxn2a
RGD1562136
Rtn4rl1
AI227919
BE108794
AI236136
Cbx8

## Part 3: Identify Gene Ontology functional groups

- Copy the list of genes near your gene on the heat map (start with about ~50 of them, the smaller the list the less clusters you will see)
- Paste them into the **Enter IDs box**: http://pantherdb.org
- **Select List type**: ID (most likely)
- **Select Organism:** check for this on the dataset you are using (Homo sapiens, Mus musculus, drosophila, etc)
- Select Analysis: **Functional classification viewed in a pie or bar chart**
- Click on **Biological Process or Pathway** and you can find the pathways these genes that cluster together belong to.
- You can export the pie chart or bar chart

PANTHER PIE CHART
Bar Chart   Export

Features:

- Mouse-over pie chart section to see category name and statistics
- Click on a pie chart section to drill down to child categories
- Click on chart legend link to retrieve gene list for each category
- Click on a color key in chart legend to choose your favorite color for the category NEW!
- Click on ⬚ in chart legend to highlight your selection in pie chart NEW!
  Click on ⬚ to reset

Select Ontology: [Pathway ▾]   View: [100% ▾]

PANTHER Pathway
Total # Genes: 238   Total # pathway hits: 70
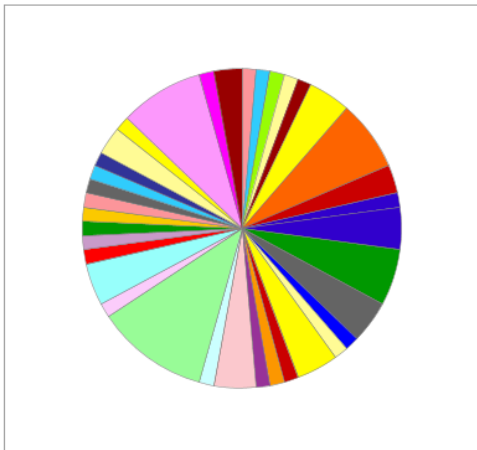
Click to get gene list for a category:
- 5HT1 type receptor mediated signaling pathway (P04373)
- 5HT2 type receptor mediated signaling pathway (P04374)
- 5HT3 type receptor mediated signaling pathway (P04375)
- 5HT4 type receptor mediated signaling pathway (P04376)
- Adenine and hypoxanthine salvage pathway (P02723)
- Alzheimer disease-presenilin pathway (P00004)
- Angiogenesis (P00005)
- B cell activation (P00010)
- Blood coagulation (P00011)
- CCKR signaling map (P06959)
- Cadherin signaling pathway (P00012)
- Cytoskeletal regulation by Rho GTPase (P00016)
- De novo purine biosynthesis (P02738)
- FAS signaling pathway (P00020)
- Gonadotropin-releasing hormone receptor pathway (P06664)
- Heme biosynthesis (P02746)
- Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026
- Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P0002
- Huntington disease (P00029)
- Hypoxia response via HIF activation (P00030)
- Inflammation mediated by chemokine and cytokine signaling pathway (P00031)
- Insulin/IGF pathway-protein kinase B signaling cascade (P00033)
- Integrin signalling pathway (P00034)
- Ionotropic glutamate receptor pathway (P00037)
- Nicotinic acetylcholine receptor signaling pathway (P00044)
- PI3 kinase pathway (P00048)
- Parkinson disease (P00049)
- T cell activation (P00053)
- TGF-beta signaling pathway (P00052)
- Toll receptor signaling pathway (P00054)
- Ubiquitin proteasome pathway (P00060)
- VEGF signaling pathway (P00056)
- Vitamin D metabolism and pathway (P04396)
- Wnt signaling pathway (P00057)
- p53 pathway feedback loops 2 (P04398)
- p53 pathway (P00059)

Web

**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Pathway hits