Pathway Analysis

November 02, 2023

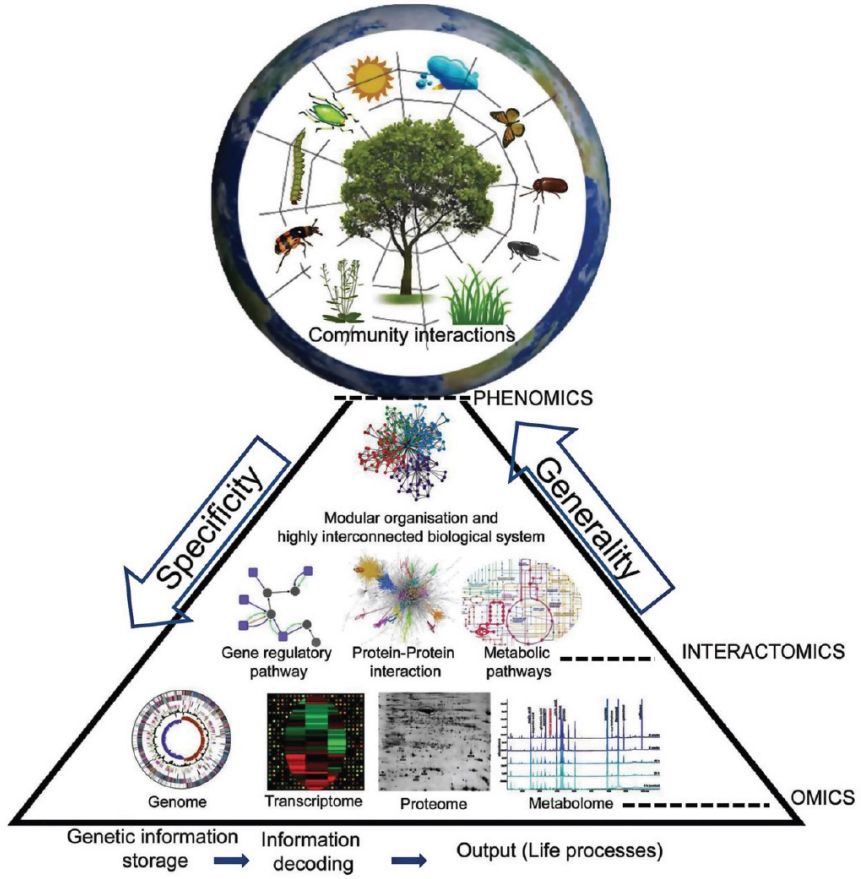
Why pathway analysis – Goals for today

- Purpose of systems approach
- Biological notion of pathway
- Informatics notion of pathway – what is testable
- Outcome of pathway analysis

cribioinfo@uic.edu

Systems biology

- Broadly: modeling complex biological systems holistically, rather than focusing on specific individual molecules
  - Functioning of biological pathways
  - Multi-omics: Interplay between genes, proteins, lipids, metabolites, and other small molecules
  - Hierarchy of systems
- Practically: comparing experimental gene lists to annotated lists from pathways or other sources



https://www.ntnu.edu/biology/research/molecular-systems-biology

cribioinfo@uic.edu

Benefits of a systems biology approach

- Data reduction technique (100s-1000s of genes to 10s-100s of pathways)
- May be more accurate in modeling disease than using molecular biomarkers
- Offers a framework for integrating multiple -omics data sets
- Helps with molecular data interpretability, spanning from molecular to biological processes
- Useful tool for hypothesis generation

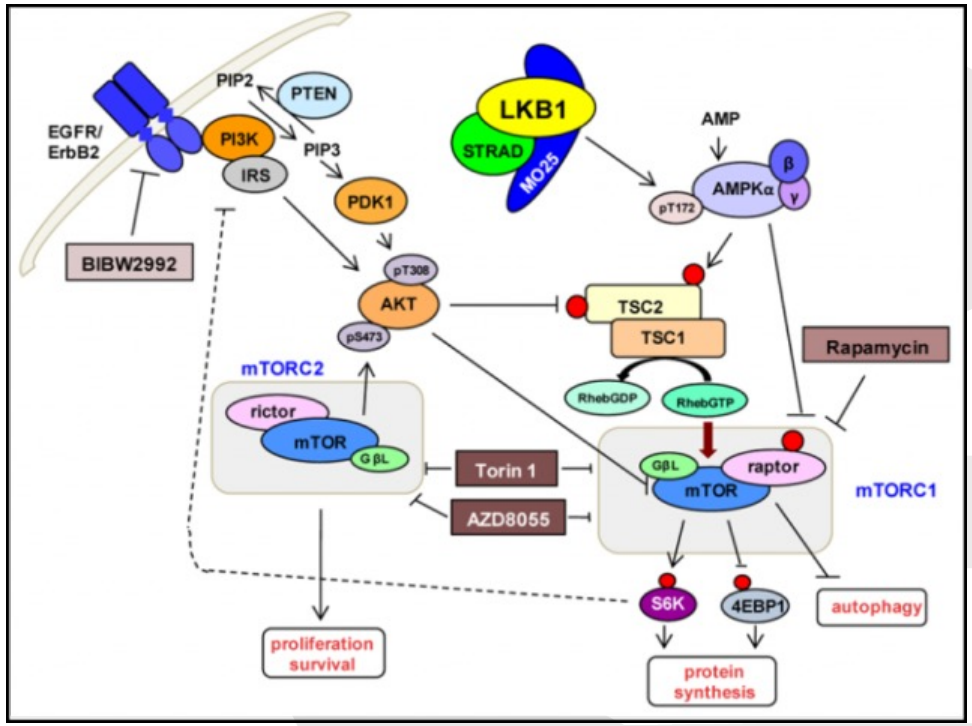
Limitations of systems biology approach

- "Pathways" may have fluid, subjective definitions
  - Different databases will have different pathway names, and different molecules in those pathways
  - Subject to expert curation (there are attempts at curation by machine learning – "natural language processing")
- Results are descriptive/qualitative: difficult to model in a quantitative way
  - How to predict the effect of differential expression on a pathway's function?
- Pathway information is surely incomplete; some molecules are better annotated than others, which can bias hypothesis generation
  - Isoforms generally not differentiated

What is a “Pathway”?

- “Pathways” are a set of molecules (genes, proteins, metabolites) that work together (interact) to perform a biological function.
- Pathway structures vary a lot by context
  - Signaling
  - Metabolism
  - Gene regulation
  - Many different types of interactions
- In context of pathway **analysis**, it is simpler to treat each pathway as a list of molecules

Pathway (biology)



Pathway (bioinformatics) (23 molecules)
EGFR, ErbB2, PI3K, PTEN, IRS, PIP2, PIP3, PDK1, AKT, rictor, raptor, mTOR, STRAD, LKB1, MO25, AMPKa, TSC2, TSC1, Rheb, Torin1, AZD8055, S6K, 4EBP1

https://www.ebi.ac.uk/training/online/course/reactome-exploring-biological-pathways/what-reactome/what-reactome-pathway

Molecule Lists

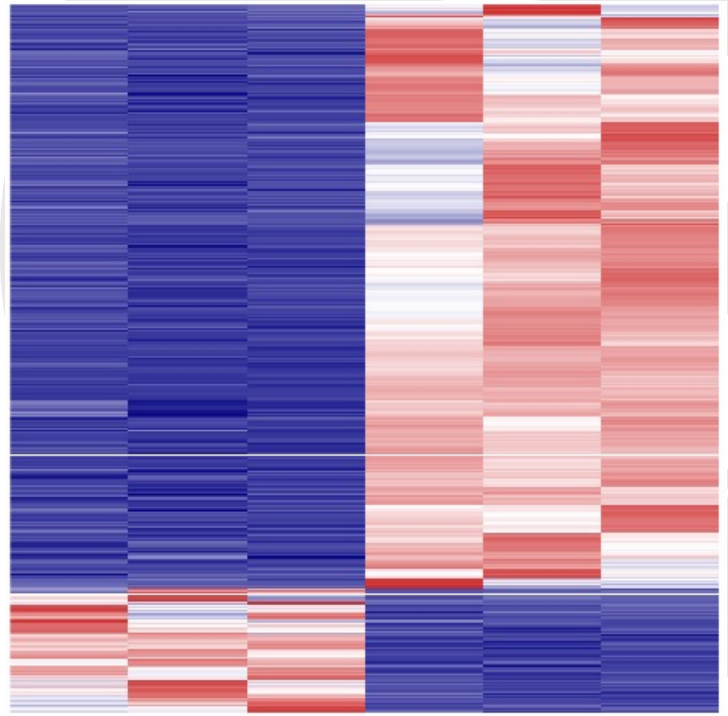
Where does our “data” come from in pathway analysis?

Getting a molecule list

- Determine interesting molecules from experimental data
  - Should be an untargeted (unbiased) experiment
- Examples:
  - Differentially expressed genes (RNA-seq), proteins (LCMS TMT), metabolites (untargeted LCMS)
  - Genes with damaging variants (DNA-seq)
  - Genes with higher levels of methylation, or bound by a TF (ChIP-seq)
  - Sensitivity in genome-wide CRISPR screen
  - Protein list from non-quantitative proteomics

Gene Lists from expression data

- Differentially expressed genes (based on some cutoff)
  - Up regulated genes
  - Down regulated genes
  - FDR corrected p-value (q-value) ≤ 0.05
- Gene clusters
  - Cluster genes based on expression patterns
  - Select the cluster(s) of interest for generating gene list(s)
- Marker genes from single-cell clustering analysis
- Similar approaches could apply for proteomics, metabolomics, lipidomics



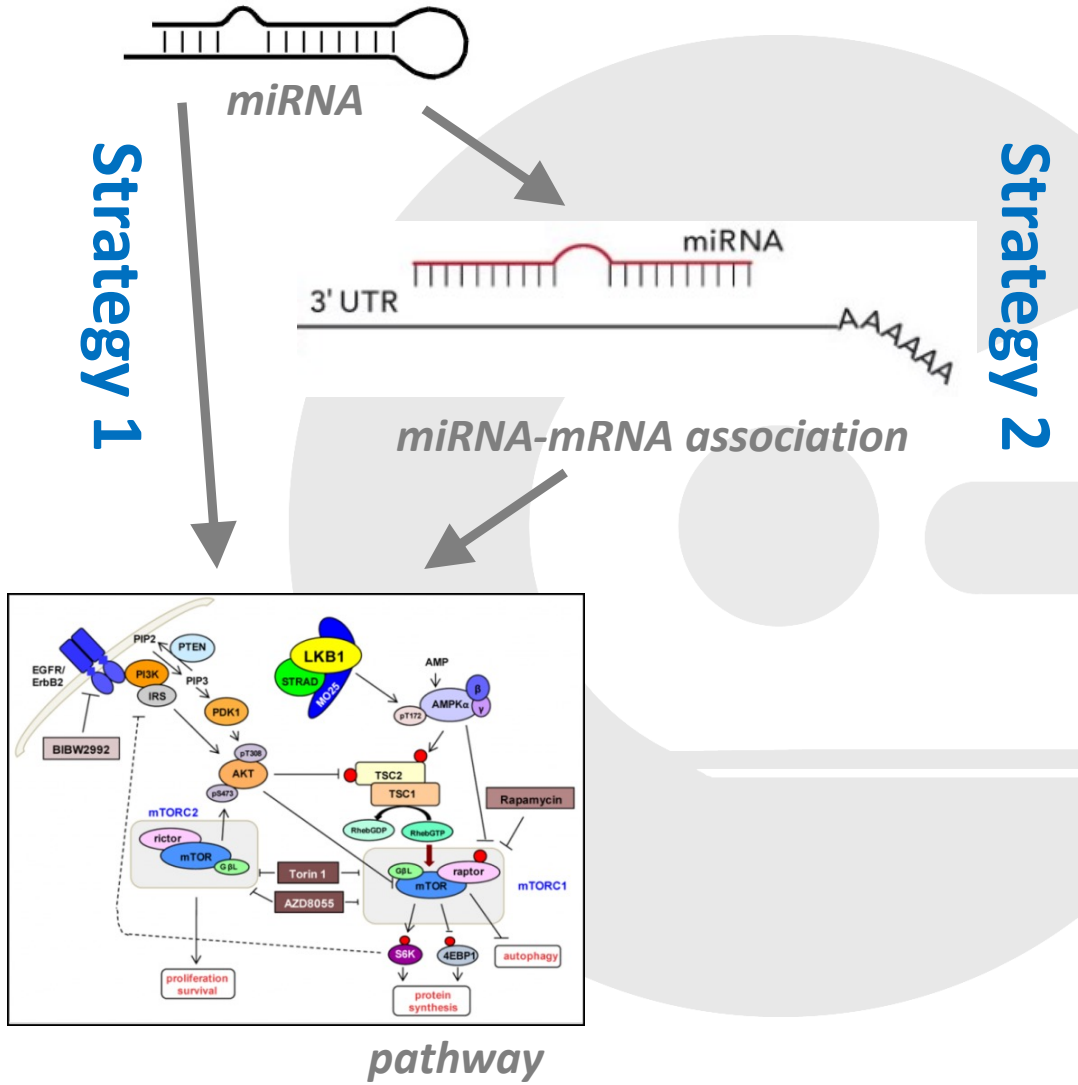
cribioinfo@uic.edu

Other options for gene lists – EXAMPLES

- Chromosomal location in relation to gene expression
- Presence of promotor sequence/motif
- Rate of mRNA gradation vs transcript length, GC content, codon usage
- Relationship between between evolutionary gene age in response to stress adaptation
- *In general, could be any list you want*

miRNAs

- Strategy 1: use pathway database with miRNAs included
  - More direct approach
  - Easier to integrate with other –omics data sets
  - Often only a few miRNAs annotated to a given pathway
  - Not all database include miRNAs
- Strategy 2: map miRNAs to target genes first
  - Requires intermediate step for miRNA-mRNA annotation
  - Need to balance predicted targets vs different levels of evidence
  - Allows for broader application to different pathway databases



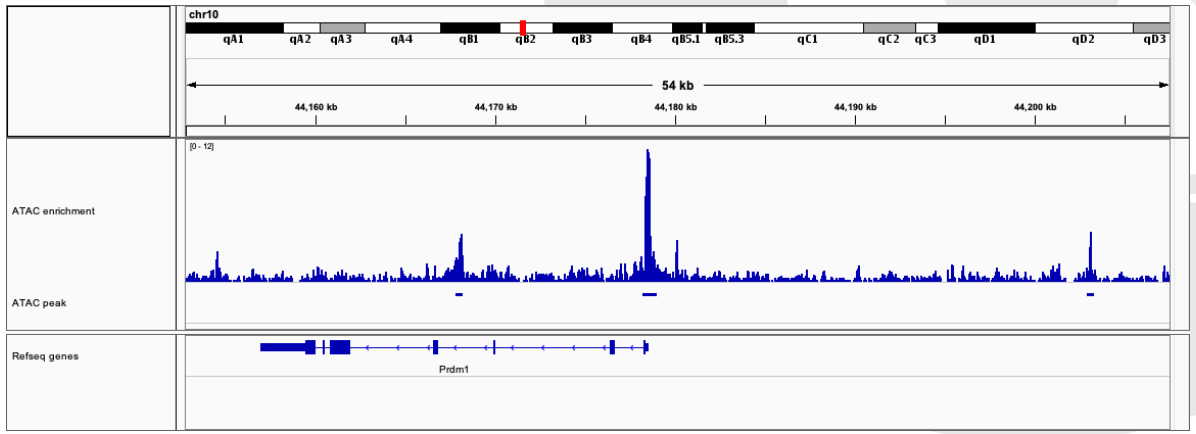
cribioinfo@uic.edu

Proteomics

- Typically mass spectrometry (LC-MS-MS) based experiments
- Untargeted vs. Targeted Proteomics
  - What is the bias of your experiment?
- Quantitative approaches (TMT, iTRAQ, SILAC, etc.)
  - Filter differentially expressed proteins (up vs. down, $q \leq 0.05$)
  - Cluster data and use cluster(s) of interest
- Qualitative approaches (protein ID)
  - Set threshold on number of detected peptides/spectra
  - Compare intersection or difference of conditions.
- Post translational modifications (PTMs)
  - Some pathway analysis tools may have PTM annotations/databases.

Epigenomics: ATAC/ChIP-seq (peaks), methylation (CpGs)

- Gene list based on genic location of epigenomic features
- Features:
  - ChIP-seq or ATAC-seq peaks
  - CpG sites or groups of CpGs (e.g., islands)
  - Combination of data sets
- Genic location:
  - Promoter
  - Gene body
  - Distal, known enhancers, TADs



cribioinfo@uic.edu

Variant calling: SNPs/indels

- Location in gene (exon/intron)
- Effect on gene (synonymous, nonsynonymous, damaging prediction)
- Tumor mutational burden
- Copy number variation burden
- Rare variants
- eQTLs

Metabolites/lipids/other small molecules

- Typically mass spectrometry (LC-MS-MS) based experiments
- Untargeted vs. Targeted experiments
  - What is bias of your experiment?
- Most experiments are quantitative
  - Filter on differential analysis results
  - Cluster data and use cluster(s) of interest
- Depending on the experiment/assay, compound identifications can be ambiguous

Molecular Identifiers

How do we “name” the molecules in our list?

Molecular identifiers

- How do you identify your molecules?
  - Information about genes/proteins annotations are typically limited to either humans, mouse, rat and model organisms
  - Functional annotations are usually based on these representative model organisms
- Existing databases are mostly based on orthologous gene/proteins.
- Each database has its own unique identifier and there can be substantial loss of information during ID conversion

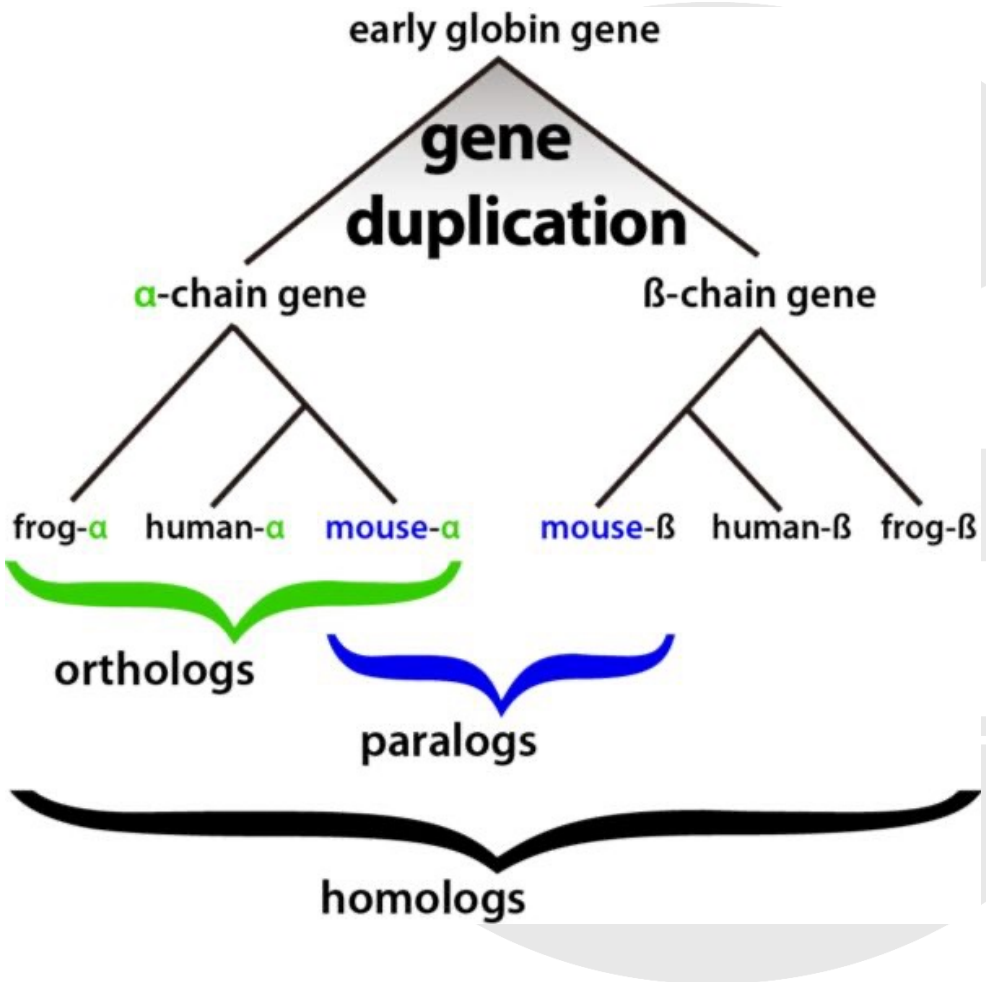
Sources of molecular identifiers: genes/proteins

- Genes:
  - Official gene symbol
  - Ensembl
  - Refseq
  - Entrez
- Proteins:
  - Gene symbol
  - Refseq
  - Uniprot

- Conversions may not be not one-to-one!
  - Gene symbol (BTBD8) (may match many species)
  - Ensembl (ENSG00000284413 and ENSG00000189195) (human version)
- Gene symbols can be problematic due to synonyms, capitalization
  - IL7R = Il7r, IL-7R, IL-7RA, IL7RA, IL7alpha, CDW127, CDw127

Orthologous Genes

- Ortholog – Same gene/function in different organism
- Orthologous IDs
  - Identify shared genes/functions across disparate organisms
  - Examples: KEGG (K00960), EC number (2.7.7.6)
- Caveat: When looking at a pathways in a database – are they organism-specific or generic?



https://sites.google.com/site/jkim339n/part2a

Small molecule Identifiers

- Small molecules are molecules that are not genes/proteins, i.e. metabolites, drugs, lipids.
- Molecular formula are often ambiguous (isomers) and names are not always standardized.
- Database/Accession IDs are best
  - CAS Registry Number, e.g. 50-28-2
  - PubChem CID, e.g. 5757
  - KEGG Compound ID, e.g. C00951
  - Human metabolite database (HMDB), e.g. HMDB0000151
- Molecular structure based identifiers can be used
  - InChI string/key are best. Methods exists to generate standard InChI strings.
  - Be careful with SMILES strings, they are not guaranteed to be consistent from same molecular structure.

Molecular identifiers in pathway databases

- Each database has its own internal identifier
  - May be the same as another database
  - May be unique to that database
- Not all molecules are annotated in pathways
  - These should be ignored in the enrichment statistics
- Total number of molecules may decrease (up to 10-20%), or in some instances increase, after conversion between ID/name types

Exercises 1.1-1.3: Converting IDs

- Converting ID types using
  - 1. DAVID
  - 2. UniProt
  - 3. BioMart

Pathway Databases

What am I comparing my data to?

Sources/databases

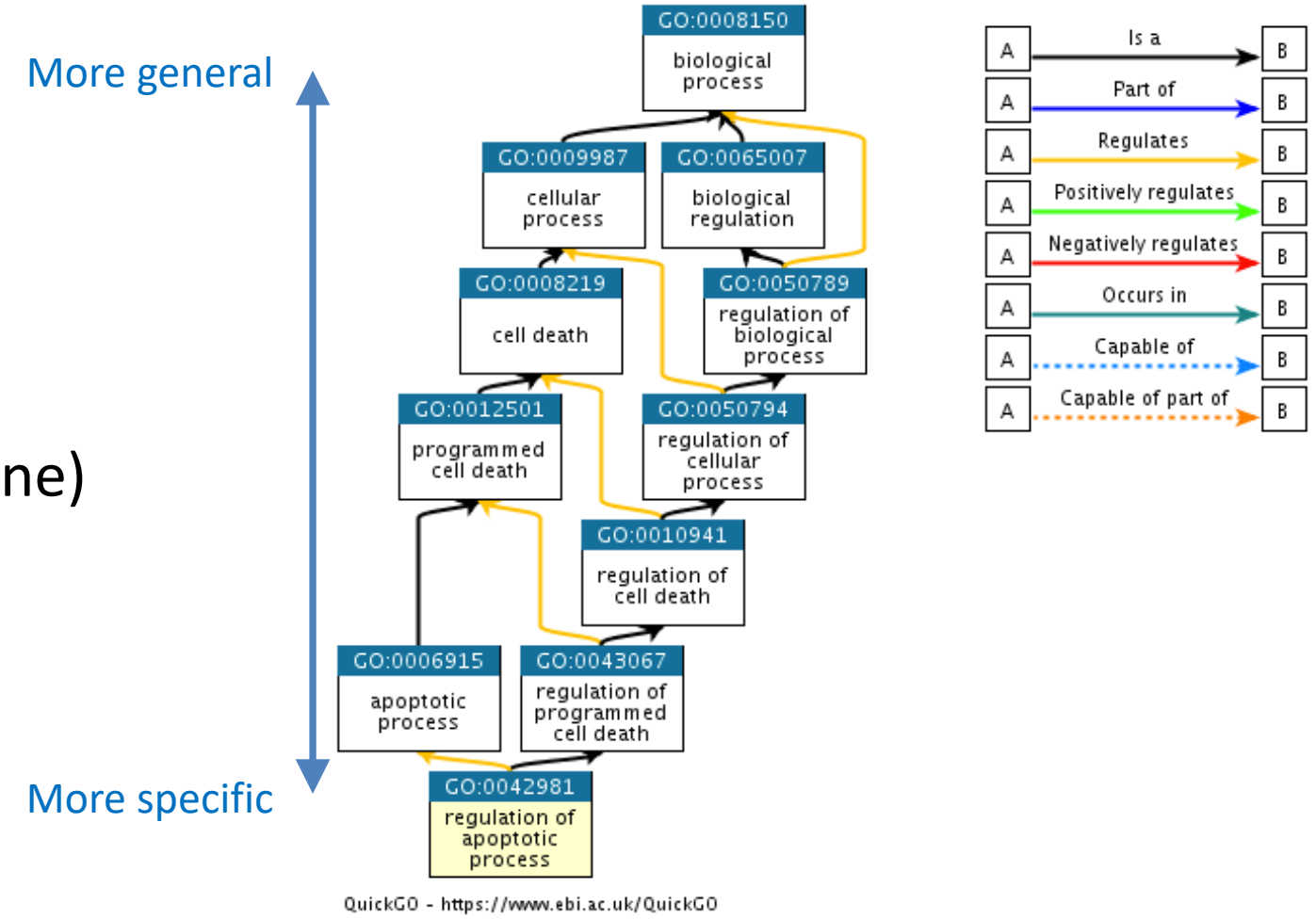
- Gene Sets Databases
  - Gene Ontology, The Molecular Signatures Database (MSigDB)
  - No molecular interactions (pathway = gene list)
- Biochemical Pathway Databases – with reactions/structure
  - KEGG (Kyoto Encyclopedia of Genes & Genomes), Reactome
  - IPA, MetaCore
- Pathway meta-databases (collect detailed pathway descriptions from multiple databases)
  - Pathway commons, WikiPathways

Gene Ontology (GO)

- Provides ‘terms’ / ‘classes’ to describe functions of gene products and their associations. 3 main databases:
- BP: Biological Process
  - ‘Biological programs’ made up of the activities of multiple gene products
  - Typical choice for “pathways”
  - *Example: positive regulation of interleukin-10 biosynthetic process*
- MF: Molecular Function
  - Molecular activities of gene products
  - Corresponds to single gene product
  - *Example: DNA-binding transcription factor activity*
- CC: Cellular Component
  - Cellular structure where gene product performs function
  - Classes/terminology refers to cellular anatomy
  - *Example: nuclear nucleosome*

Gene Ontology (GO)

- Hierarchy of terms
- The GO vocabulary is species agnostic
- Freely available
- Can access from many online sources
  - QuickGO (browse hierarchy online)
  - Amigo
  - DAVID
  - Panther
  - MSigDB
  - On some commercial systems (MetaCore)



cribioinfo@uic.edu

Web Interfaces to Access the GO

- Web-based tools for searching and browsing the Gene Ontology database
  - AmiGO – (http://amigo.geneontology.org)
  - QuickGo (http://www.ebi.ac.uk/QuickGO)
- Accessing GO using API
  - http://geneontology.org/docs/tools-guide/

Exercise 1.4: QuickGo

- Browse GO database on QuickGo
  - Search by term or gene
  - View hierarchy
  - Download gene list

Molecular Signatures Database (MSigDB)

- Collection of annotated gene sets
- 22596 gene sets
- 8 major collections
  - **H: hallmark gene sets**
    - relatively non-redundant collection
    - Used by many pathway enrichment methods
  - C1: positional gene sets
  - **C2: curated gene sets**
  - C3: regulatory target gene sets
  - C4: computational gene sets
  - **C5:** ontology gene sets, GO: Gene Ontology gene sets
  - C6: oncogenic signatures gene sets
  - C7: immunologic signatures genes
  - C8: cell type signature gene sets
- **Gene Set Enrichment Analysis** (GSEA) – uses this database by default the for the enrichment analysis

Exercise 1.5: MSigDB

- Browse MSigDB database
  - Browse by different collections (Hallmark, KEGG, etc.)
  - Search by term or gene
  - Download gene list

KEGG: Kyoto Encyclopedia of Genes and Genomes

- Categorizes genes/orthologs, proteins and compounds in different groups
  - Pathways – “Canonical” pathways
  - Modules – Functional units of pathways, reactions, or phenotypic features
  - BRITE categories – Hierarchical organization of pathways
- Contains molecular level information, disease associated gene sets for e.g. ‘pathways in cancer’, information for drugs and chemical substances.
- It has data for different species across different domains of life
  - Eukaryotes: 536, Bacteria:5612, Archaea: 317
- Website (https://www.kegg.jp/) free for academic use.

KEGG Database at Glance

1. **Systems information**
   - KEGG PATHWAY KEGG pathway maps
   - KEGG BRITE BRITE hierarchies and tables
   - KEGG MODULE KEGG modules
2. **Genomic information**
   - KO (KEGG Orthology) Functional orthologs
   - KEGG GENOME KEGG organisms (complete genomes)
   - KEGG GENES Genes and proteins
3. **Chemical information (KEGG LIGAND)**
   - KEGG COMPOUND Small molecules
   - KEGG GLYCAN Glycans
   - KEGG REACTION Reactions and reaction classes
   - KEGG ENZYME Enzyme nomenclature
4. **Health information**
   - KEGG NETWORK Disease-related network elements
   - KEGG DISEASE Human diseases
   - KEGG DRUG Drugs and drug groups
   - KEGG MEDICUS Health information resource [Drug labels search]

Exercise 1.6: Browse KEGG

- Browse KEGG database
  - Find a pathway in KEGG
  - Look at details for a gene
  - View organism specific version of the pathway

Broader types of “pathways”

- Regulator targets
- Drug targets
- Biomarkers
- Surface markers, molecule types (GO MF and GO CC)
- Protein domains (Interpro)
- **Custom – make your own gene list**

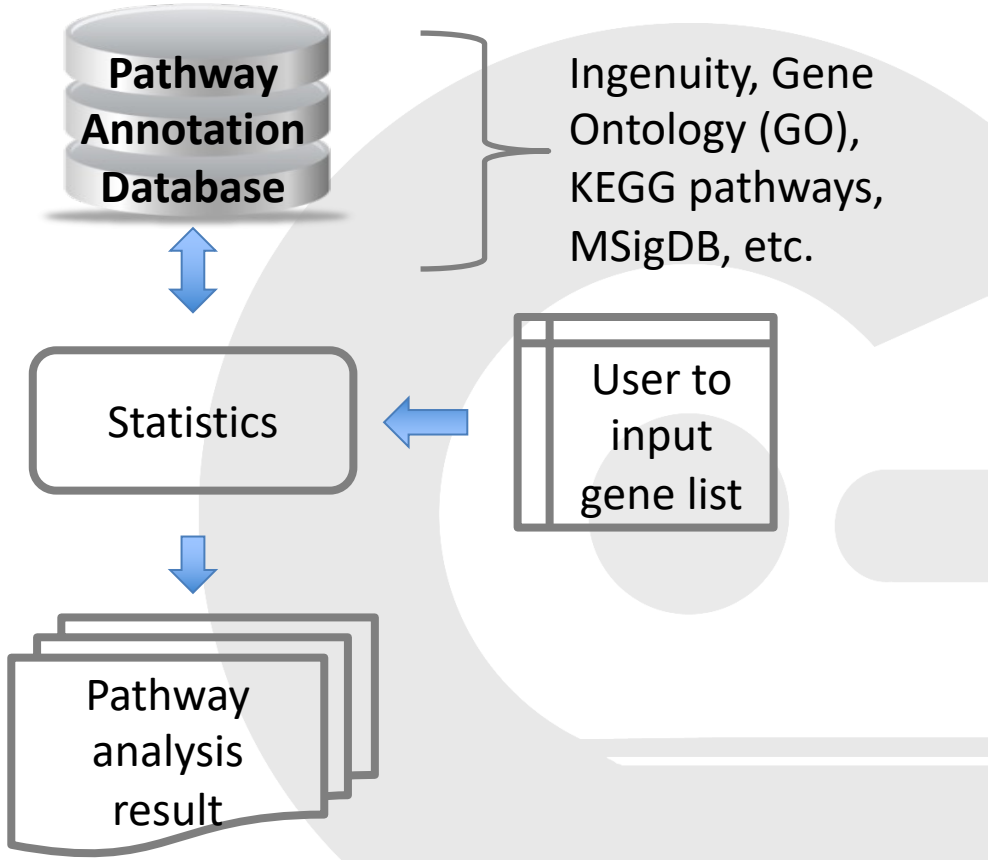
BREAK

Statistical analysis of pathways

Making comparisons between our data and pathways

Pathway enrichment

- Given: a **data set** of interest (e.g., molecule/gene list), and a **pathway database**
- Goal: determine which pathways in the database are enriched in the molecule list
- **In principle, the choice of enrichment statistic is independent of the pathway database**
  - Limitations are due to implementation and convenience
  - Commercial systems like IPA don't let you download the database, can only use built-in statistics



cribioinfo@uic.edu

Input molecule lists

- Experimentally determined
- Differential expression
  - Statistical significance (FDR)
  - Fold-change
  - Clustering analysis
- Examples:
  - Genes with FDR < 0.05
  - Genes with FDR < 0.05 and fold-change > 2
- Non-quantitative:
  - Non-quantitative proteomics
  - Genes with damaging SNPs
  - Genes with peak in promoter

Enrichment statistics

- Most common statistics:
  - **Fisher's Exact test (FET)**: Compare intersection between subset molecule list and pathways with
    - AKA: hypergeometric test, Over-Representation Analysis (ORA)
    - Requires us to pick a subset of molecules of interest
  - **Gene Set Enrichment Analysis (GSEA):** Look at ranking of genes in pathway in complete transcriptome data set
    - Requires a ranking statistic
- There are some others that we will review also
- **Always need to run FDR correction**

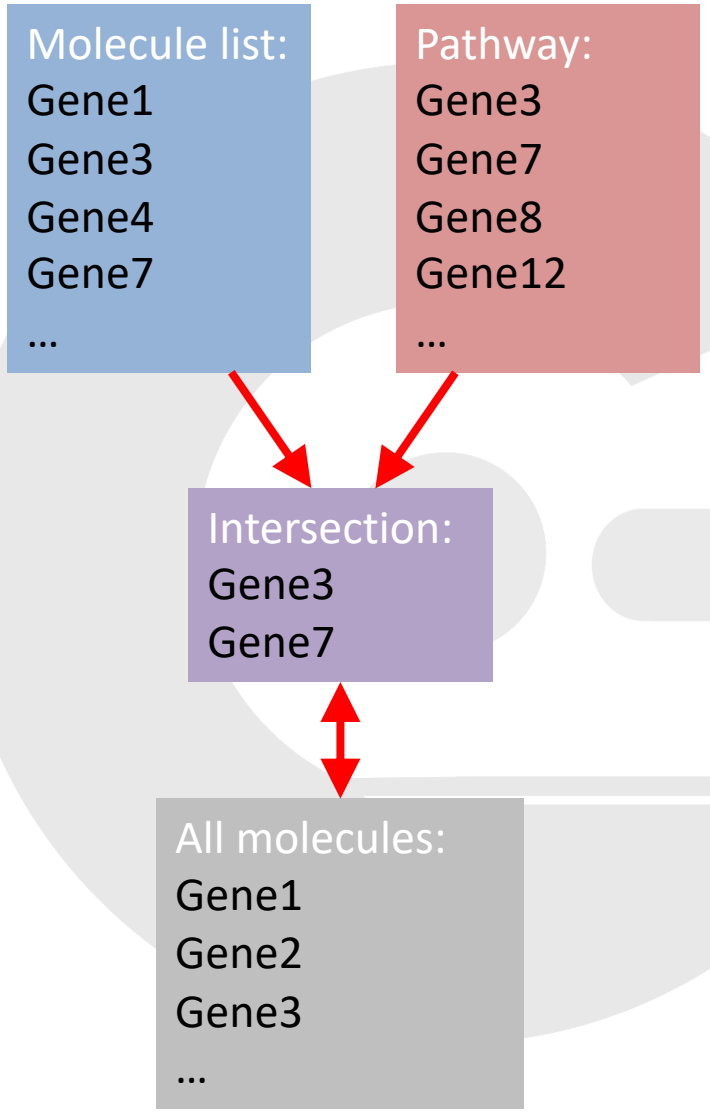
Typical results

- List of terms and p-values/fold enrichment, plus FDR correction for multiple testing (table/Excel file)

Category	Term	Fold Enrichment	PValue	FDR
GOTERM_MF_FAT	GO:0030695~GTPase regulator activity	1.866364949	2.56E-13	4.24E-10
GOTERM_MF_FAT	GO:0030554~adenyl nucleotide binding	1.376011116	4.01E-13	6.64E-10
GOTERM_MF_FAT	GO:0001882~nucleoside binding	1.37285846	4.02E-13	6.64E-10
GOTERM_MF_FAT	GO:0001883~purine nucleoside binding	1.371364102	5.79E-13	9.58E-10
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	1.384433726	5.81E-13	9.61E-10
GOTERM_MF_FAT	GO:0005524~ATP binding	1.382215399	1.09E-12	1.80E-09
GOTERM_MF_FAT	GO:0017076~purine nucleotide binding	1.317529549	4.64E-12	7.67E-09
GOTERM_MF_FAT	GO:0032553~ribonucleotide binding	1.321934278	7.57E-12	1.25E-08
GOTERM_MF_FAT	GO:0032555~purine ribonucleotide binding	1.321934278	7.57E-12	1.25E-08
GOTERM_CC_FAT	GO:0005856~cytoskeleton	1.424110951	8.05E-12	1.19E-08
GOTERM_MF_FAT	GO:0000166~nucleotide binding	1.26884698	1.04E-10	1.72E-07
GOTERM_MF_FAT	GO:0003779~actin binding	1.838128773	3.02E-10	5.00E-07
GOTERM_MF_FAT	GO:0005083~small GTPase regulator activity	1.966265756	3.14E-10	5.19E-07
INTERPRO	IPR001849:Pleckstrin homology	1.865430688	8.13E-10	1.47E-06
INTERPRO	IPR011993:Pleckstrin homology-type	1.81428899	9.47E-10	1.71E-06
GOTERM_CC_FAT	GO:0042995~cell projection	1.547963308	1.24E-09	1.84E-06
GOTERM_MF_FAT	GO:0008092~cytoskeletal protein binding	1.62743417	6.44E-09	1.06E-05

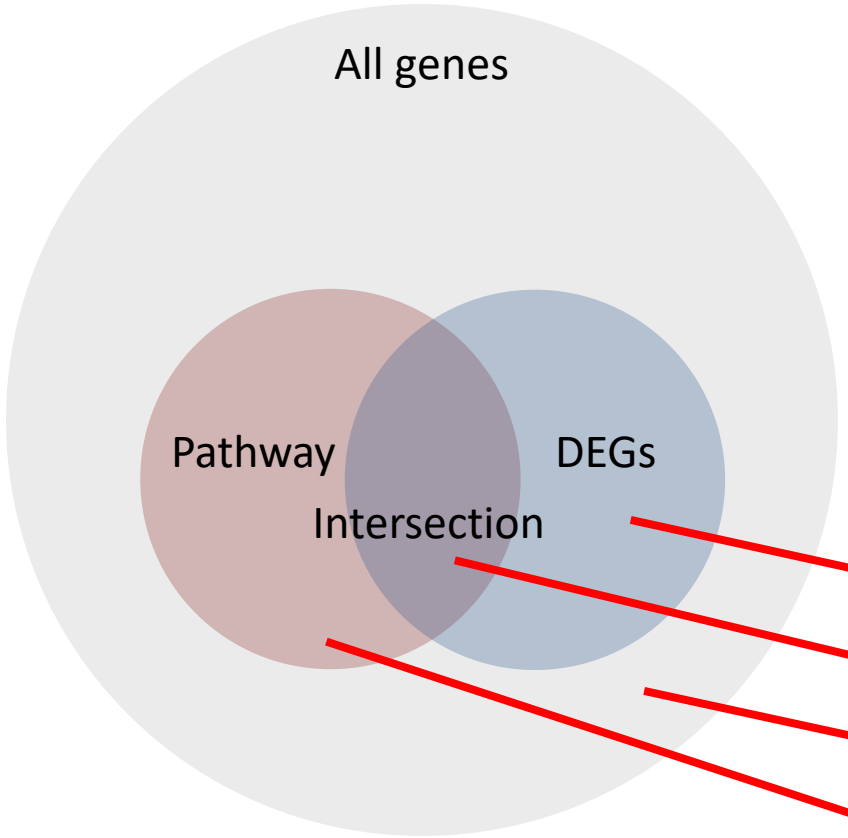
Fisher's Exact test (FET)

1. Get **molecule list** of interest (**Data**)
   - DEGs, ChIP-seq target genes, mutated genes, etc.
2. Get **pathways** (molecule sets) from database
   - GO, KEGG, MSigDB, IPA, etc.
3. Check **intersection** between molecule list and pathways
4. Compare to expected intersection given overall size of transcriptome/proteome/metabolome (**all molecules**)



cribioinfo@uic.edu

Contingency table



Fisher's Exact test:

Odds ratio: $\frac{35/50}{540/16{,}540} = \frac{0.7}{0.03} = 20.8$

(20.8 fold-enrichment)

Log2 Odds ratio: log2(20.8) = 4.4

P-value: 4.3e-29

(very significant)

	In pathway	Not in pathway
DEG	35	540
Not DEG	50	16,040

cribioinfo@uic.edu

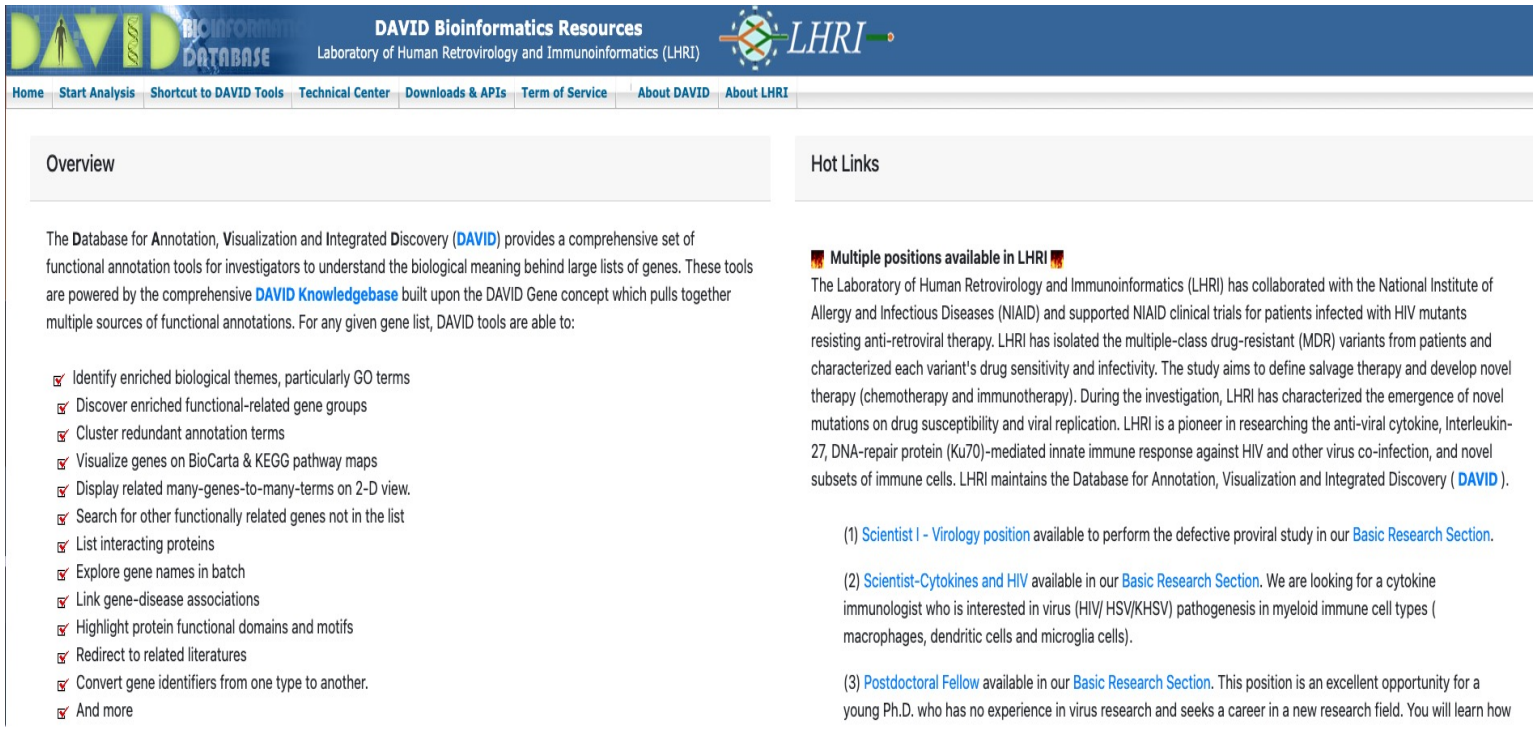
Considerations for FET

- Input is a molecule/gene list
  - No consideration of fold-change, significance, etc.
  - Stats are only based on what's in the list
- You must decide ahead of time what thresholds to set
  - FDR and/or fold-change
  - Separate up and down-regulated?
- Gives you flexibility
  - Genes don't have a "signal" (e.g., all genes with high mutation burden; ChIP-seq targets)
  - Gene clusters – many different combinations of signals
- But, loss of information
  - Gene with fold-change of 2 and q-value of 0.05 treated the same as a gene with fold-change of 10 and FDR of 0.000001

Implementations

- Most common statistical test in many free and commercial packages
  - DAVID (http://david.abcc.ncifcrf.gov) (free)
  - MetaScape (https://metascape.org/gp/index.html) (free)
  - PANTHER (http://pantherdb.org) (uses as similar binomial test) (free)
  - Reactome (free)
  - Ingenuity Pathway Analysis (commercial)
  - Metacore (commercial)
- Easy to use in R also
  - Would need to supply your own pathway database
  - Most useful for checking custom pathways/molecule lists

DAVID



- Fisher's Exact test enrichments
- GO, KEGG, and a number of other pathway/gene signature databases
- Also includes a gene ID conversion

cribioinfo@uic.edu

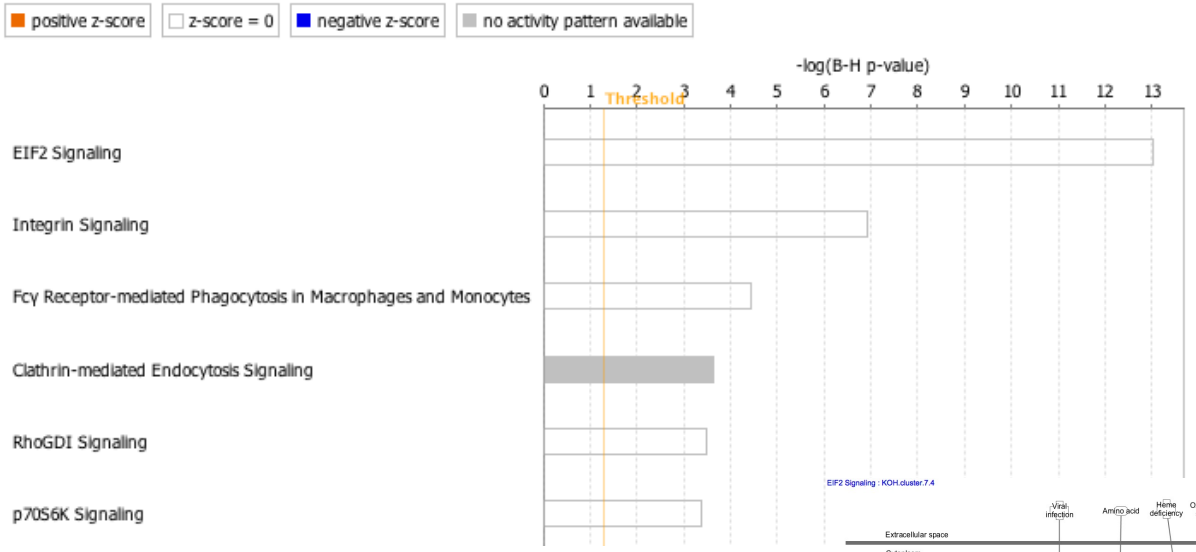
Exercise 1.7: Run analysis in DAVID

- Run pathway enrichment in DAVID
  - We'll use a practice data set comparing normal brain tissue to that with a viral infection
- Export results

Ingenuity Pathway Analysis

- Built on curated molecular interactions from the literature
- 2 primary functions:
  - "Pathway" databases: Enrichment of experimental molecule set of interest against a curated molecule lists
    - Biological pathways, upstream regulators, disease biomarkers, drug targets, …
  - Network building: expand/interrogate functions relating to key molecules of interest
- Many other analyses also available, such as comparing your data set to other experiments

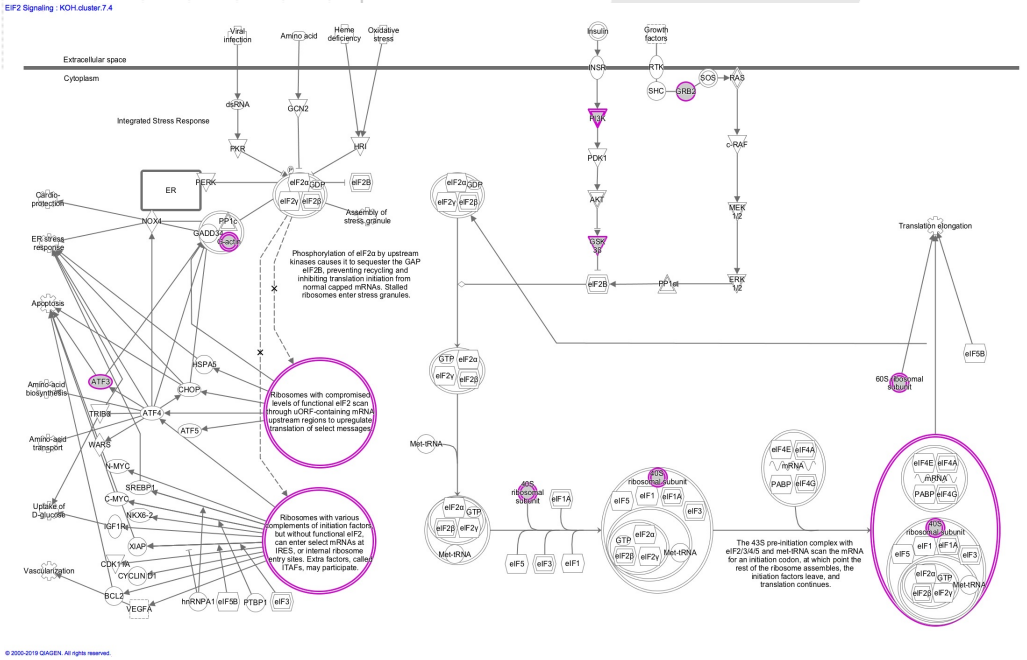
IPA Pathways and networks



FDR-corrected p-values from canonical pathways

Ability to look at structure of pathways (EIF2 Signaling)

- **Interactive**
- **Molecules from experimental gene list in purple**



cribioinfo@uic.edu

Exercise 1.8 (instructor-only): IPA

- Import data set into IPA
- Look at results from canonical analysis
- Export results

Introduction to Ingenuity Pathway Analysis (IPA) Workshop

When: Wednesday, Nov 8, 2023

Where: VIRTUAL and IN-PERSON
Virtual: Via Zoom
In-Person: Molecular Biology Research Building (MBRB), Rm 1017

Time: 9 am to 1 pm

Cost: FREE [**Registration Required**]

Laptop Requirement:
This will be a hands-on workshop! Please bring a laptop.

Questions? Email:
resinfocore@uic.edu

Ingenuity Pathway Analysis (IPA) is a commercial pathway and network analysis tool for systems biology analysis. Applications include:

- Pathway enrichment analysis
- Network building
- Biomarker discovery
- Elucidation of biological mechanisms
- Visualization of complex trends
- Hypothesis generation
- Literature review

Access to an institutional license for IPA is administered by the Research Informatics Core (RIC) and the Research Resources Center (RRC). A Qiagen IPA representative will be offering this FREE hands-on training workshop. Temporary access to IPA will be provided for free for the duration of the workshop.

Registration:
go.uic.edu/ipaworkshop

Please register by Nov 2nd to ensure sufficient time for your account for the workshop will be created.

Using FET in R

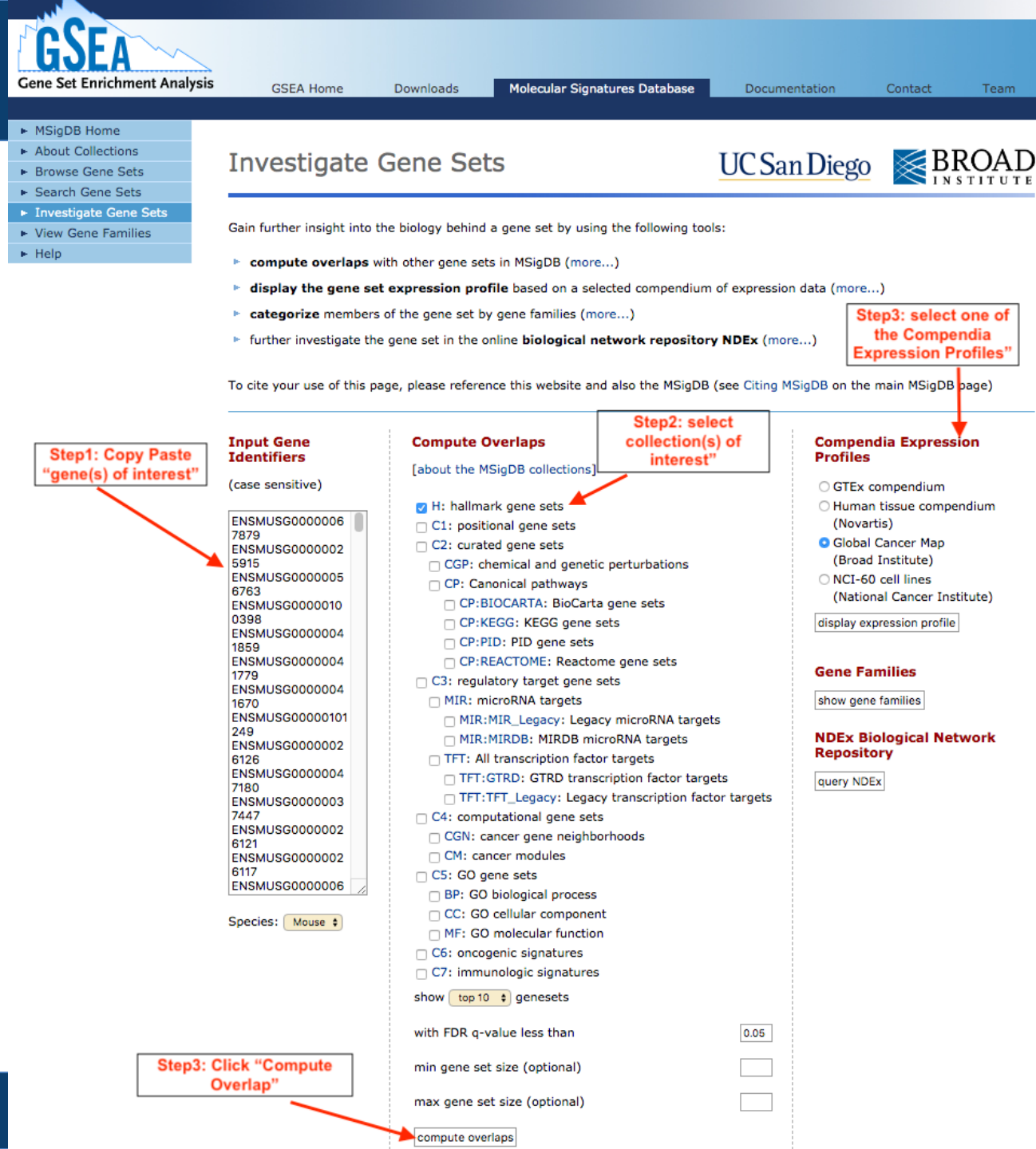
- We can easily run Fisher's Exact Test in R/R studio
  - Need to import pathway/molecule set to test against also
  - If testing many different pathways, will need to store all p-values in a data frame and run FDR correction
- There are R packages for running GO enrichments, but associated databases may not be updated
  - Check how the package is set up
- Most useful application: testing against a custom molecule list

Exercise 1.9: FET in R

- Pathway enrichment for custom gene set
  - Our practice data set is from a viral infection in brain tissue
  - We will compare to a list of anti-viral genes

Fisher's Exact Test in MSigDB

- MSigDB includes a simple interface for FET statistics
  - **Investigate Gene Sets** selection
- Choose combination of databases
- Limited features
  - Copy-paste gene list
  - Species: human/mouse/rat
  - Max 2000 input genes

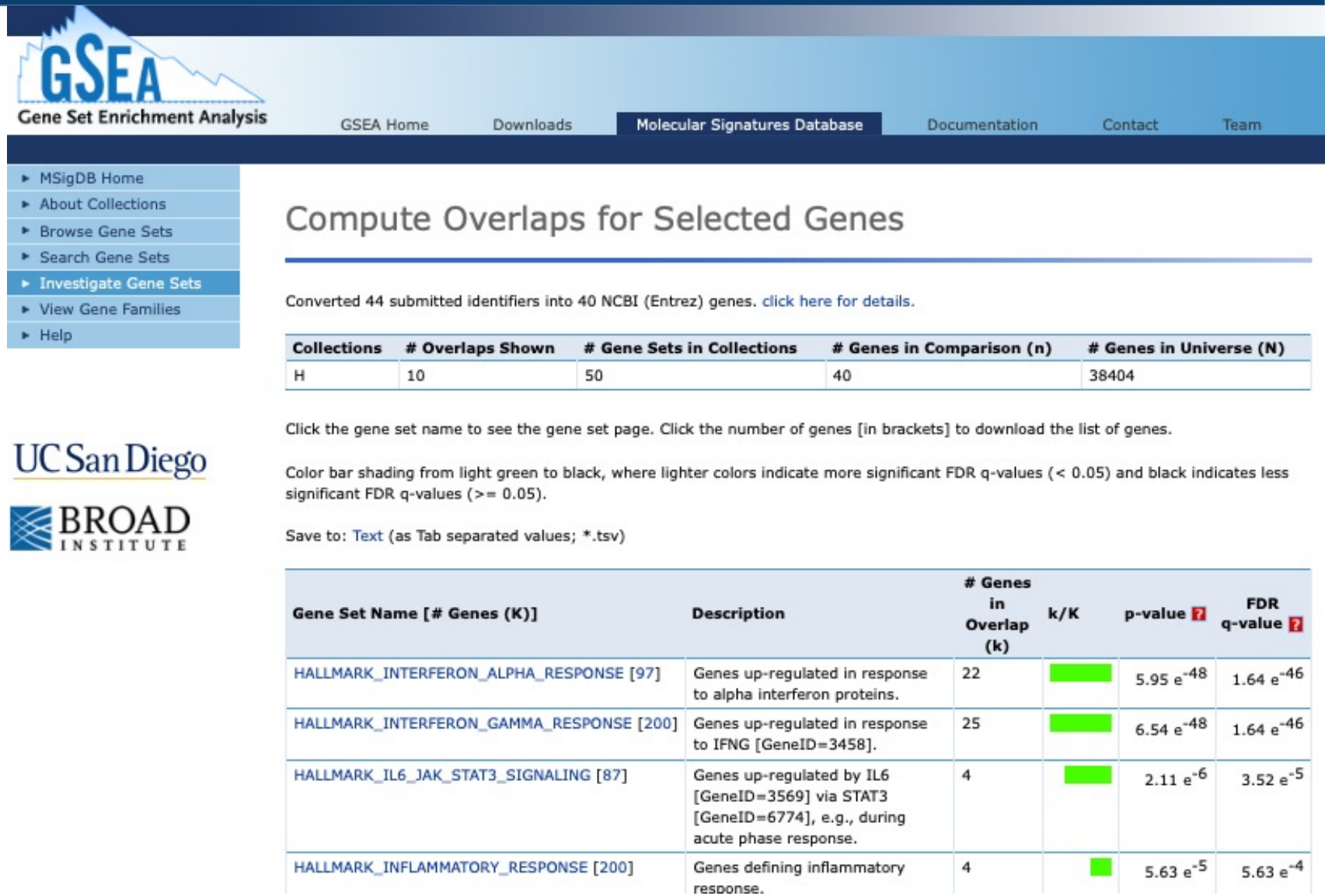


cribioinfo@uic.edu

Fisher's Exact Test in MSigDB

Results:

- P-value, FDR, and enrichment ratio (k/K)
- Can save as text file



GSEA
Gene Set Enrichment Analysis

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact | Team

- MSigDB Home
- About Collections
- Browse Gene Sets
- Search Gene Sets
- Investigate Gene Sets
- View Gene Families
- Help

UC San Diego
BROAD INSTITUTE

Compute Overlaps for Selected Genes

Converted 44 submitted identifiers into 40 NCBI (Entrez) genes. click here for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
H	10	50	40	38404

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

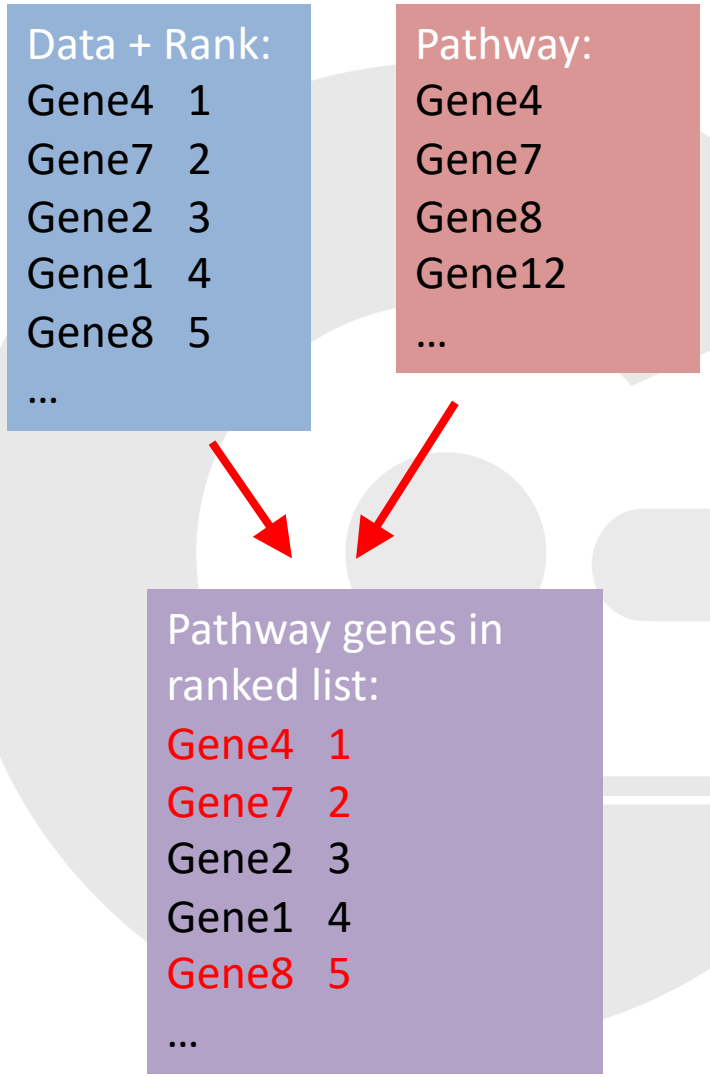
Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Text (as Tab separated values; *.tsv)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	22		$5.95\ e^{-48}$	$1.64\ e^{-46}$
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	25		$6.54\ e^{-48}$	$1.64\ e^{-46}$
HALLMARK_IL6_JAK_STAT3_SIGNALING [87]	Genes up-regulated by IL6 [GeneID=3569] via STAT3 [GeneID=6774], e.g., during acute phase response.	4		$2.11\ e^{-6}$	$3.52\ e^{-5}$
HALLMARK_INFLAMMATORY_RESPONSE [200]	Genes defining inflammatory response.	4		$5.63\ e^{-5}$	$5.63\ e^{-4}$

Gene Set Enrichment Analysis (GSEA)

1. Genome-wide data with signals (**Data**)
   - **Rank-order** genes based on expression
2. Get **pathways** (molecule sets) from database
   - GO, KEGG, **MSigDB**, ~~IPA~~, etc.
3. Check ranked values for genes in the pathway
   - More high ranks = up-regulated
   - More low ranks = down-regulated



cribioinfo@uic.edu

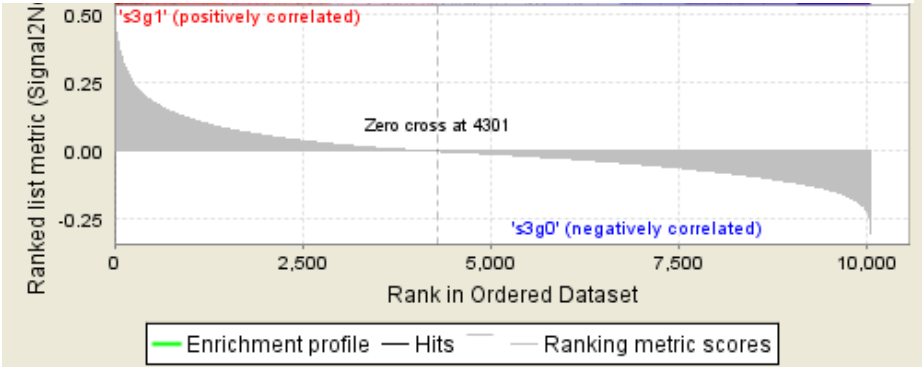
GSEA enrichment overview

- "Leading edge" statistic
  1. Rank order genes
  2. Look at ranks of genes from pathway
  3. Get cumulative rank "score" of genes in pathway: enrichment score (ES)
  4. Compute p-value for "leading edge" (max/min of ES) based on permutation test

4. Find leading edge, test height with permutation test

2. Rank of pathway genes (black lines)

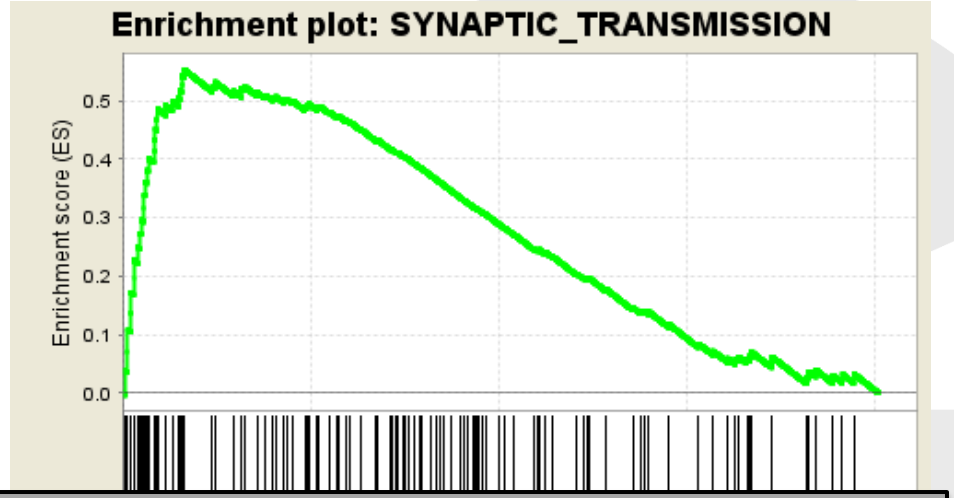
1. Rank of all genes



cribioinfo@uic.edu

GSEA enrichment overview

- "Leading edge" statistic
  1. Rank order genes
  2. Look at ranks of genes from pathway
  3. Get cumulative rank "score" of genes in pathway: enrichment score (ES)
  4. Compute p-value for "leading edge" (max/min of ES) based on permutation test



Result is p-value, enrichment score, PLUS indication of "up" or "down" regulated

- Based on pathway genes clustering towards left or right of rank list

Considerations for GSEA

- No need to set thresholds: input is whole transcriptome
  - You do need to decide how to rank*
- Limited to a pair-wise comparison
  - With multiple groups, or no signal, still need to rank somehow, or can't use GSEA
- Enrichment score, and up vs down for pathway, is a somewhat simplistic model
  - What if you have a mix of activators and repressors?
- Statistical test is permutation based*
  - Significance limited by number of permutations
  - Execution time is longer with more permutations
  - Can permute gene names or sample labels

**** More discussion on next slides***

Rank-ordering statistics

- Goal is to rank genes from most up-regulated (1st rank) to most down-regulated (last rank)
  - Non-differential genes in the middle
- Default: "signal to noise"
  - Similar to t-statistic

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

$\mu_{A/B}$ = mean expression of A or B

$\sigma_{A/B}$ = standard deviation of A or B

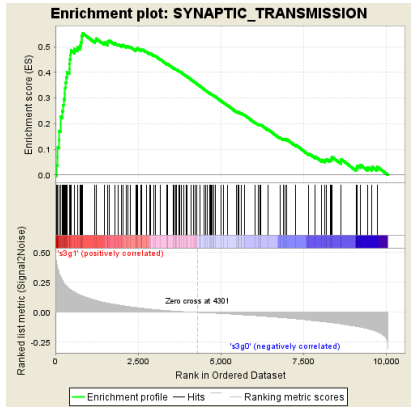
  - Better than fold-change as it includes variance also
  - Input is normalized expression levels, plus list of sample groups
- Several other options are available within GSEA as well

cribioinfo@uic.edu

DIY ranking

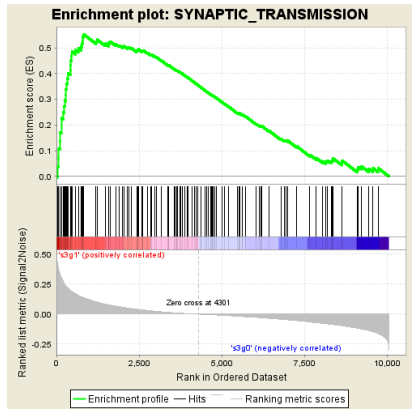
- You can also supply your own ranks to GSEA, instead of having it do the calculation
- Signal-to-noise developed for microarray, some modifications may be needed for RNA-seq
  - Eliminate very low-expressed genes first
  - Use log CPM table for calculation
- Alternative 1: compute (log FC) * (-log P-value)
  - Rank on this statistic
  - Accounts for counts-based nature of data
- Alternative 2: rank on p-value
  - This groups up- and down-regulated together, may be appropriate in certain circumstances

Permutation test



Compute ES for data set

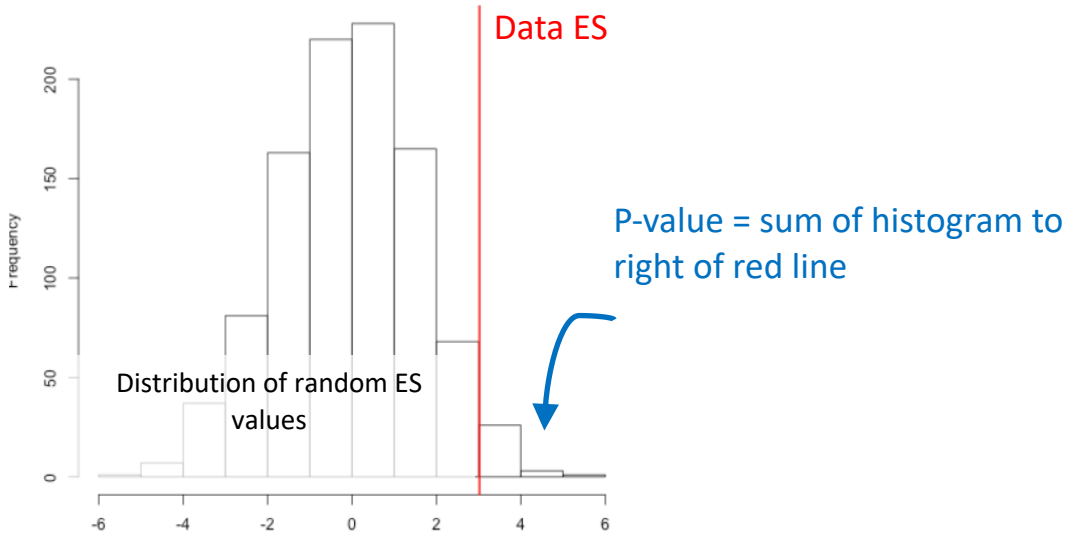
Randomly shuffle gene names OR sample phenotypes



Compute ES for shuffled data set

Many times (~10,000)

- Compare data ES to random ES list
  - P-value: % of random ES values > data ES



- NOTE: smallest p-value is 1/(# of permutations)
  - 100 permutations: smallest p-value is 0.01
  - 1000 permutations: smallest p-value is 0.001
    - Takes 10x longer to run than 100 permutations
- Compute normalized ES to account pathway size

cribioinfo@uic.edu

What to permute?

- By default, GSEA permutes the **phenotype labels**
  - E.g.: randomly shuffle case and control labels
- If you don't have very many samples, there aren't very many ways to do this
  - Not recommended with <7 samples *per* phenotype
- Instead, you can permute the **genes set**
  - Randomly change which genes are in the pathway
  - May be less conservative: does not account for gene correlations that are inherent in biological pathways
- If you supplied your own ranking, then you have to do gene set permutation

Implementations

- Web-based:
  - GenePattern server (https://www.genepattern.org)
  - Always linked to up-to-date pathway databases (MSigDB)
  - Can also upload your own gene lists
- Downloadable desktop application or command-line Java application
  - https://www.gsea-msigdb.org/gsea/downloads.jsp
  - You will also need to download the databases

GSEA file formats: .gct file

Always "#1.2"

Number of genes

Number of samples

```
#1.2
21415   6
NAME        DESCRIPTION   SAMPLE1   SAMPLE2   SAMPLE3   SAMPLE4   SAMPLE5   SAMPLE6
XKR4        Xkr4          1.86      1.70      2.07      1.60      1.41      2.36
RP1         Rp1           1.15      1.11      1.43      1.50      1.01      1.27
SOX17       Sox17         88.44     61.64     85.93     75.33     60.37     59.74
MRPL15      Mrpl15        8.37      10.11     8.15      9.41      8.08      7.48
LYPLA1      Lypla1        45.15     44.86     42.55     46.36     46.72     41.23
TCEA1       Tcea1         49.47     50.43     51.15     50.46     48.70     51.35
RGS20       Rgs20         0.00      0.00      0.04      0.07      0.16      0.04
ATP6V1H     Atp6v1h       52.66     69.15     60.28     63.11     61.50     64.01
RB1CC1      Rb1cc1        158.48    132.09    139.00    122.69    152.04    139.23
```

Gene IDs

- Must be unique
- Best to use HUGO Official Gene Symbols (all caps), otherwise you'll need to supply a mapping file from your IDs to the gene symbols.

Can be anything, but you need this column

Normalized gene expression

Format for expression data

- Plain text file, suffix is .gct
- GSEA is VERY particular about these details

cribioinfo@uic.edu

GSEA file formats: .txt file

Alternate (easier) format for expression data

- Plain text file, suffix is .txt
- Same as .gct, without the 2 header lines

NAME	DESCRIPTION	SAMPLE1	SAMPLE2	SAMPLE3	SAMPLE4	SAMPLE5	SAMPLE6
XKR4	Xkr4	1.86	1.70	2.07	1.60	1.41	2.36
RP1	Rp1	1.15	1.11	1.43	1.50	1.01	1.27
SOX17	Sox17	88.44	61.64	85.93	75.33	60.37	59.74
MRPL15	Mrpl15	8.37	10.11	8.15	9.41	8.08	7.48
LYPLA1	Lypla1	45.15	44.86	42.55	46.36	46.72	41.23
TCEA1	Tcea1	49.47	50.43	51.15	50.46	48.70	51.35
RGS20	Rgs20	0.00	0.00	0.04	0.07	0.16	0.04
ATP6V1H	Atp6v1h	52.66	69.15	60.28	63.11	61.50	64.01
RB1CC1	Rb1cc1	158.48	132.09	139.00	122.69	152.04	139.23

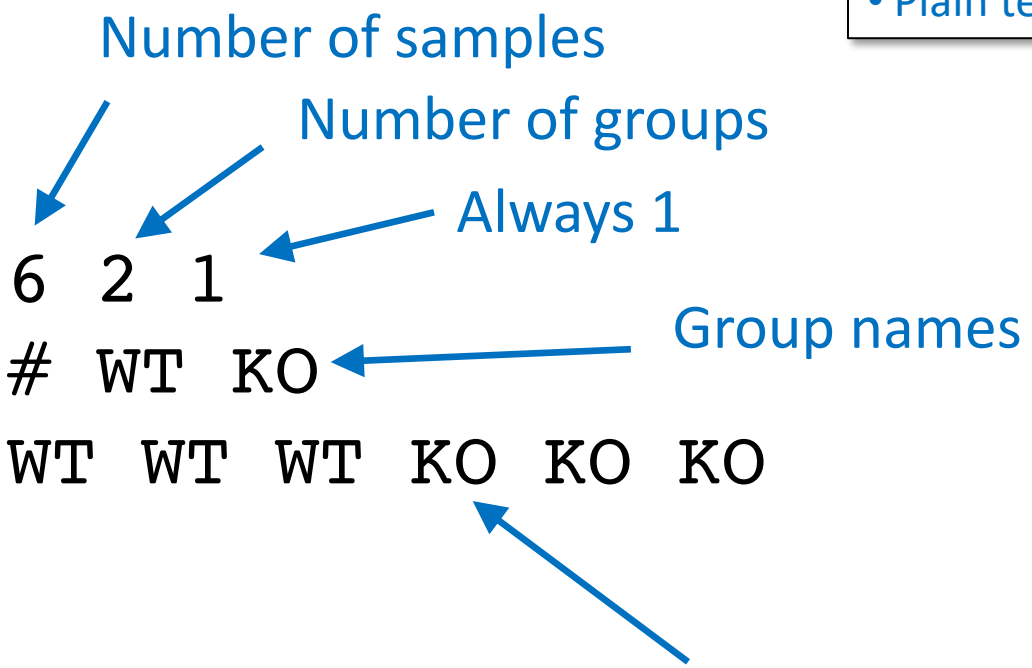
.gct is the “original” input format, and most tutorials/descriptions for GSEA use this format

No need to use .gct now, since .txt is easier.

GSEA file formats: .cls file

Format for metadata/phenotypes data

- Plain text file, suffix is .cls



cribioinfo@uic.edu

GSEA file formats: .rnk file

Format for pre-ranked gene list

- Plain text file, suffix is .rnk

```
XKR4        8.5849
RP1         12.7449
SOX17       6549.66
MRPL15      9.9225
LYPLA1      1410
TCEA1       2129.82
RGS20       24.6016
ATP6V1H     3055.88
RB1CC1      17956
```

Value to rank on

- Genes don't need to be sorted already
- GSEA will rank on this value

Gene IDs

- Must be unique
- Best to use HUGO Official Gene Symbols (all caps), otherwise you'll need to supply a mapping file from your IDs to the gene symbols.

GSEA file formats: .gmt file

Format for user-supplied gene sets (i.e., your own pathway)

- Plain text file, suffix is .gmt

```
GO:0000002  mitochondrial_genome_maintenance     AKT3          DNA2    TYMP    RNASEH1  SLC25A4
GO:0000003  reproduction                         ADA           GNPDA1  ZGLP1   SCXA     SYCE1L
GO:0000012  single_strand_break_repair           LOC100133315  APLF    SIRT1   LIG4     APTX
GO:0000018  regulation_of_DNA_recombination      RAD50         PRDM7   CHEK1   PPP4R2   PAXIP1
GO:0000019  regulation_of_mitotic_recombination  RAD50         MLH1    MRE11A  BLM
GO:0000022  mitotic_spindle_elongation           PRC1          KIF23
```

Pathway ID/name (must be unique)

Pathway description

Genes in this pathway: tab-delimited list (1 column per gene)

NOTE – if you're using your own gene sets:

- Make sure the identifiers in gene sets match the data
- Not necessary to use gene symbol as identifier

GSEA file formats: .chip file

Format mapping alternate IDs to gene symbols

- Plain text file, suffix is .chip
- Must be 3 columns with these exact names

Probe Set ID	Gene Symbol	Gene Title
ENSMUSG00000051951	XKR4	XKR4
ENSMUSG00000025900	RP1	RP1
ENSMUSG00000025902	SOX17	SOX17
ENSMUSG00000033845	MRPL15	MRPL15
ENSMUSG00000025903	LYPLA1	LYPLA1

Gene ID (must be unique)

Gene symbol (official HUGO, all caps). It's OK if there are duplicates.

Any description, or NA

NOTE: GSEA has many built-in .chip files, including a mapping for ENSEMBL IDs for human, mouse, and rat.

Running GSEA

- You need to input:
  - .gct OR .txt plus .cls file
    –OR–
  - .rnk file
- Optional inputs:
  - Custom gene sets (pathways) – .gmt file
  - Gene ID to gene symbol mapping file (if gene identifiers are not HUGO gene symbols in all caps) – .chip file
- More information:
  - GSEA user's guide:

https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html?Run_GSEA_Page

  - Data formats

http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

Exercise 1.10: GSEA

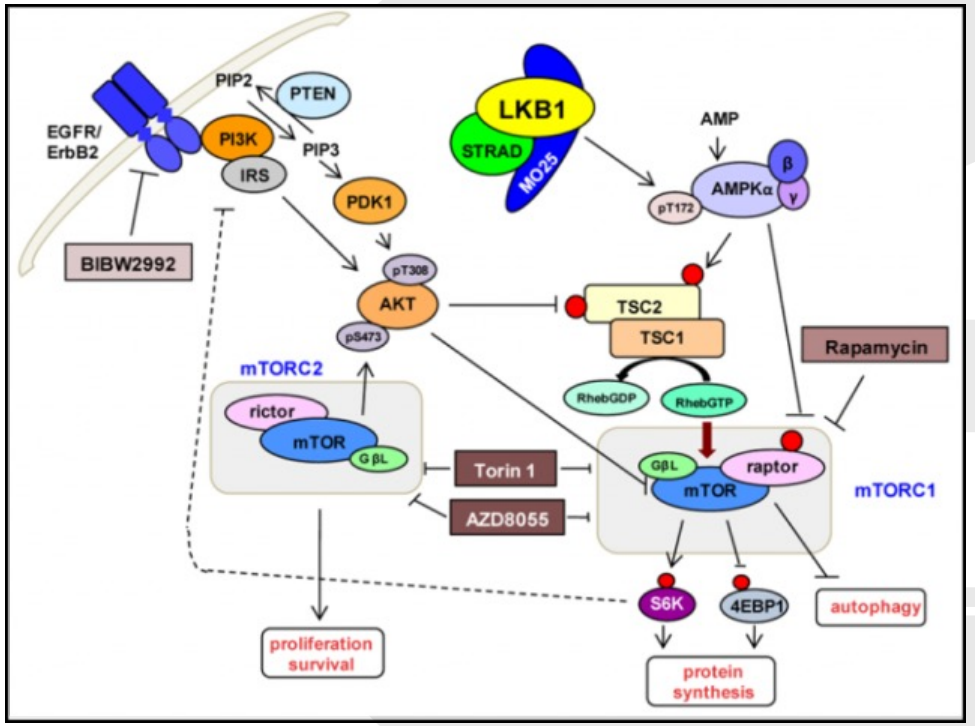
- Run GSEA with our RNA-seq data set

Fisher's Exact Test vs GSEA

- Flexibility:
  - Fisher's Exact test allows you to make more general molecule lists
    - Choose to separate or not separate up- and down-regulated genes
    - Define gene lists based on a combination of features from a clustering analysis
    - Pick genes from annotation against a ChIP-seq data set
  - GSEA limited to pair-wise comparisons
    - For >2 groups, will just run multiple pair-wise comparisons
- Thresholds:
  - Fisher's Exact Test requires you to set a threshold. All molecules are treated equally if they are above the threshold.
  - No thresholds required for GSEA, it's the extent of up- or down-regulation that matters
- **In both cases: FDR correction for multiple testing!**
  - You should ONLY consider the corrected p-value (aka FDR/Q-value/B-H p-value/adjusted p-value) for pathway significance
  - (PANTHER uses Bonferroni, which is overly conservative)

Other methods: “Topology” enrichment tests

- Inclusion of pathway structure in model
  - Recognize effects of up-regulators, down-regulators, and interactions in pathway
  - Relies on having a pathway with known interactions
- Challenges:
  - Integration of diverse regulatory effects
  - Balancing of effects of from different molecules
  - Completeness and accuracy of pathways



cribioinfo@uic.edu

Topology-ish scoring in IPA

- Pathway Z-score in IPA
  - Compare gene fold-change to what you would expect based on that gene's function in pathway
  - Aggregate scores over all genes in pathway for positive/negative z-score (i.e., pathway is up/down regulated)
  - Not a statistical test (no p-value)
  - Not computable for some pathways ("no activity pattern available")

cribioinfo@uic.edu

Topology Methods

- Others:
  - Signaling Pathway Impact Analysis (SPIA) (PMID: 17785539, 18990722)
  - EnrichNet (PMID 22962466)
  - GGEA (PMID 21685094)
  - TopoGSA (PMID 20335277)

Review of statistical methods

- Most common methods used are Fisher's Exact Test (FET) and GSEA
- In principle, choice of statistical method is independent of choice of database
  - Sometimes choice driven by convenience/availability
  - MSigDB integrated with GSEA
  - Commercial systems like IPA usually only use FET
- There are many many available tools on the web and in R
- **We recommend:**
  - IPA for most in-depth analysis
  - DAVID for a flexible free tool using FET
  - GenePattern for GSEA

LUNCH

Visualizations

Visualization of pathway results

- Depends on what you want to show:
  - Overview of enriched pathways
  - Involvement of specific biological processes
  - Comparison of pathway enrichments between gene lists
- Data to plot:
  - Fold-enrichment or enrichment score for magnitude of enrichment
  - -log10(FDR) for significance
- Plotting decision is part of your interpretation of the results
- Plan to include table of complete results in a supplementary file

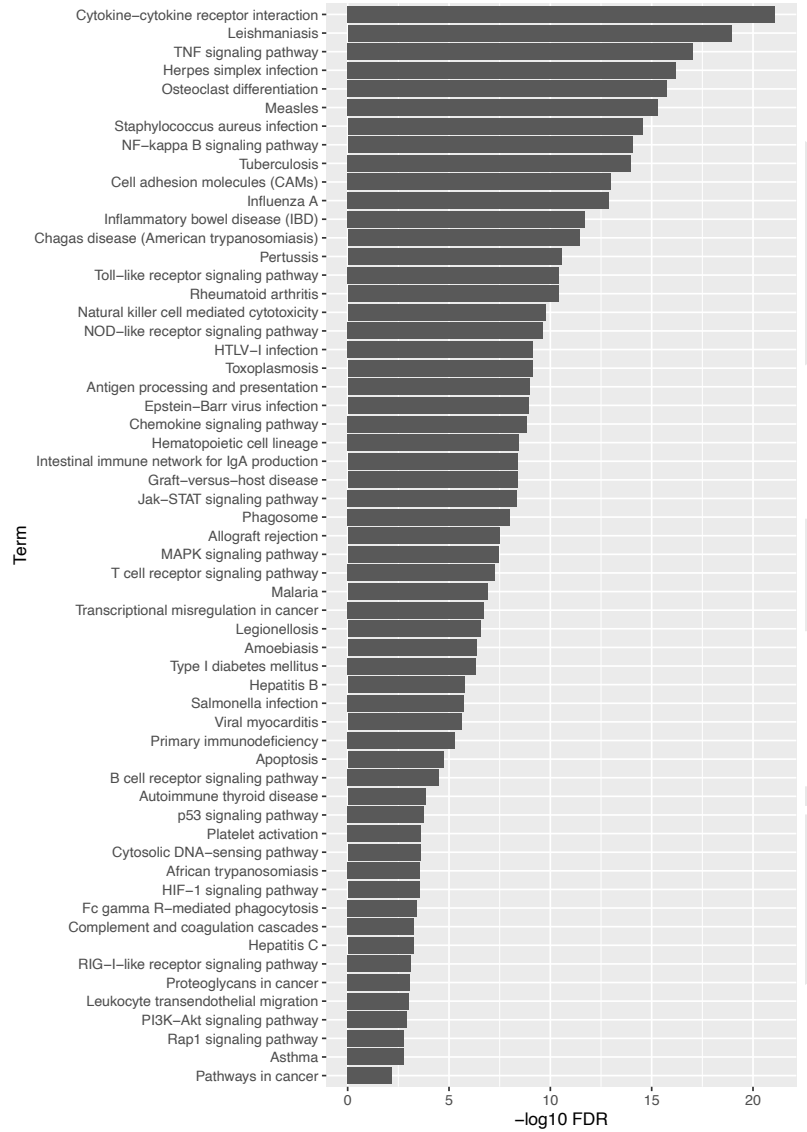
Reminder – typical results file

- Plot (log) fold-enrichment and/or –log10 FDR
- Use term names/descriptions as labels

Category	Term	Fold Enrichment	PValue	FDR
GOTERM_MF_FAT	GO:0030695~GTPase regulator activity	1.866364949	2.56E-13	4.24E-10
GOTERM_MF_FAT	GO:0030554~adenyl nucleotide binding	1.376011116	4.01E-13	6.64E-10
GOTERM_MF_FAT	GO:0001882~nucleoside binding	1.37285846	4.02E-13	6.64E-10
GOTERM_MF_FAT	GO:0001883~purine nucleoside binding	1.371364102	5.79E-13	9.58E-10
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	1.384433726	5.81E-13	9.61E-10
GOTERM_MF_FAT	GO:0005524~ATP binding	1.382215399	1.09E-12	1.80E-09
GOTERM_MF_FAT	GO:0017076~purine nucleotide binding	1.317529549	4.64E-12	7.67E-09
GOTERM_MF_FAT	GO:0032553~ribonucleotide binding	1.321934278	7.57E-12	1.25E-08
GOTERM_MF_FAT	GO:0032555~purine ribonucleotide binding	1.321934278	7.57E-12	1.25E-08
GOTERM_CC_FAT	GO:0005856~cytoskeleton	1.424110951	8.05E-12	1.19E-08
GOTERM_MF_FAT	GO:0000166~nucleotide binding	1.26884698	1.04E-10	1.72E-07
GOTERM_MF_FAT	GO:0003779~actin binding	1.838128773	3.02E-10	5.00E-07
GOTERM_MF_FAT	GO:0005083~small GTPase regulator activity	1.966265756	3.14E-10	5.19E-07
INTERPRO	IPR001849:Pleckstrin homology	1.865430688	8.13E-10	1.47E-06
INTERPRO	IPR011993:Pleckstrin homology-type	1.81428899	9.47E-10	1.71E-06
GOTERM_CC_FAT	GO:0042995~cell projection	1.547963308	1.24E-09	1.84E-06
GOTERM_MF_FAT	GO:0008092~cytoskeletal protein binding	1.62743417	6.44E-09	1.06E-05

Overview plot: bar plot of enrichments

- Simple plot to view the top pathways and their enrichment levels
- Typically plot –log10 FDR
  - Sometimes also plot enrichment statistic or log enrichment



cribioinfo@uic.edu

Exercise 2.1: Barplots

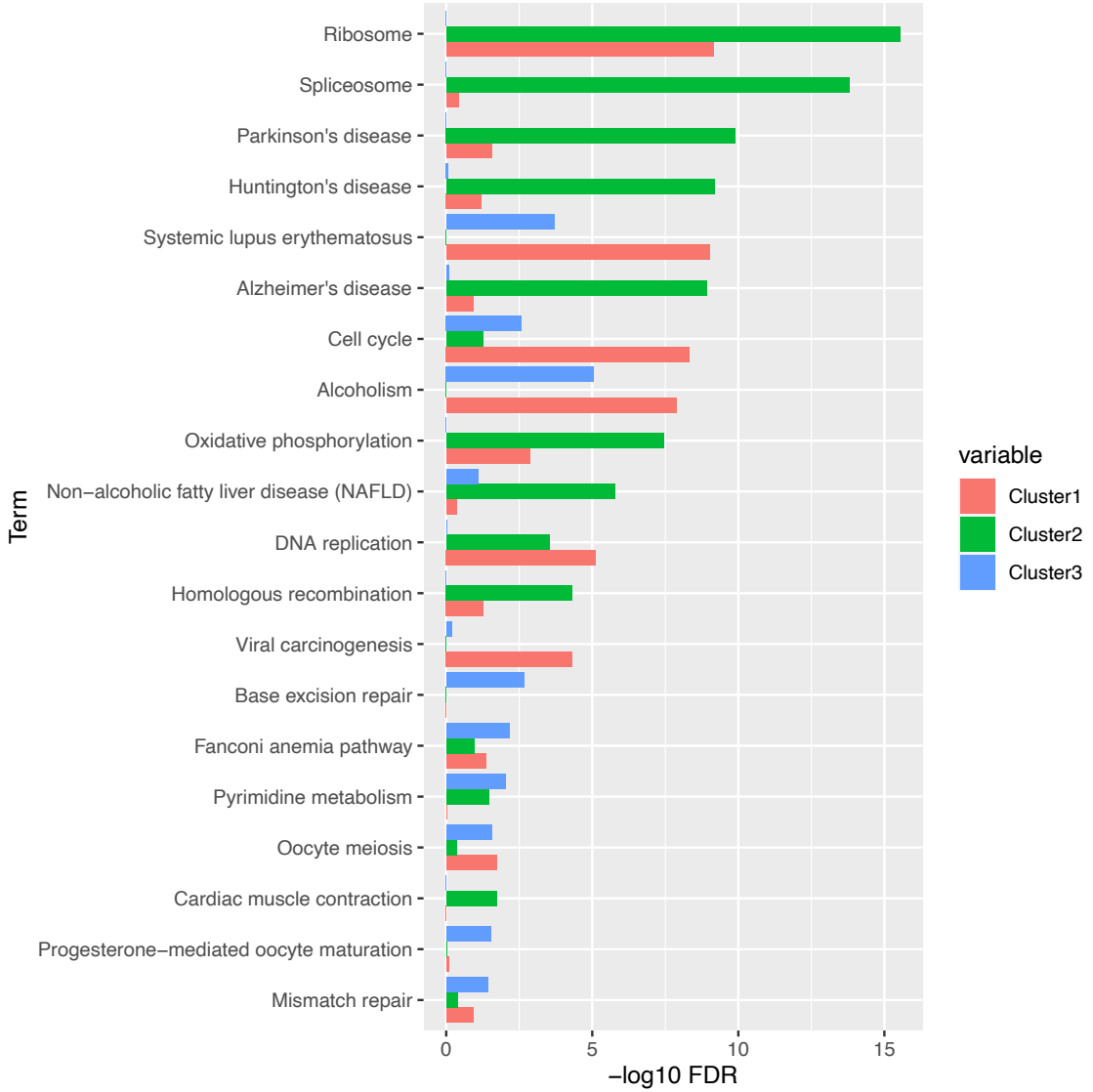
- Make barplots of
  1. Fold-enrichment
  2. -log10 FDR

Comparison of pathway enrichments

- Compare pathway enrichment for multiple molecule lists
  - Up vs genes
  - Differedown-regulatednt gene clusters
  - Different –omics data sets
  - Different downstream processing strategies
- Want to compare which pathways are enriched, and to what extent

Side-by-side barplots

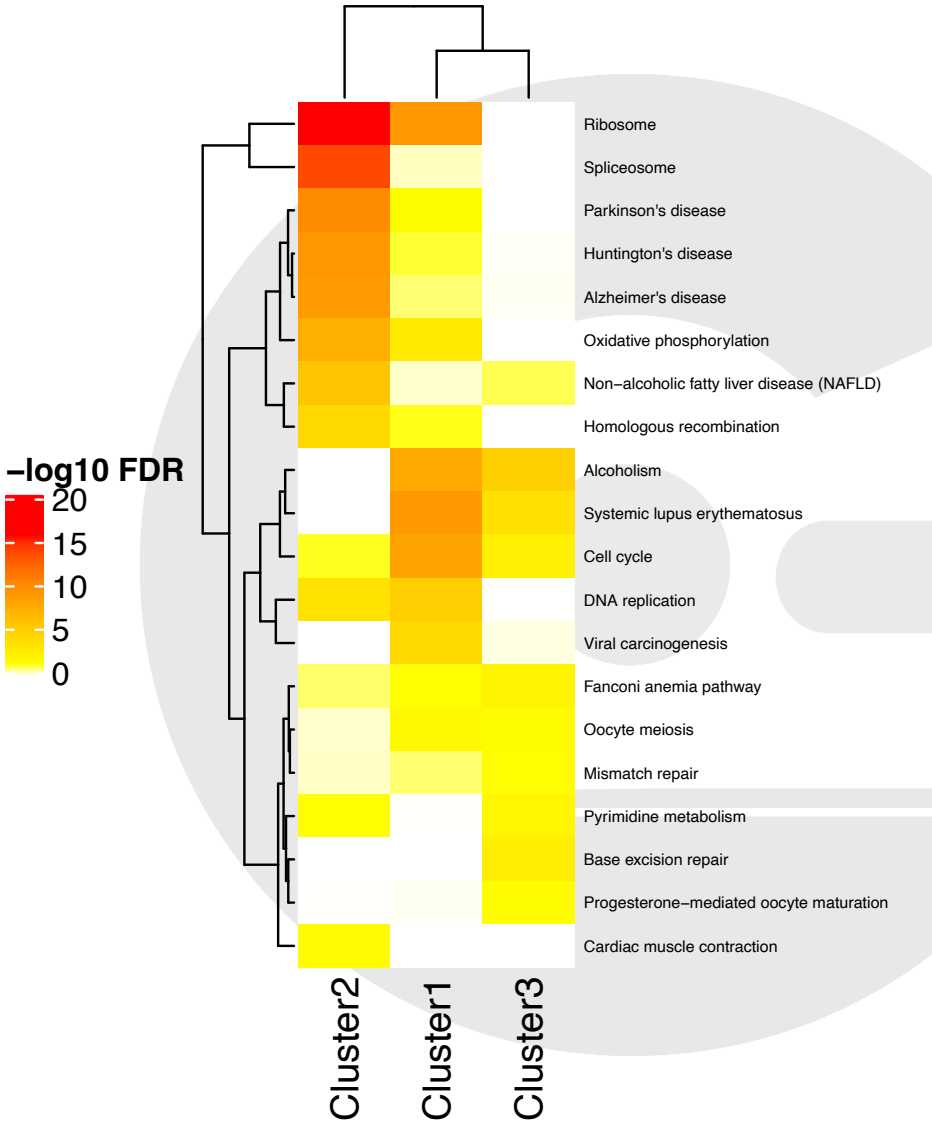
- Easily see relative enrichment & significance levels
- Becomes unreadable for too many data sets



cribioinfo@uic.edu

Heatmaps

- Comparison of many molecule lists
- Clustering of terms and data sets based on enrichment similarity
- Need to adjust color scale appropriately



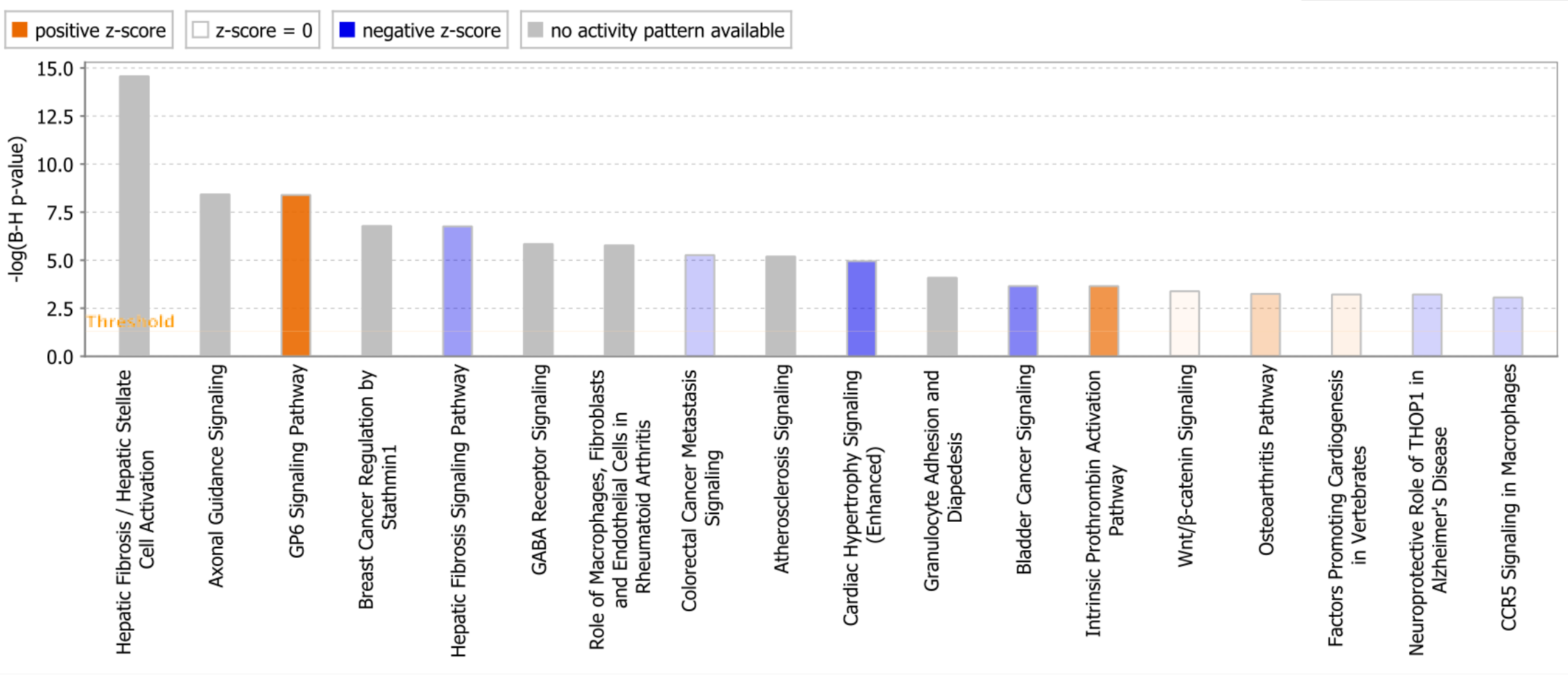
cribioinfo@uic.edu

Exercise 2.2: Comparison plots

- Side-by-side plot for 2 gene lists
- Heatmap for many gene lists

Visualization in IPA: barplots

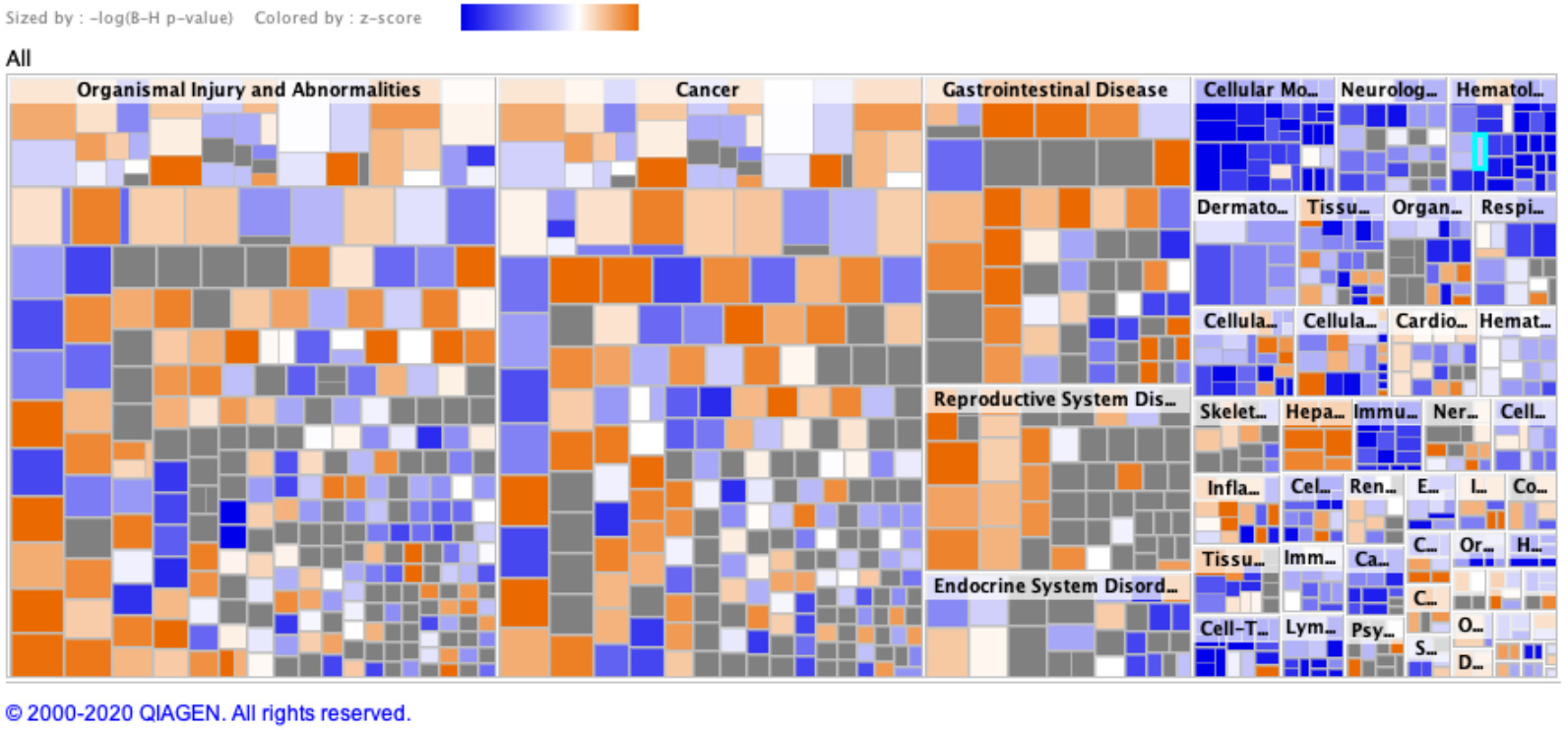
Pathway enrichment in –log10 scale, Z-score shown by color shading



cribioinfo@uic.edu

Visualization in IPA: “tree heatmap”

- Diseases and functions
  - Square size = -log10 B-H p-value (significance)
  - Square color = z-score
  - Grouped by rough category

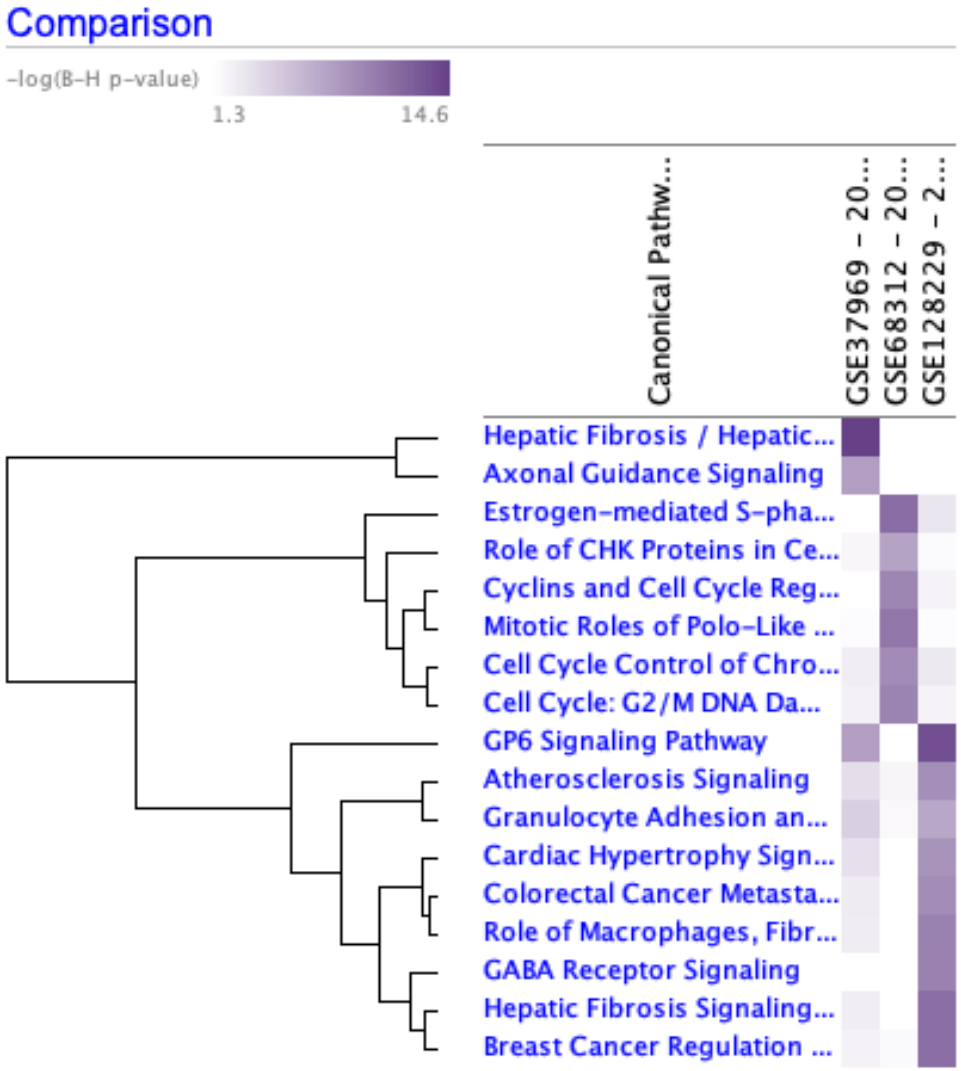


Kind of like a barplot, but more confusing…

Barplot visualization is available as well

Visualization in IPA: comparison heatmap

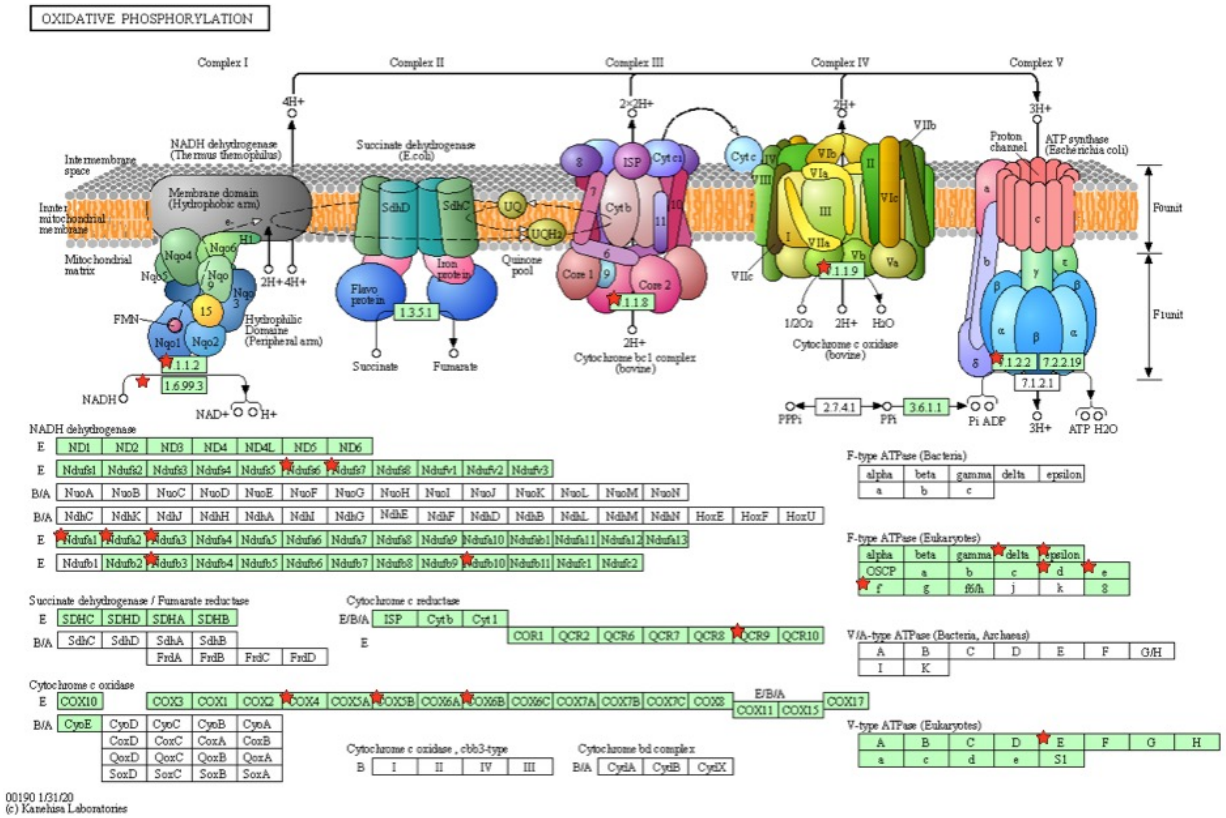
- Same idea as our comparison heatmap
- To compare multiple data sets:
  - Select >1 core analysis runs
  - Right-click > New Comparison Analysis
- Comparisons visualized in heatmaps, can export as images or tables



cribioinfo@uic.edu

Visualization of specific pathway maps

- Pathway map
  - Visualization for discussion of pathway function
  - Show where the molecules in your list are



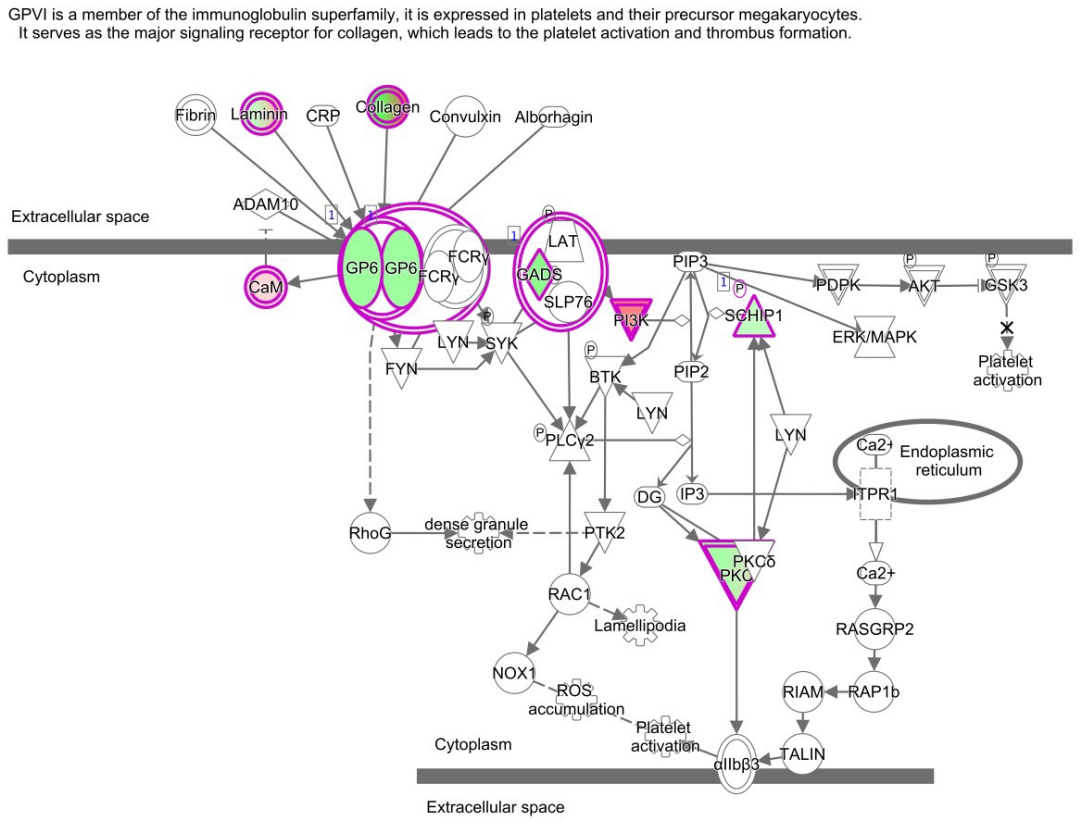
KEGG pathway for Oxidative Phosphorylation

- Can be generated from DAVID
- ★s are molecules in your data set

cribioinfo@uic.edu

Visualization of specific pathway maps

- Pathway map
  - Visualization for discussion of pathway function
  - Show where the molecules in your list are



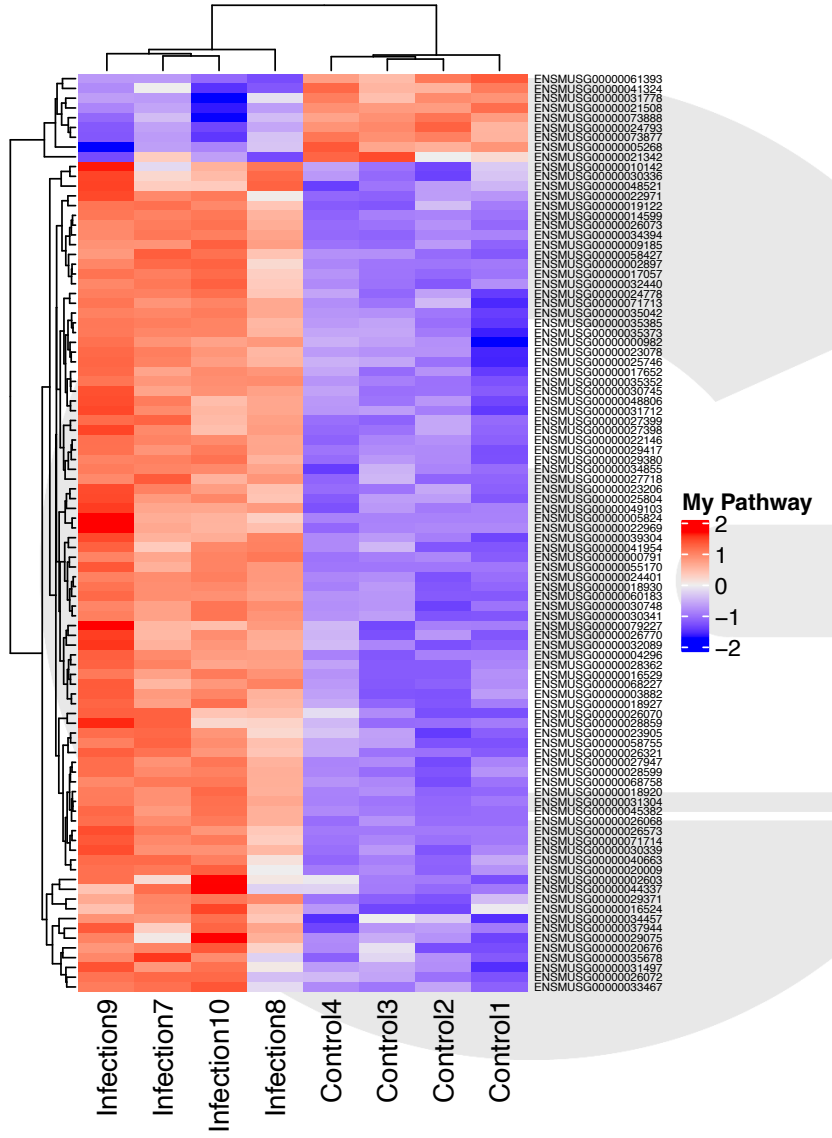
IPA pathway for GP6 signaling

- Purple outlines are molecules in your data set

cribioinfo@uic.edu

Expression heatmap for a pathway

- Show expression patterns of molecules in your heatmap
- Reminder: GSEA produces these automatically
  - Click on the “Details” link in the detailed results
  - Scroll to the bottom (top is the leading edge figure)



cribioinfo@uic.edu

Exercise 2.3: Pathway expression heatmaps

- Plot expression patterns for differentially expressed genes in the top KEGG pathway

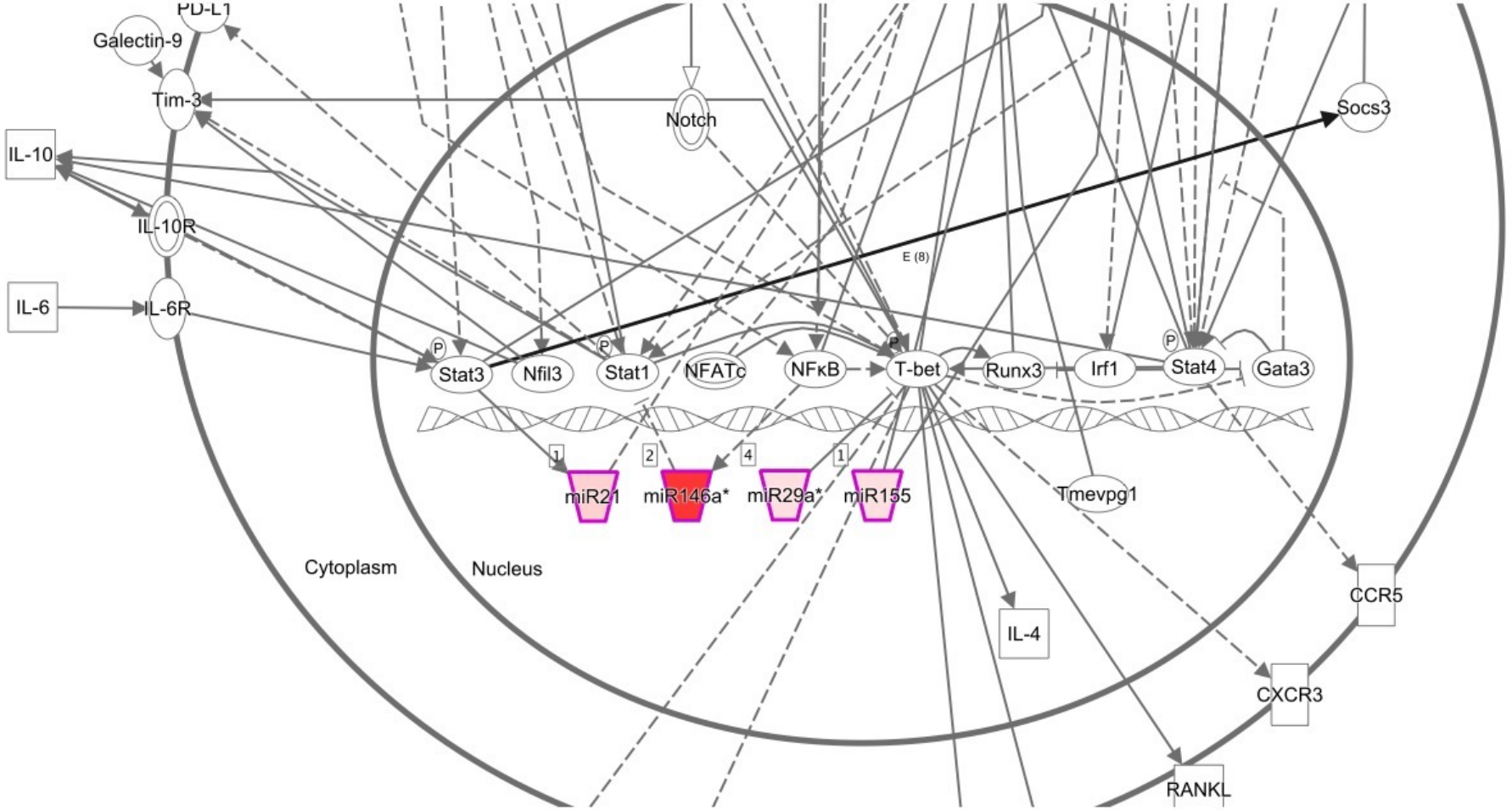
cribioinfo@uic.edu

Pathway analysis with other –omics data

- Most pathways are annotated based on genes/proteins
  - Easiest to annotate with gene or quantitative protein data
  - Isoforms probably need to be considered at gene-level
- *Some* databases include other molecules
  - miRNAs (IPA – validated targets)
  - Metabolites/small molecules (IPA, KEGG)
  - Post-translational modifications (IPA)
- Other –omics data needs to be mapped to genes
  - Epigenomics: TF binding sites, open chromatin, CpGs
  - Genomics: variant calls
  - miRNAs: map to target genes (decide validated vs predicted)

Example of other molecules in pathways

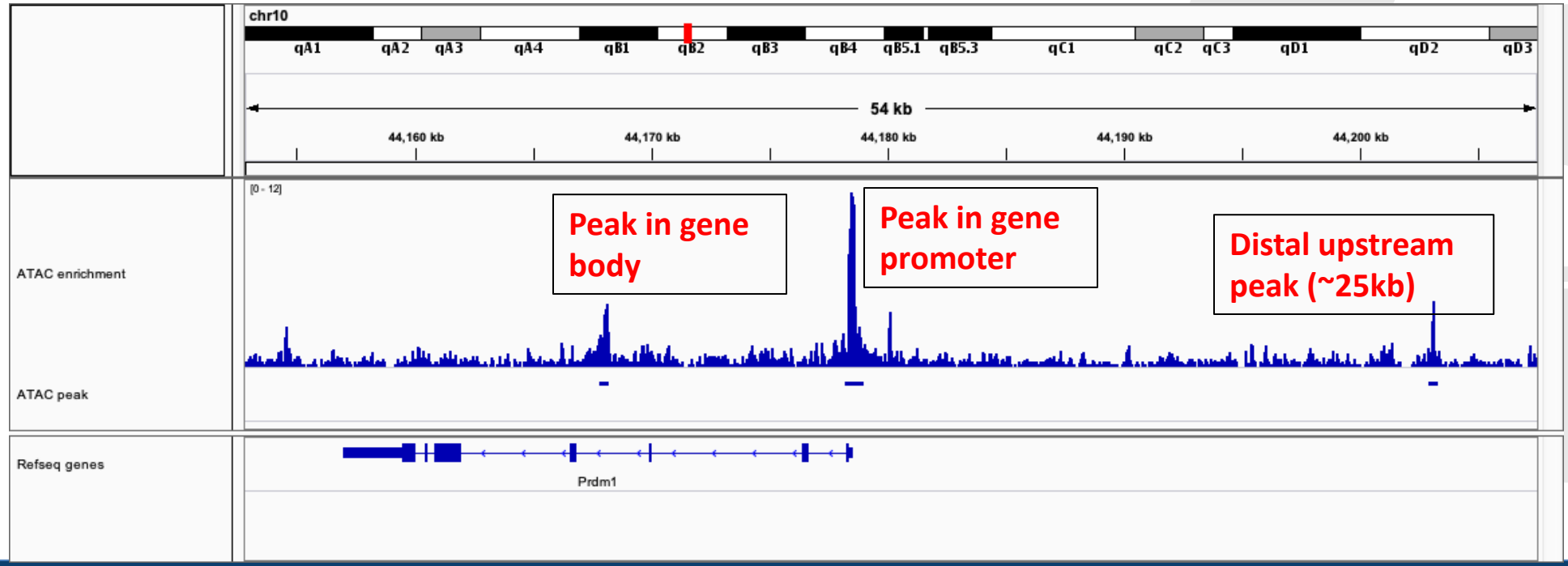
- miRNAs (highlighted) in Th1 pathway (IPA)



cribioinfo@uic.edu

Epigenomics: Annotating to gene level

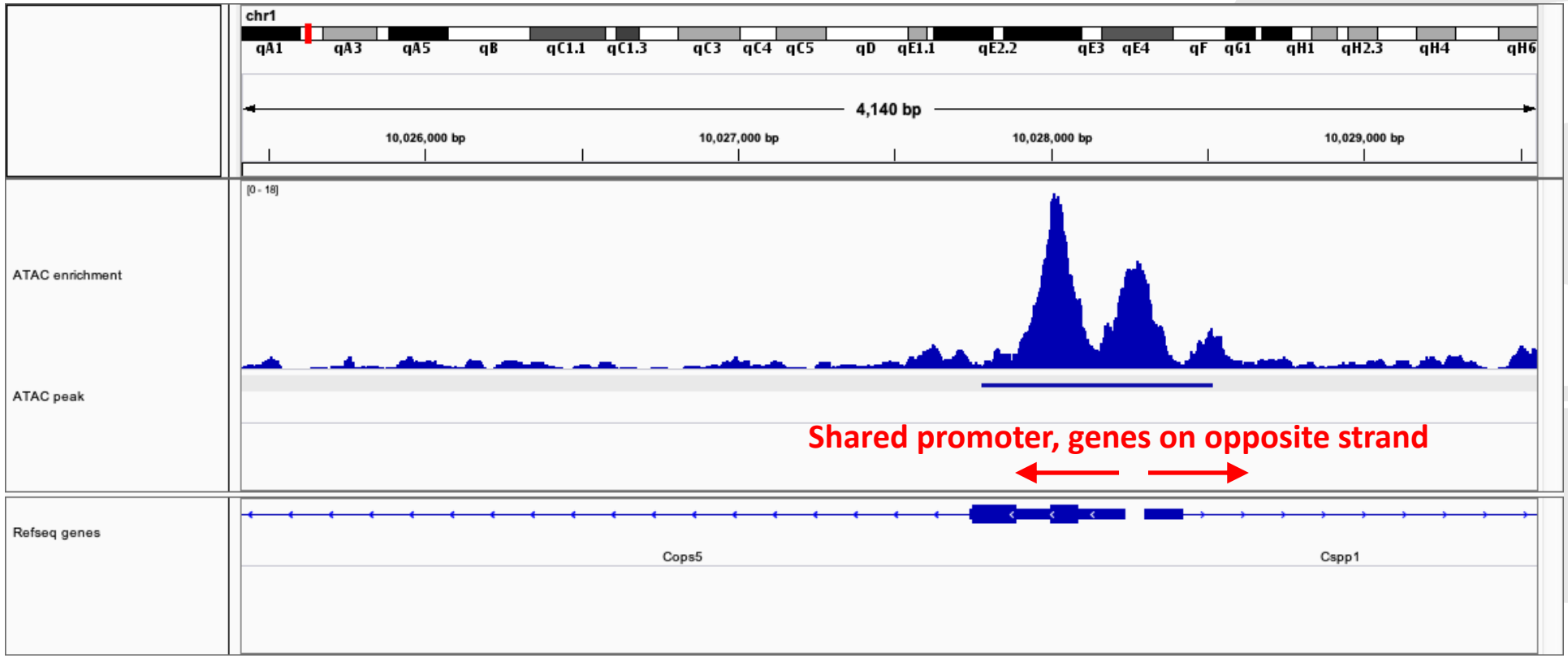
- ChIP-seq or ATAC-seq
  - Quantified features are “peaks”: regions of enrichment in genome
- DNA methylation (bisulfite)
  - Quantified features are CpGs, or group of CpGs
- In both cases: need to map a genomic interval/location to a gene
  - Annotation may not be one-to-one



cribioinfo@uic.edu

Annotation options

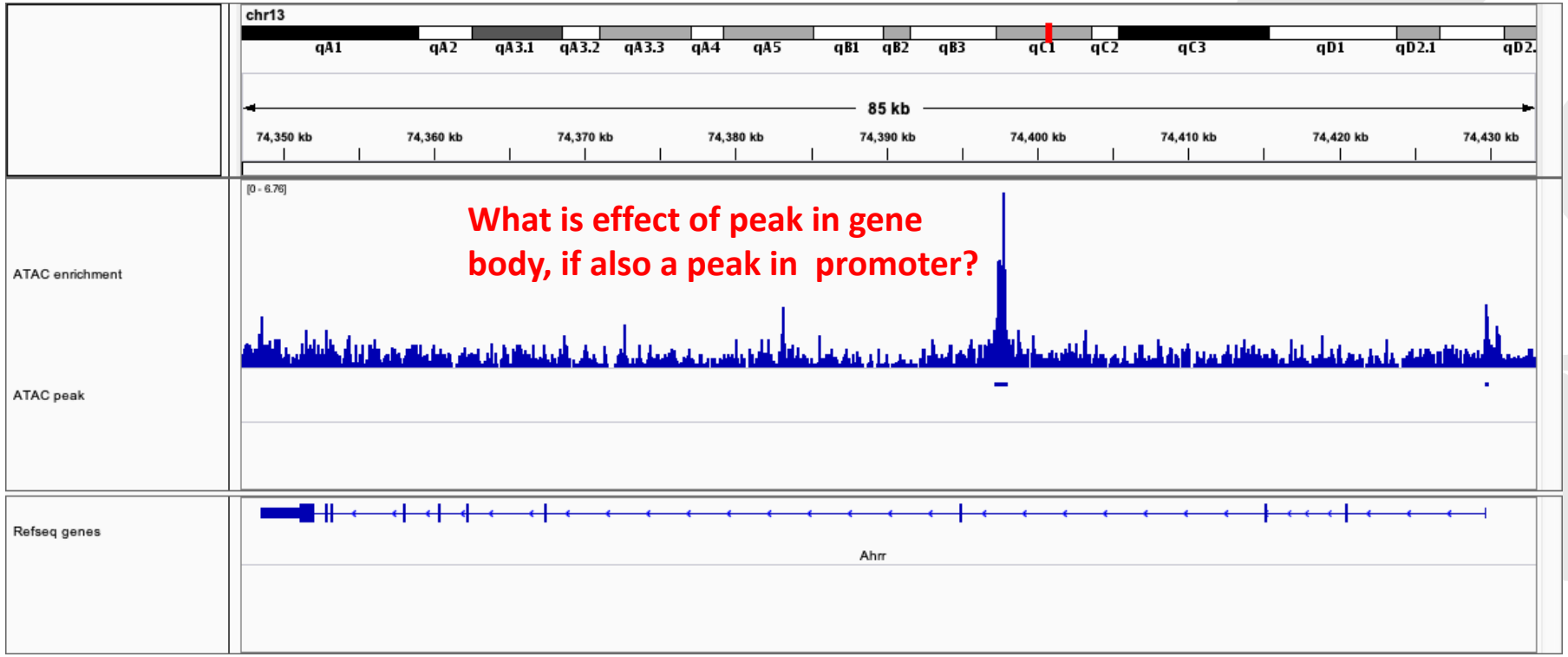
- Promoter (+/-2kb from TSS)
  - *Usually* one-to-one
  - Many peaks will be intergenic



cribioinfo@uic.edu

Annotation options

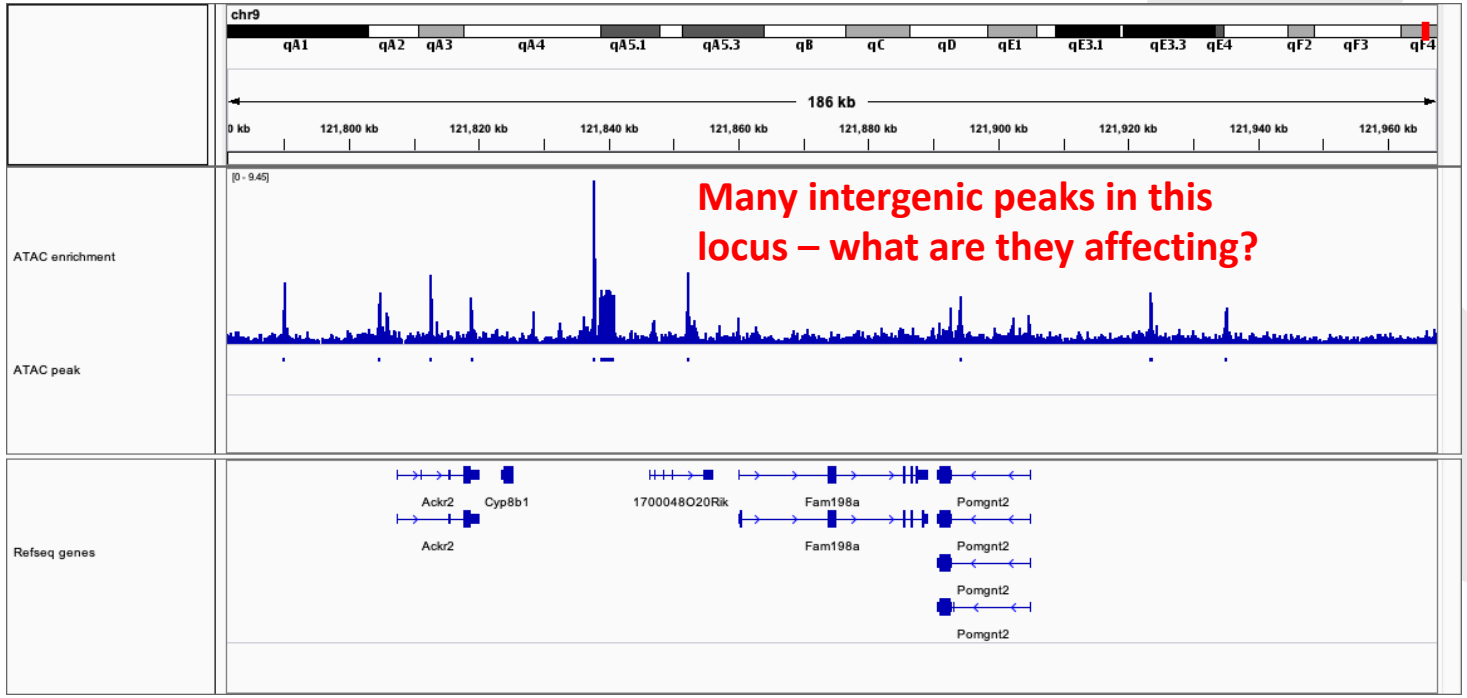
- Gene body (between TSS and TSE)
  - *Usually* one-to-one; some genes overlap on opposite strands
  - Relationship of peaks far from TSS may be unclear



cribioinfo@uic.edu

Annotation options

- Distal
  - Will capture the most relationships, but will be many-to-many
  - Expect weaker effects on average
  - Set a distance cutoff
  - (If data available) Use TAD data from Hi-C, or associations from CHIA-PET



cribioinfo@uic.edu

Annotation strategy

- Simplest option: annotate to promoters
- More advanced: include some distal annotations
  - Look at correlations in expression vs enrichment between pairs of genes and (distal) nearby peaks
  - Other strategies may depend on context of epigenetic mark & expected function

Pathway analysis for genomic variants

- SNPs, indels, CNVs, etc.: infer effect on biological systems
- Association with genes is more straightforward: probably include just exonic variants
  - Possibly look at intronic if it affects splicing
  - Possibly look at promoter or other annotated regulatory element if it affects TF binding, etc.
- **Primary challenge: filtering variants to get gene list**

Variant features

- Effect on translated exonic sequence
  - Synonymous/nonsynonymous (+stop gain, stop loss, frameshift)
  - Predicted effect on protein function (models)
    - SIFT
    - PolyPhen
    - Dozens more…
- Mutation burden
  - Number of nonsynonymous/damaging variants per gene

Exercise 2.4: pathway enrichment for variant calling

- Start with variant calls in tabular format
- Filter with 2 strategies:
  - Genes with damaging variants from SIFT model
  - Genes with high mutation burden of non-synonymous variants
- Run both through pathway analysis (DAVID GO BP)
- Compare to each other

BREAK

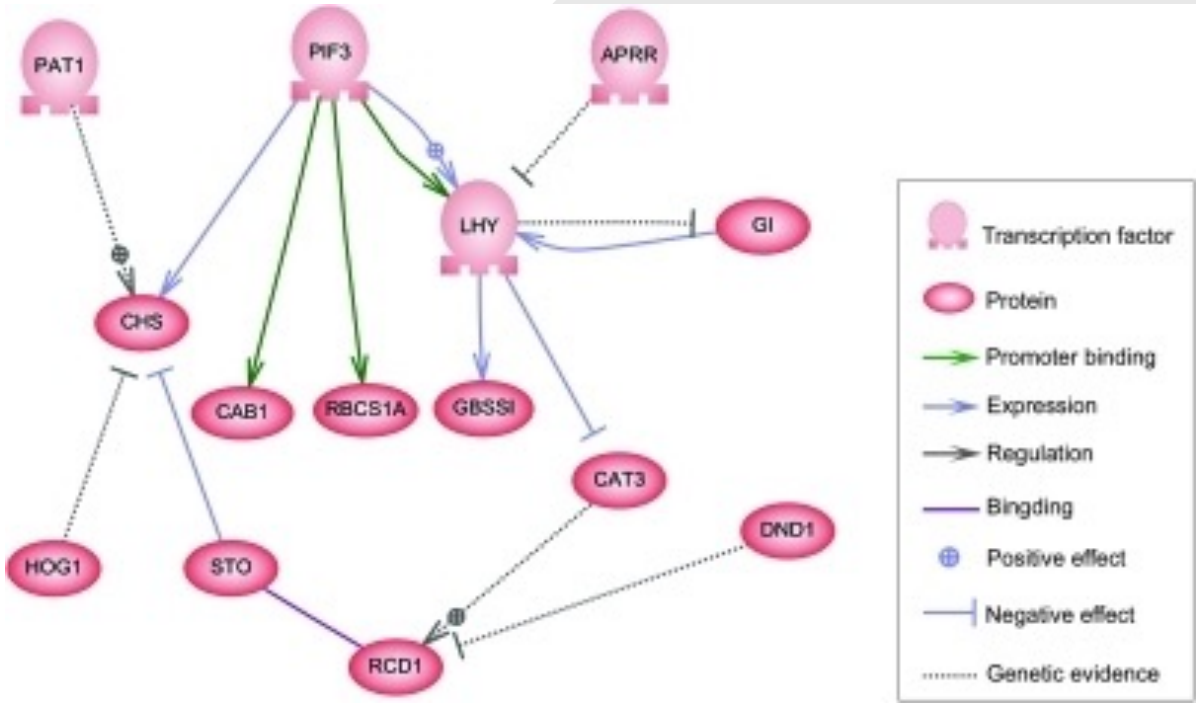
Please complete our workshop survey

http://go.uic.edu/RICWorkshopSurvey

Network Analysis

Network structure

- Each node is a biological entity
  - Molecule, complex, etc.
- Each edge is an interaction
  - Binding, catalysis, phosphorylation, expression, etc.
  - Activating or inhibiting
  - May be directed or undirected
- Information is curated from literature/experiments
- Higher-level process not curated
  - Network can be as big or small as you like



cribioinfo@uic.edu

Pathway vs. Network

Pathway

- Curated/defined group of molecules and interactions for a specific biological process, e.g. EGFR pathway, glycolysis
- Stored as list of molecules and pathway memberships

Network

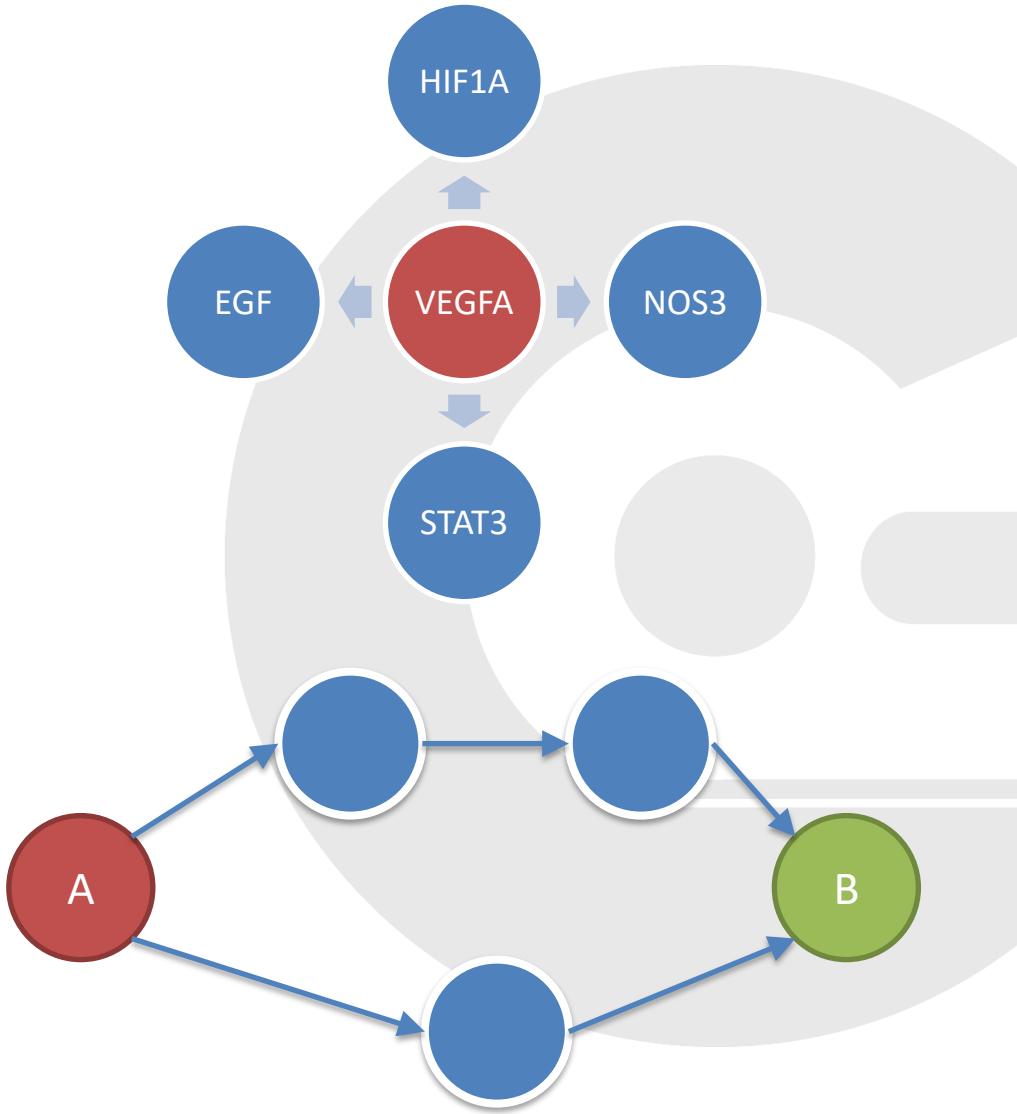
- All possible biological interactions between molecules.
  - Molecules are **nodes**
  - Interactions/relationships between molecules are **edges**
- Relationship data are not constrained to defined pathways.
  - Can be subset or superset of pathway(s)

Goals of a network analysis

- Build on interaction knowledgebase
  - Unconstrained by pathway curations
- Formulate broader hypotheses to test:
  - What else interacts with molecules of interest?
  - What interactions connect molecules of interest?
  - How do these fit into my organism/dataset?

Strategies in Network Analyses

- **Grow:** Start with one or a few molecules and grow/expand network to reveal interacting molecules. Can filter based on…
  - Type of interactions (activation, inhibition, catalysis, binding, phosphorylation, etc.)
  - Type of molecules (transcription factor, receptor, miRNA, metabolite, etc.)
- **Find Paths:** Have two sets of molecules and find "paths" connecting the molecules. Can use similar filters as with "grow" strategy.



cribioinfo@uic.edu

Source and Databases

- Different types of interaction databases
- Protein Protein Interactions (PPI)
  - STRING, WikiPathways, IntAct, BioGRID
- Chemical-Protein Interactions
  - STITCH
- miRNA–mRNA Interactions
  - miRbase, miRNet, HMDD (the Human microRNA Disease Database)
  - https://tools4mirs.org/ (set of computational and annotation resources)
- Commercial tools – IPA, MetaCore (include all of the above)

STRING: known & predicted protein-protein interactions

- Includes both direct (physical) and indirect (functional) associations
- Interactions are derived from:
  - Genomic context predictions
  - High throughput experiments
  - Automated Text mining
  - Previous knowledge base
  - Conserved co-expression data
- Coverage
  - Currently has 24,584,628 proteins from 5090 organisms (4445 Bacteria, 477 Eukaryotes, 168 Archaea).
- Web user interface supports
  - Single protein search, multiple proteins
- STITCH: similar to STRING, but also includes small molecules (metabolites, drugs, etc.)

Exercise 2.5 & 2.6: STRING and STITCH

- STRING: Search protein by name
  - ER-beta
- STRING: Search by Multiple proteins
  - antiviral gene list.
- STITCH: Search by molecule name
  - Estradiol (endogenous small molecule)
  - ER-beta (protein)
  - Tamoxifen (drug)

miRNA annotations

- miRbase (http://www.mirbase.org)
  - Repository of annotated miRNAs
  - Contains mostly published miRNAs
  - Provides direct experimental evidence
  - Has search feature
- miRNet (https://www.mirnet.ca/)
  - miRNA-target interaction networks
  - functional enrichment analysis
  - miRNA target identification
  - Collects data from multiple sources (including miRbase)
  - Network building and visualization
- miRTarBase (http://mirtarbase.cuhk.edu.cn/php/index.php)
  - miRNA-target interactions
  - Has search feature for e.g. search by miRNA, Target Gene, Pathway (KEGG)

Exercise 2.7: miRNet

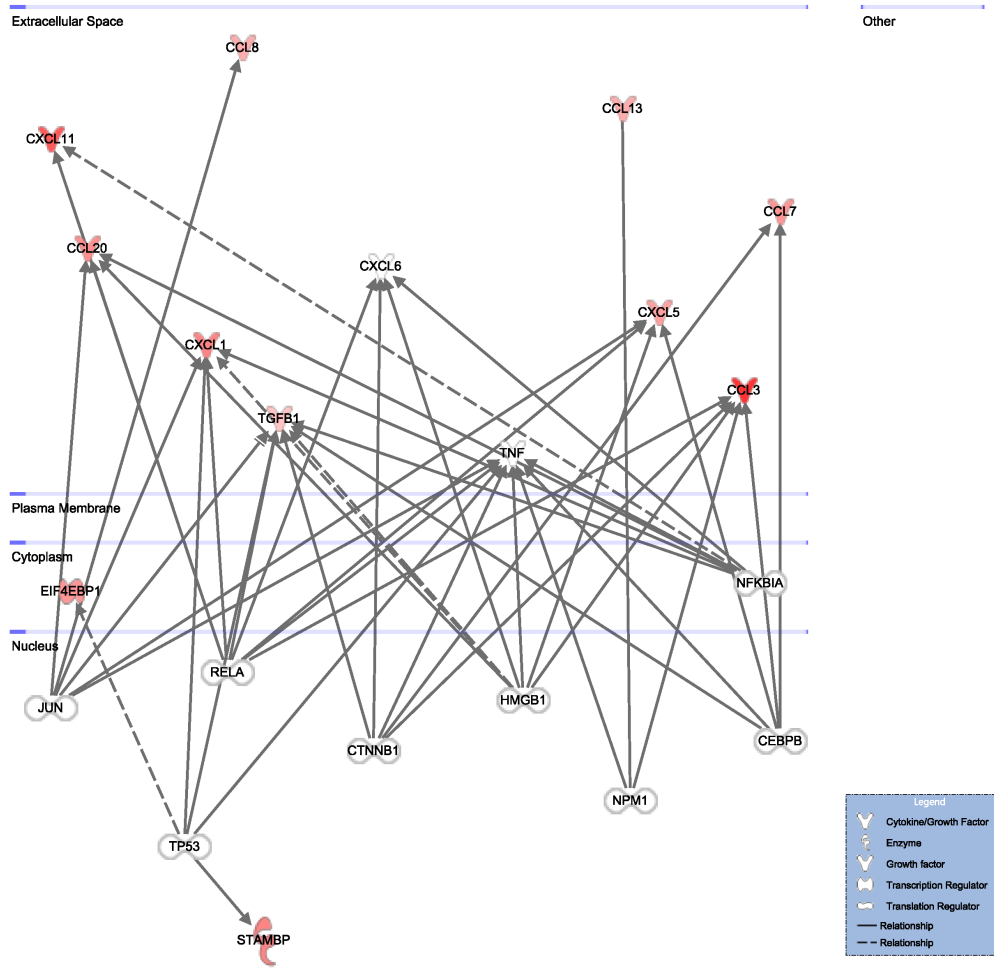
- For a given list of miRNA – identify gene and lncRNA target
  - Browse interactions for miRNA - gene target
  - Browse interactions for miRNA – lncRNA target
  - Build and visualize “Minimum Network” with seed nodes.
- Detailed tutorial
  - https://www.mirnet.ca/miRNet/docs/Tutorial.xhtml

Ingenuity Pathway Analysis (IPA)

- All content is built on molecular interaction database
  - Manually curated
- One can search for gene, proteins, drugs and chemicals
- Data for model organisms: human, mouse and rat
- Network analysis with data or without data
  - Without data
    - Use molecule(s) of interest as seed. Grow the network around this molecule.
  - With data
    - Input your gene(s)/molecules or analysis
    - Perform core analysis for pathway enrichment for your data
    - Use pathway of interest to expand/grow

IPA network analysis

- **GROW** Tool – Expand a network from one or more entities (akin to *More* feature in STRING/STITCH)
- **PATH EXPLORER** – Find shortest paths between two sets of molecule(s)
- Both tools can use a number of filters to find specific relationships/molecules.
  - Grow example: Find upstream transcription regulators of a protein/gene
  - Path explorer example: Find other transcriptional regulators associated with regulation of a gene(s).
- Can use *Overlay* functions of IPA to color nodes with experimental results, e.g. log2FC or FDR corrected p-values.
- Can use *Path Designer* to graphically edit the pathway (change shapes, colors, add images, etc.)



Exercise 2.8: IPA network analysis

- Grow a network from a single molecule
  - Find downstream growth factors that are (transcriptionally) regulated by ER-beta
- Find a path between two molecules
  - Find transcription regulators that mediate a relationship from ER-beta to p53.

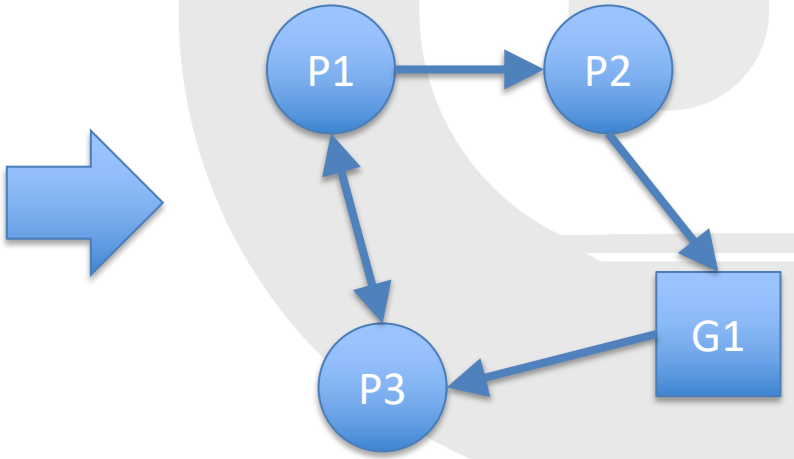
Visualization Tools

- Many visualization tools are available for networks visualizations
  - Many resources include these tools
- Some popular ones:
  - Cytoscape: Network visualization and analysis tool, including >200 plugins
  - STITCH: interaction networks of chemicals and proteins
  - COPASI: Biochemical System Simulator
- Basic input is a list of interactions

From	To	Type
Protein1	Protein2	Phosphorylation
Protein2	Gene1	Transcriptional regulation
Gene1	Protein3	Transcription
Protein3	Protein1	Binding

From/To are essential

Other features may include type, direction, activation/inhibition, weight (strength)



Summary

- A number of different databases and tools can be used for systems biology
- There are two primary modes of systems biology, which to use depends on the goal/question of the project.
- *Pathway enrichment*: Finding over-represented pathways (molecule sets) in experimental data (RNAseq, ChIPseq, proteomics, metabolomics)
  - *What pathways are effected by the experimental conditions?*
  - Different tests can be used (Fisher's exact, GSEA leading edge, etc.)
  - Key step is defining your "molecules of interest" (q value, log2FC cutoffs)
- *Network analysis*: Finding interactions between molecules
  - What are the upstream transcription factors in common for my set of molecules?
  - What are the downstream kinases for this receptor?
  - What are the intermediate molecules between two sets of proteins?