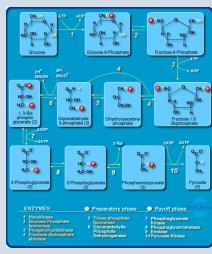
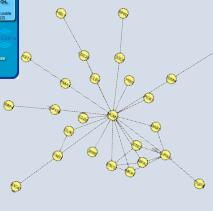


Module 1
Introduction to Pathway and Network Analysis
of Gene Lists

Gary Bader
Pathway and Network Analysis of -omic Data
June 13, 2016

 bioinformatics.ca

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

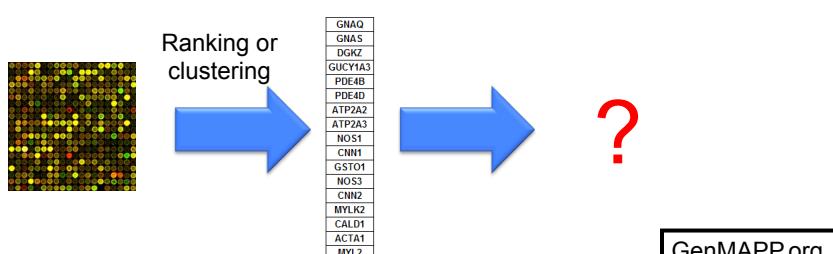

Donnelly Centre
for Cellular + Biomolecular Research

UNIVERSITY OF TORONTO


<http://baderlab.org>

Interpreting Gene Lists

- My cool new screen worked and produced 1000 hits! ...Now what?
- Genome-Scale Analysis (Omics)
 - Genomics, Proteomics
- Tell me what's interesting about these genes



Ranking or clustering → ?

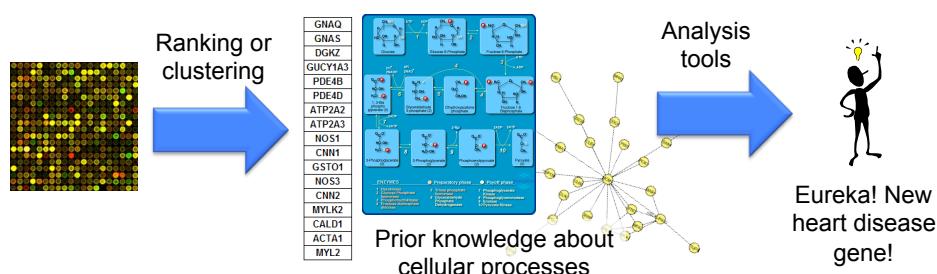
GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

GenMAPP.org

Module 1: Introduction to Pathway and Network Analysis
bioinformatics.ca

Interpreting Gene Lists

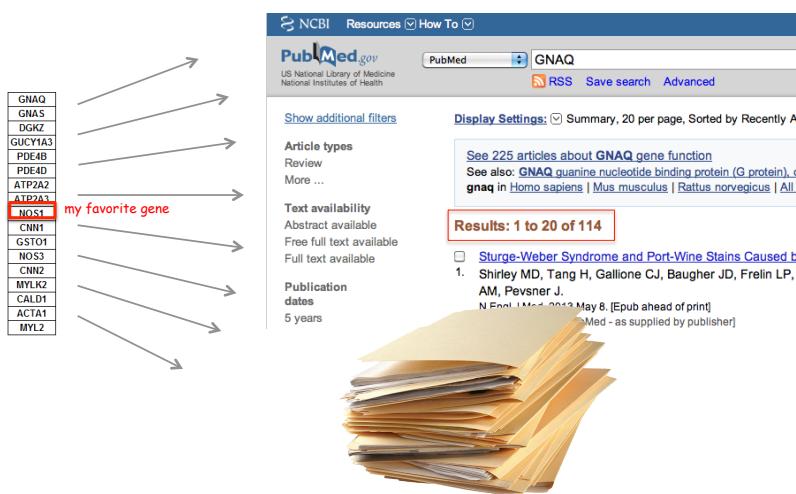
- My cool new screen worked and produced 1000 hits! ...Now what?
- Genome-Scale Analysis (Omics)
 - Genomics, Proteomics
- Tell me what's interesting about these genes
 - Are they enriched in known pathways, complexes, functions



Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Pathway and network analysis

- Save time compared to traditional approach



Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Pathway and Network Analysis

- Helps gain mechanistic insight into ‘omics data
 - Identifying a master regulator, drug targets, characterizing pathways active in a sample
- Any type of analysis that involves pathway or network information
- Most commonly applied to help interpret lists of genes
- Most popular type is pathway enrichment analysis, but many others are useful

Pathway analysis example 1

Autism Spectrum Disorder (ASD)

- Genetics
 - highly heritable
 - monozygotic twin concordance 60-90%
 - dizygotic twin concordance 0-10%
(depending on the stringency of diagnosis)
 - known genetics:
 - 5-15% rare single-gene disorders and chromosomal rearrangements
 - de-novo CNV previously reported in 5-10% of ASD cases
 - GWA (Genome-wide Association Studies) have been able to explain only a small amount of heritability

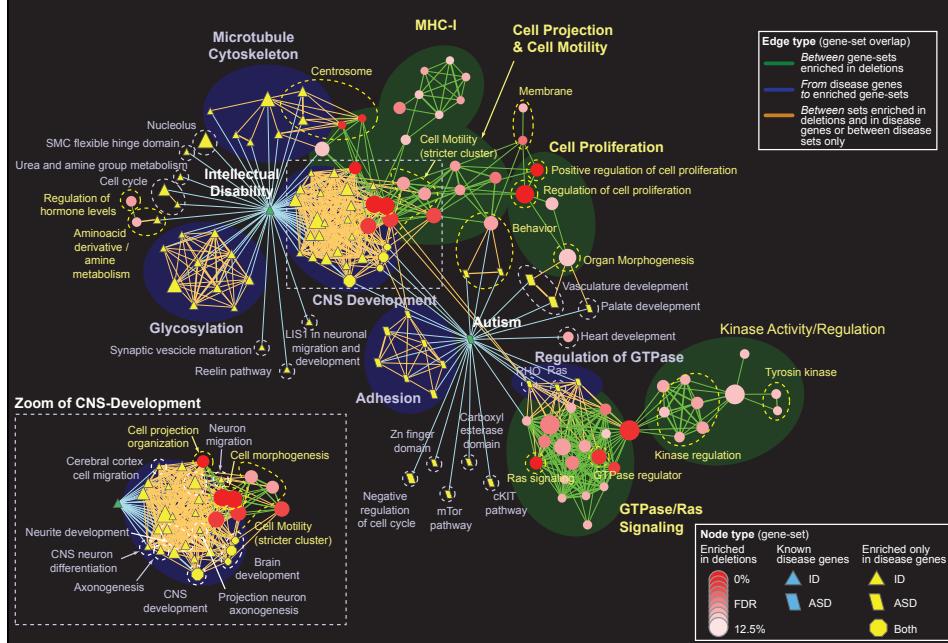
Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.

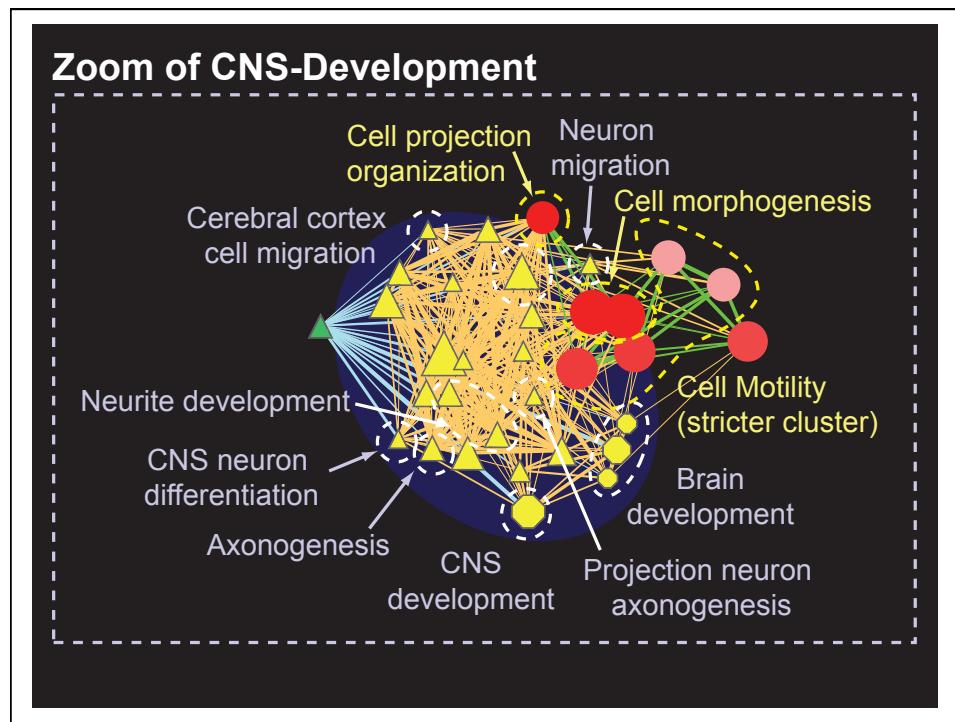
Rare copy number variants in ASD

- Rare Copy Number Variation screening (Del, Dup)
 - 889 Case and 1146 Ctrl (European Ancestry)
 - Illumina Infinium 1M-single SNP
 - high quality rare CNV (90% PCR validation)
 - identification by three algorithms required for detection
 - QuantiSNP, iPatter, PennCNV
 - frequency < 1%, length > 30 kb
- Results
 - average CNV size: 182.7 kb, median CNVs per individual: 2
 - > 5.7% ASD individuals carry at least one de-novo CNV
 - Top ~10 genes in CNVs associated to ASD

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Pathways Enriched in Autism Spectrum





Pathway analysis example 2

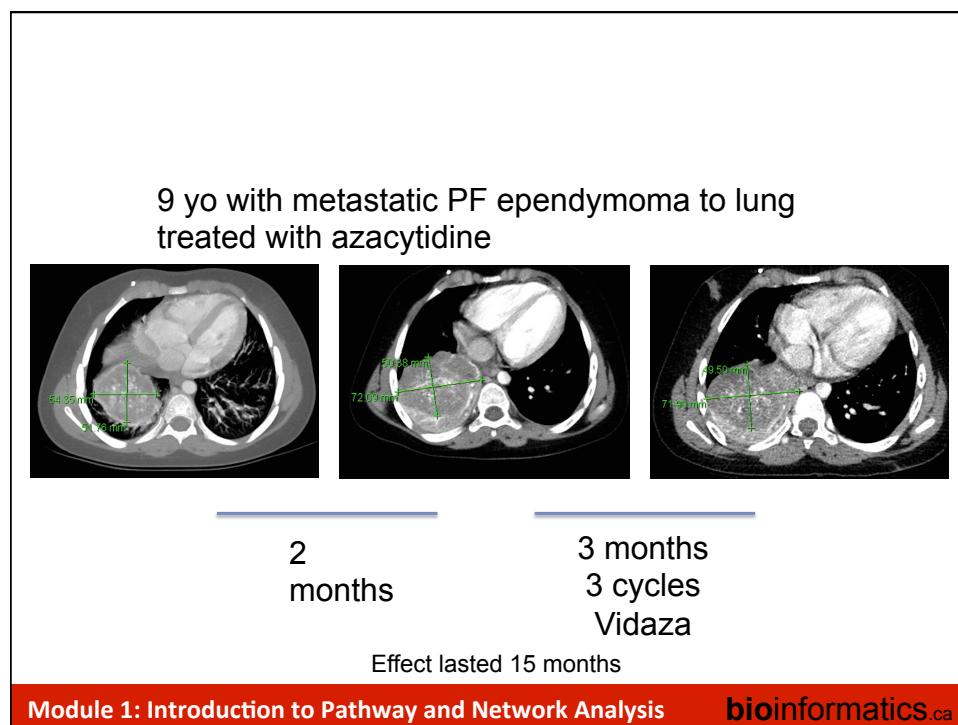
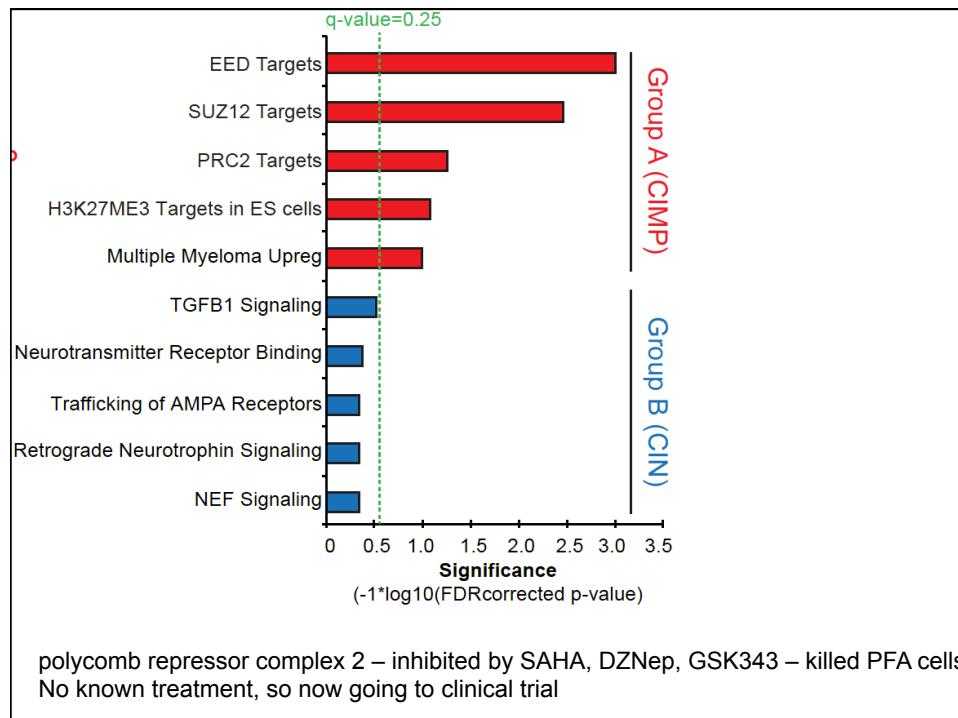
Ependymoma Pathway Analysis

- Ependymoma brain cancer - most common and morbid location for childhood is the posterior fossa (PF = brainstem + cerebellum)
- Two classes: PFA - young, dismal prognosis, PFB - older, excellent prognosis. Determined by gene expression clustering.
- Exome sequencing (42 samples), WGS (5 samples) showed almost no mutations, however methylation arrays showed clear clustering into PFA and PFB (79 samples)
- PFA more transcriptionally silenced by CpG methylation

Witt et al., Cancer Cell 2011

Nature. 2014 Feb 27;506(7489):445-50

Steve Mack, Michael Taylor, Scott Zuyderduyn



Benefits of Pathway Analysis

vs. transcripts, proteins, SNPs...

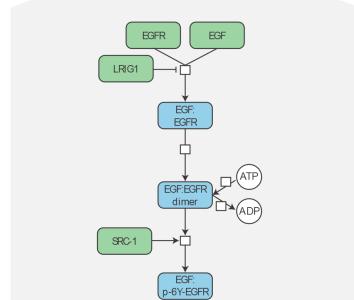
- Easier to interpret
 - Familiar concepts e.g. cell cycle
- Identifies possible causal mechanisms
- Predicts new roles for genes
- Improves statistical power
 - Fewer tests, aggregates data from multiple genes into one pathway
- More reproducible
 - E.g. gene expression signatures
- Facilitates integration of multiple data types

Module 1: Introduction to Pathway and Network Analysis

bioinformatics.ca

Pathways vs. Networks

EGFR-centered Pathway



- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

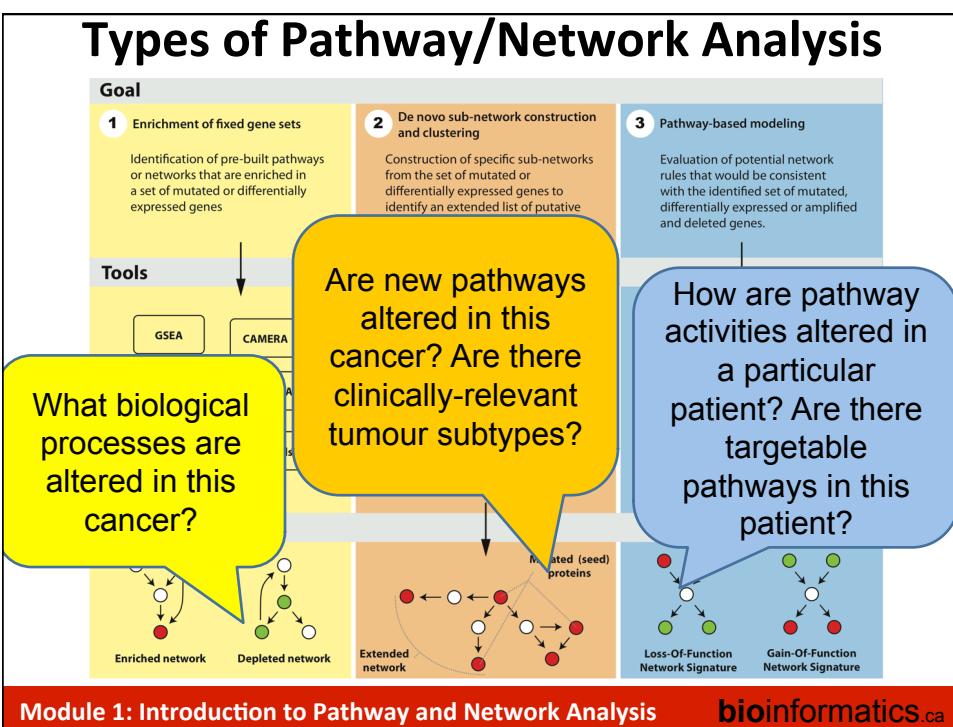
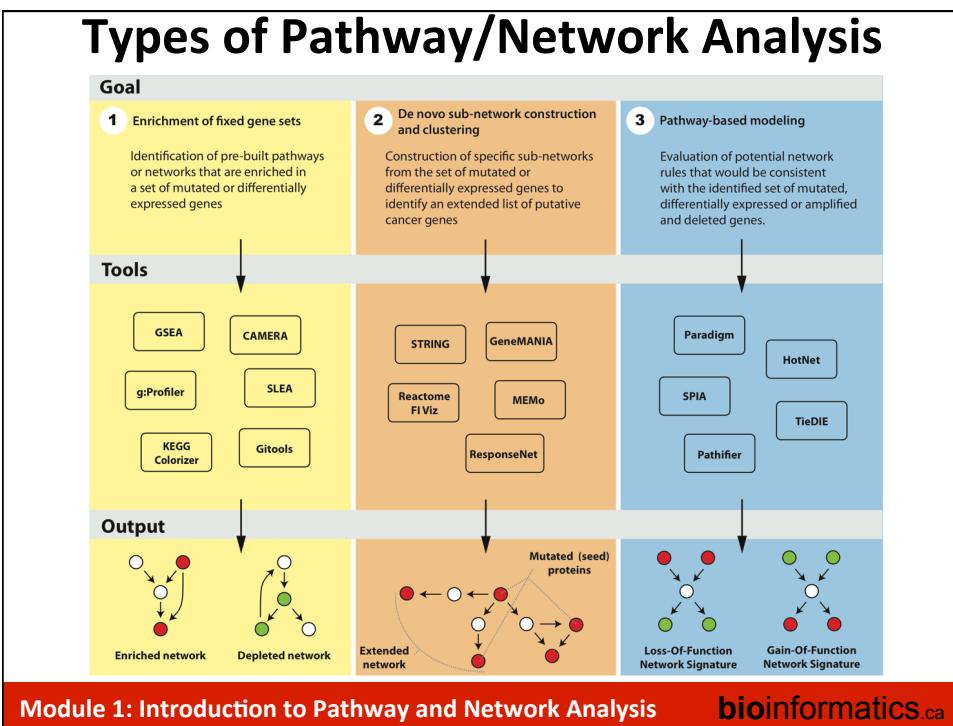
EGFR-centered Network



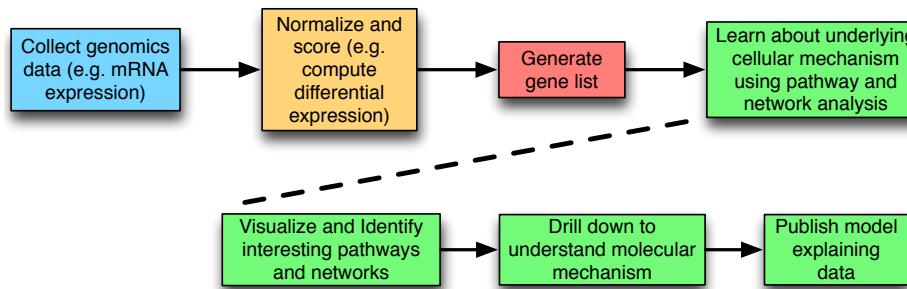
- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration

Module 1: Introduction to Pathway and Network Analysis

bioinformatics.ca

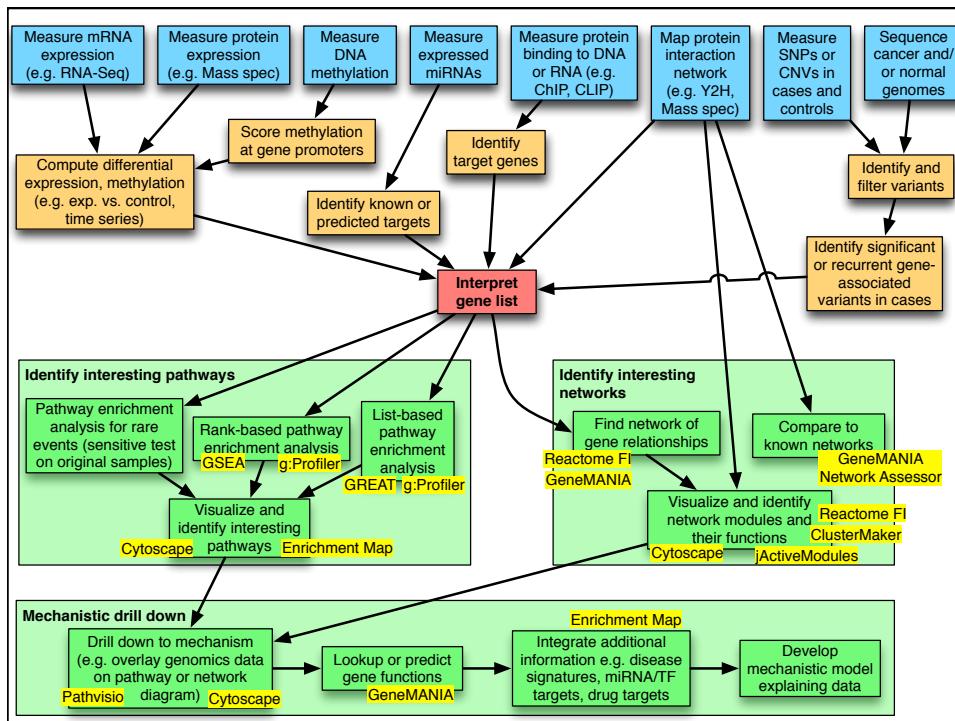


Pathway analysis workflow overview



Module 1: Introduction to Pathway and Network Analysis

bioinformatics.ca



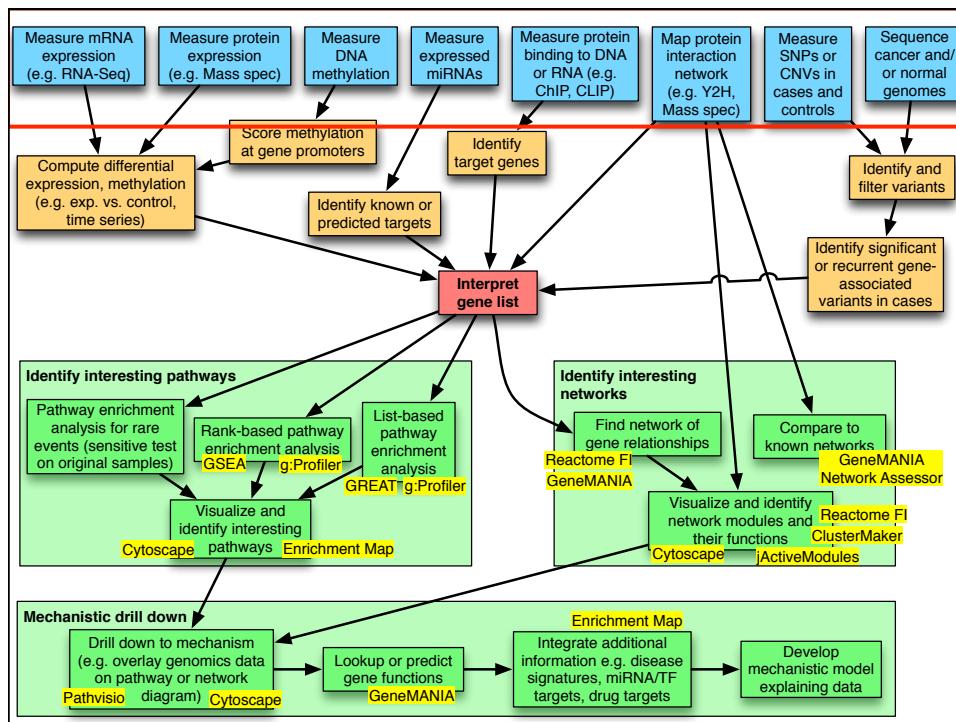
Where Do Gene Lists Come From?

- Molecular profiling e.g. mRNA, protein
 - Identification → Gene list
 - Quantification → Gene list + values
 - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, microRNA targets, transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
 - Single nucleotide polymorphisms (SNPs)
 - Copy number variants (CNVs)

Other examples?

What Do Gene Lists Mean?

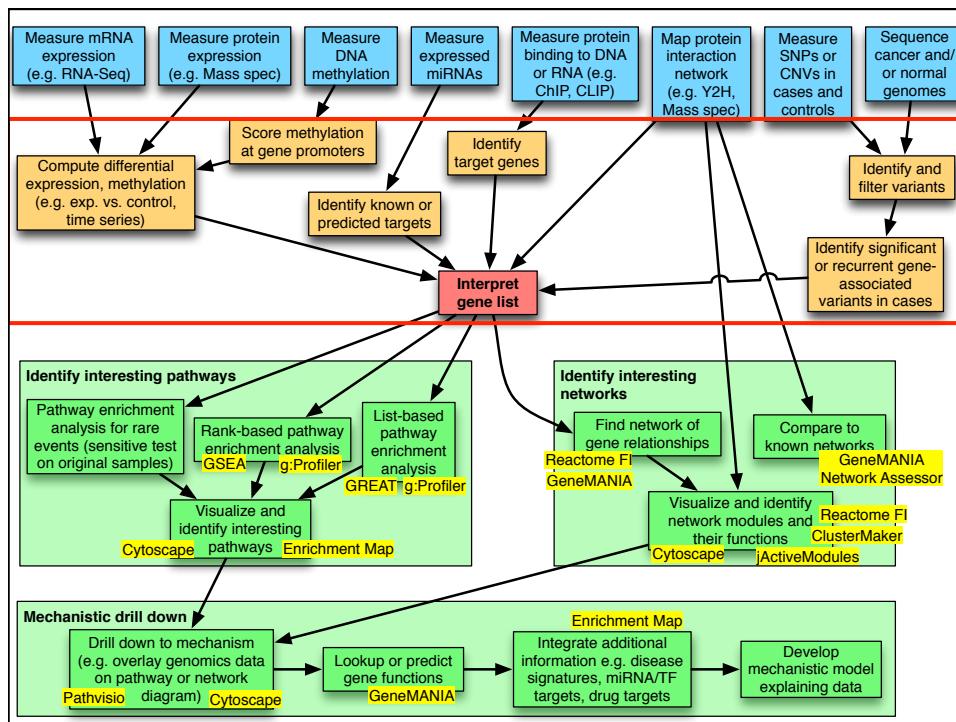
- Biological system: complex, pathway, physical interactors
- Similar gene function e.g. protein kinase
- Similar cell or tissue location
- Chromosomal location (linkage, CNVs)



Before Analysis

- ✓ Normalization
- ✓ Background adjustment
- ✓ Quality control (garbage in, garbage out)

- ✓ Use statistics that will increase signal and reduce noise specifically for your experiment
- ✓ Gene list size
- ✓ Make sure your gene IDs are compatible with software



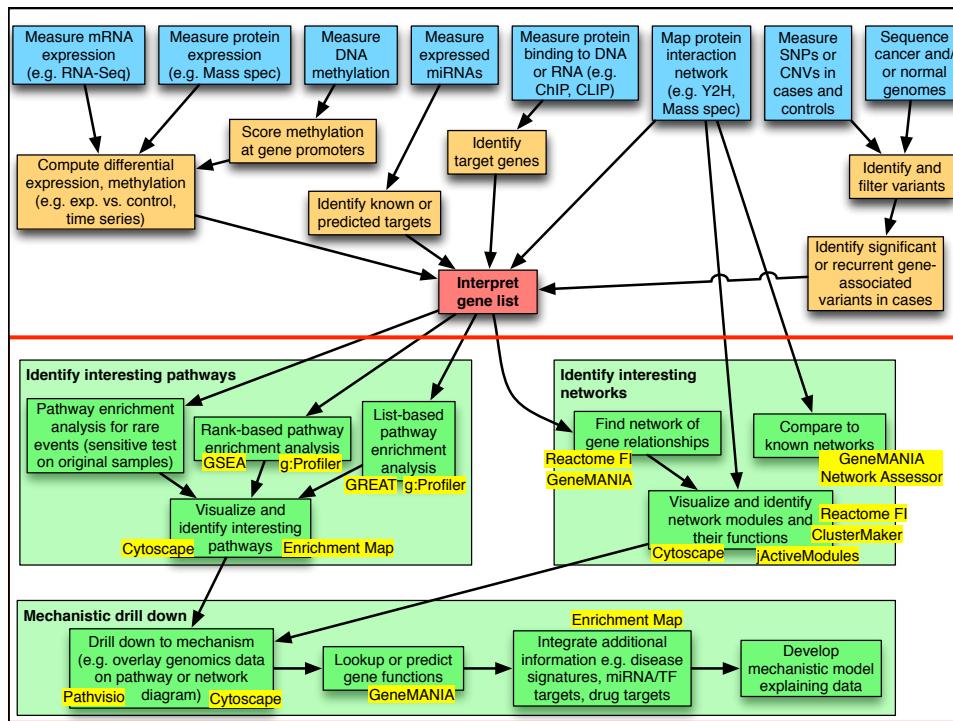
Biological Questions

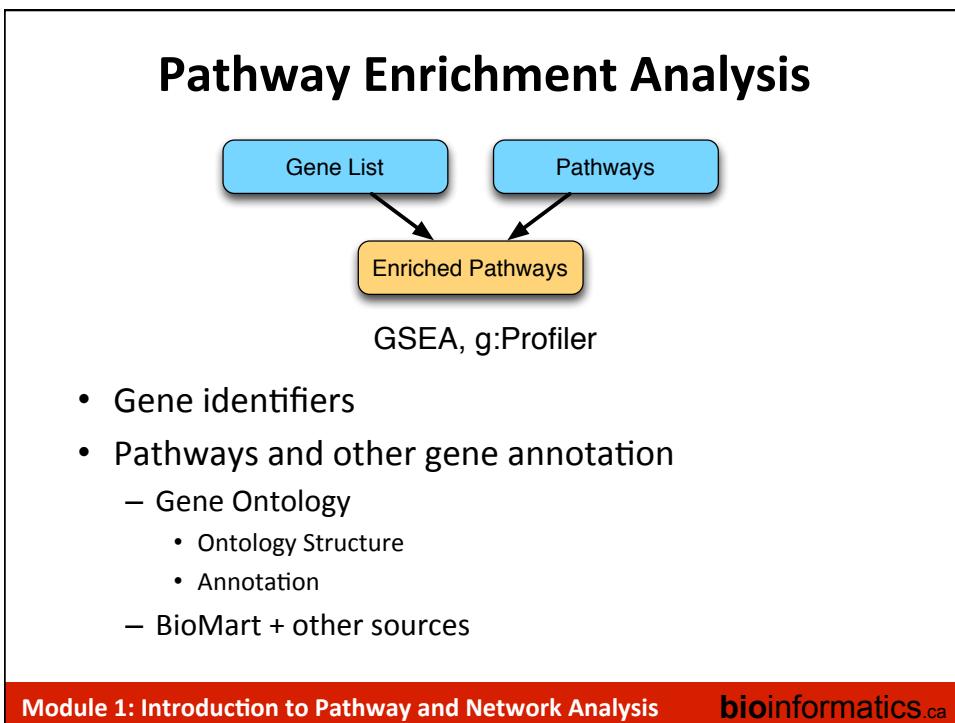
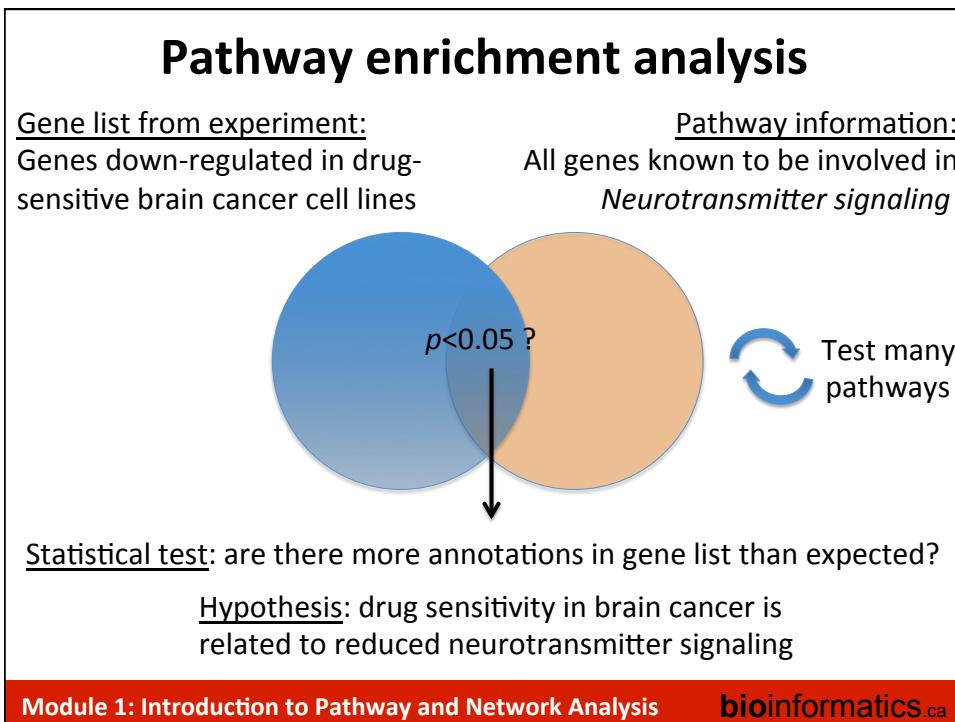
- Step 1: What do you want to accomplish with your list (hopefully part of experiment design! ☺)
 - Summarize biological processes or other aspects of gene function
 - Perform differential analysis – what pathways are different between samples?
 - Find a controller for a process (TF, miRNA)
 - Find new pathways or new pathway members
 - Discover new gene function
 - Correlate with a disease or phenotype (candidate gene prioritization)
 - Find a drug

Biological Answers

- Computational analysis methods we will cover
 - Day 1: Pathway enrichment analysis: summarize and compare
 - Day 2: Network analysis: predict gene function, find new pathway members, identify functional modules (new pathways)
 - Day 3: Regulatory network analysis: find and analyze controllers

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca





Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
 - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
 - Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
 - Important to recognize the correct record type
 - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins e.g. in RefSeq, which stores sequence.

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

Common Identifiers

Gene

[Ensembl](#) ENSG00000139618
[Entrez Gene](#) 675
 Unigene Hs.34012

Species-specific

HUGO HGNC BRCA2
 MGI MGI:109337
 RGD 2219
 ZFIN ZDB-GENE-060510-3
 FlyBase CG9097
 WormBase WBGene00002299 or ZK1067.1
 SGD S000002187 or YDL029W

Annotations

InterPro IPR015252
 OMIM 600185
 Pfam PF09104
 Gene Ontology GO:0000724
 SNPs rs28897757

Experimental Platform

Affymetrix 208368_3p_s_at
 Agilent A_23_P99452
 CodeLink GE60169
 Illumina GI_4502450-S

Red = Recommended

Protein

Ensembl ENSP00000369497
[RefSeq](#) NP_000050.2
[UniProt](#) BRCA2_HUMAN or
 A1YBP1_HUMAN
 IPI IPI00412408.1
 EMBL AF309413
 PDB 1MIU

Identifier Mapping

- So many IDs!
 - Software tools recognize only a handful
 - May need to map from your gene list IDs to standard IDs
- Four main uses
 - Searching for a favorite gene name
 - Link to related resources
 - Identifier translation
 - E.g. Proteins to genes, Affy ID to Entrez Gene
 - Merging data from different sources
 - Find equivalent records

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

ID Mapping Services

Input gene/protein/transcript IDs (mixed)

TP53 MDM2 207105_S_AT P60484

Type of output ID

>> Static URL Come back later					
g#	initial alias >> g:Goot >> g:Sorter >> g:Orth >> g:Cocoa	converted alias >> g:GOST >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	name >> g:GOST >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	description >> description	namespace >> namespace
1	TP53	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]	UNIPROT_GN, ENTREZGENE, VEGA_GENE, DBRASSS, DBASS3, HGNC, WIKIGENE	
2	MDM2	MDM2	MDM2 proto-oncogene, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:6973]	UNIPROT_GN, ENTREZGENE, VEGA_GENE, HGNC, WIKIGENE	
3	207105_S_AT	PIK3R2	phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:9880]	AFFY_HG_U133A_PLUS_2, AFFY_HG_FOCUS, AFFY_HG_U133A_2, AFFY_HG_U133A	
4	P60484	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9388]	UNIPROTSWISSPROT	

- **g:Convert**
- <http://biit.cs.ut.ee/gprofiler/gconvert.cgi>
- **Ensembl Biomart**
- <http://www.ensembl.org>

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Beware of ambiguous ID mappings

The screenshot shows the gProfiler web interface. At the top, there's a navigation bar with links like 'Welcome!', 'About', 'Contact', 'Beta', 'Archives', and 'Logout'. Below the navigation is a search bar with dropdown menus for 'Organism' (set to 'Human'), 'Query' (set to 'genes, proteins, probes, term') with a value 'TP53 MDM2 207105_S_AT p60484', and 'Options' (checkboxes for 'Significant only', 'Order by query', 'Chromosomal regions', 'Hierarchical sorting', and 'Show all terms (no filtering)'). There's also a 'Submit' button and a 'Clear' button.

On the right side, there's a large list of biological concepts with their corresponding GO IDs and descriptions. Some items are highlighted in blue, such as 'Gene Ontology', 'Biological process', 'Cellular component', 'Molecular function', 'Protein from experiment [EXP]', 'Direct acting [DAO]', 'Mutant phenotype [MP]', 'Genetic interaction [GIG]', 'Physical interaction [PI]', 'Traceable author [TAS]', 'Non-traceable author [NAS]', 'Inferred by curator [IC]', 'Expression pattern [IEP]', 'Sequence or structural similarity [ISS]', 'Genomic context [IGC]', 'Biological aspect of ancestor [BA]', 'Rapid divergence [RD]', 'Reviewed computational analysis [RCA]', 'Electronic annotation [EIA]', 'No biological data [ND]', 'Not annotated [NA]', 'Biological pathways', 'KEGG', 'Reactome', 'Regulatory motifs in DNA', 'TRANSFAC TFBS', 'miRNA', 'CORUM protein complexes', 'Human Phenotype Ontology', and 'BioGRID protein-protein interaction'.

A yellow box at the bottom left contains a warning message: 'Warning: Some gene identifiers are ambiguous. Resolve these manually.' It also says 'Attempt to automatically resolve symbols using a namespace (percentage of ambiguous symbols resolved in brackets):' followed by '207105_S_AT' with two entries: 'ENSG00000268173 (PIK3R2, 26 GO annot.) - phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]' and 'ENSG00000105647 (PIK3R2, 26 GO annot.) - phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]'. There's a 'Submit query' button at the bottom.

At the bottom of the page, there's a red footer bar with the text 'Module 1: Introduction to Pathway and Network Analysis' and 'bioinformatics.ca'.

ID Challenges

- Avoid errors: map IDs correctly
 - Beware of 1-to-many mappings
- Gene name ambiguity – not a good ID
 - e.g. FLJ92943, LFS1, TRP53, p53
 - Better to use the standard gene symbol: TP53
- Excel error-introduction
 - OCT4 is changed to October-4 (paste as text)
- Problems reaching 100% coverage
 - E.g. due to version issues
 - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

At the bottom of the page, there's a red footer bar with the text 'Module 1: Introduction to Pathway and Network Analysis' and 'bioinformatics.ca'.

Letters to Nature

Nature **426**, 100 (6 November 2003) | doi:10.1038/nature02141

Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

Hiroaki Kawasaki & Kazunari Taira

Nature **423**, 838–842 (2003).

In this Article, the messenger RNA that is identified to be a target of microRNA-23 (miR-23) is from the gene termed human 'homolog of ES1' (HES1), accession number Y07572, and not from the gene encoding the transcriptional repressor 'Hairy enhancer of split' HES1 (accession number NM_00524) as stated in our paper. We incorrectly identified the gene because of the confusing nomenclature. The function of HES1 Y07572 is unknown but the encoded protein shares homology with a protein involved in isoprenoid biosynthesis. Our experiments in NT2 cells had revealed that the protein levels of the repressor Hes1 were diminished by miR-23. Although we have unpublished data that suggest the possibility that miR-23 might also interact with Hes1 repressor mRNA, the explanation for the finding that the level of repressor Hes1 protein decreases in response to miR-23 remains undefined with respect to mechanism and specificity. Given the interpretational difficulties resulting from our error, we respectfully retract the present paper. Further studies aimed at clarifying the physiological role of miR-23 will be submitted to a peer-reviewed journal subject to the outcome of our ongoing research.

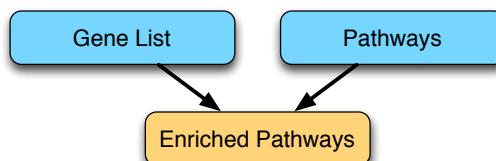
Recommendations

- For proteins and genes
 - (doesn't consider splice forms)
- Map everything to Entrez Gene IDs or Official Gene Symbols using a spreadsheet
- If 100% coverage desired, manually curate missing mappings using multiple resources
- Be careful of Excel auto conversions – especially when pasting large gene lists!
 - Remember to format cells as 'text' before pasting

What Have We Learned?

- Genes and their products and attributes have many identifiers (IDs)
- Genomics often requires conversion of IDs from one type to another
- ID mapping services are available
- Use standard, commonly used IDs to reduce ID mapping challenges

Pathway Enrichment Analysis



GSEA, g:Profiler

- Gene identifiers
- Pathways and other gene annotation
 - Gene Ontology
 - Ontology Structure
 - Annotation
 - BioMart + other sources

Pathways and other gene function attributes

- Available in databases
- Pathways
 - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
 - Gene Ontology molecular function, cell location
 - Chromosome position
 - Disease association
 - DNA properties
 - TF binding sites, gene structure (intron/exon), SNPs
 - Transcript properties
 - Splicing, 3' UTR, microRNA binding sites
 - Protein properties
 - Domains, secondary and tertiary structure, PTM sites
 - Interactions with other genes

Pathways and other gene function attributes

- Available in databases
- Pathways
 - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
 - Gene Ontology molecular function, cell location
 - Chromosome position
 - Disease association
 - DNA properties
 - TF binding sites, gene structure (intron/exon), SNPs
 - Transcript properties
 - Splicing, 3' UTR, microRNA binding sites
 - Protein properties
 - Domains, secondary and tertiary structure, PTM sites
 - Interactions with other genes

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
 - protein kinase
 - apoptosis
 - membrane
- Dictionary: term definitions
- Ontology: A formal system for describing knowledge
- www.geneontology.org

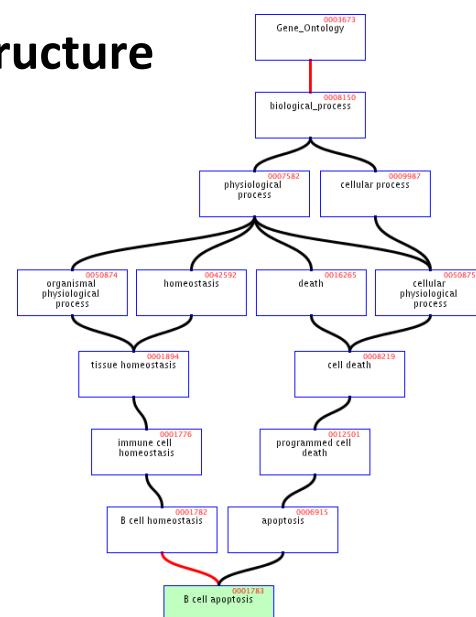


Jane Lomax @ EBI

Module 1: Introduction to Pathway and Network Analysis www.geneontology.org

- Terms are related within a hierarchy
 - is-a
 - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child

GO Structure

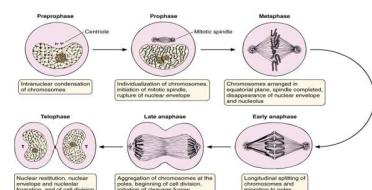
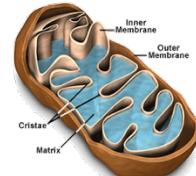


Module 1: Introduction to Pathway and Network Analysis

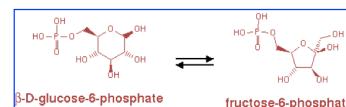
bioinformatics.ca

What GO Covers?

- GO terms divided into three aspects:
 - cellular component
 - molecular function
 - biological process



Cell division



glucose-6-phosphate
isomerase activity

Part 1/2: Terms

- Where do GO terms come from?
 - GO terms are added by editors at EBI and gene annotation database groups
 - Terms added by request
 - Experts help with major development

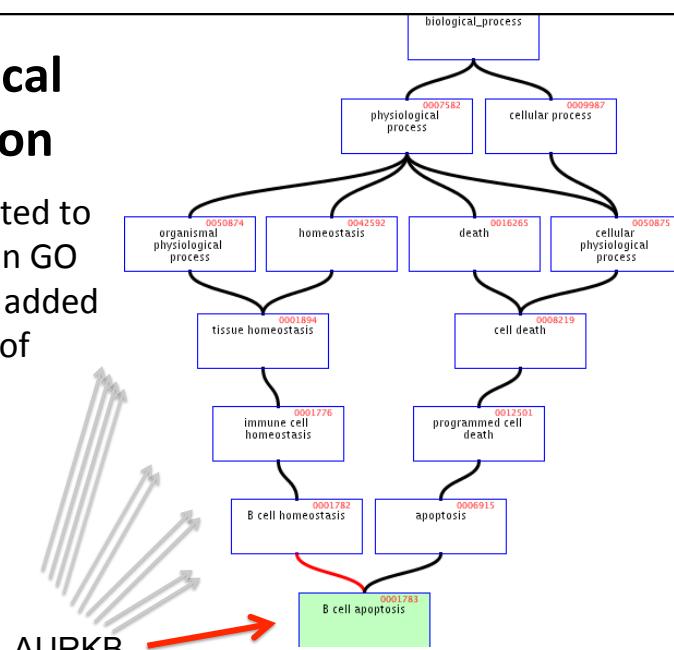
	Jun 2012	Jun 2016	increase
Biological process	23,074	29,541	28%
Molecular function	9,392	11,133	19%
Cellular component	2,994	4,082	36%
total	37,104	44,756	21%

Part 2/2: Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
 - Known as ‘gene associations’ or GO annotations
 - Multiple annotations per gene
- Some GO annotations created automatically (without human review)

Hierarchical annotation

- Genes annotated to specific term in GO automatically added to all parents of that term



Annotation Sources

- Manual annotation
 - Curated by scientists
 - High quality
 - Small number (time-consuming to create)
 - Reviewed computational analysis
- Electronic annotation
 - Annotation derived without human validation
 - Computational predictions (accuracy varies)
 - Lower ‘quality’ than manual codes
- Key point: be aware of annotation origin

For your information

Evidence Types

- | | |
|--|---|
| <ul style="list-style-type: none"> • Experimental Evidence Codes <ul style="list-style-type: none"> • EXP: Inferred from Experiment • IDA: Inferred from Direct Assay • IPI: Inferred from Physical Interaction • IMP: Inferred from Mutant Phenotype • IGI: Inferred from Genetic Interaction • IEP: Inferred from Expression Pattern  | <ul style="list-style-type: none"> • Author Statement Evidence Codes <ul style="list-style-type: none"> • TAS: Traceable Author Statement • NAS: Non-traceable Author Statement • Curator Statement Evidence Codes <ul style="list-style-type: none"> • IC: Inferred by Curator • ND: No biological Data available  |
| <ul style="list-style-type: none"> • Computational Analysis Evidence Codes <ul style="list-style-type: none"> • ISS: Inferred from Sequence or Structural Similarity • ISO: Inferred from Sequence Orthology • ISA: Inferred from Sequence Alignment • ISM: Inferred from Sequence Model • IGC: Inferred from Genomic Context • RCA: inferred from Reviewed Computational Analysis | |
| <ul style="list-style-type: none"> • IEA: Inferred from electronic annotation | |

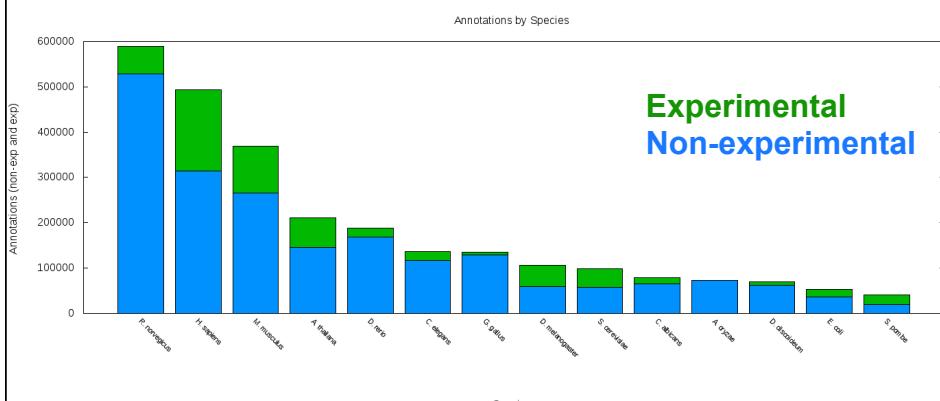
<http://www.geneontology.org/GO.evidence.shtml>

Species Coverage

- All major eukaryotic model organism species and human
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development
- Current list:
 - <http://geneontology.org/page/download-annotations>

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Variable Coverage



www.geneontology.org, Jun 2015

For your information

Contributing Databases

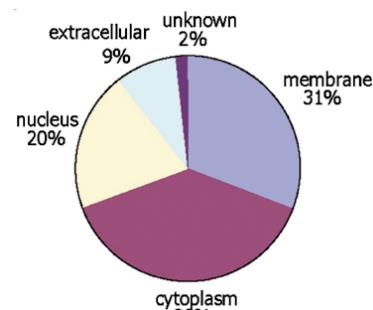
- [Berkeley *Drosophila* Genome Project \(BDGP\)](#)
- [dictyBase](#) (*Dictyostelium discoideum*)
- [FlyBase](#) (*Drosophila melanogaster*)
- [GeneDB](#) (*Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei*)
- [UniProt Knowledgebase](#) (Swiss-Prot/TrEMBL/PIR-PSD) and [InterPro](#) databases
- [Gramene](#) (grains, including rice, *Oryza*)
- [Mouse Genome Database \(MGD\)](#) and [Gene Expression Database \(GXD\)](#) (*Mus musculus*)
- Rat Genome Database (RGD) (*Rattus norvegicus*)
- [Reactome](#)
- [Saccharomyces Genome Database \(SGD\)](#) (*Saccharomyces cerevisiae*)
- [The Arabidopsis Information Resource \(TAIR\)](#) (*Arabidopsis thaliana*)
- [The Institute for Genomic Research \(TIGR\)](#): databases on several bacterial species
- [WormBase](#) (*Caenorhabditis elegans*)
- [Zebrafish Information Network \(ZFIN\)](#): (*Danio rerio*)

Module 1: Introduction to Pathway and Network Analysis

bioinformatics.ca

GO Slim Sets

- GO has too many terms for some uses
 - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
 - Generic, plant, yeast



Crockett DK et al. Lab Invest. 2005 Nov;
85(11):1405-15

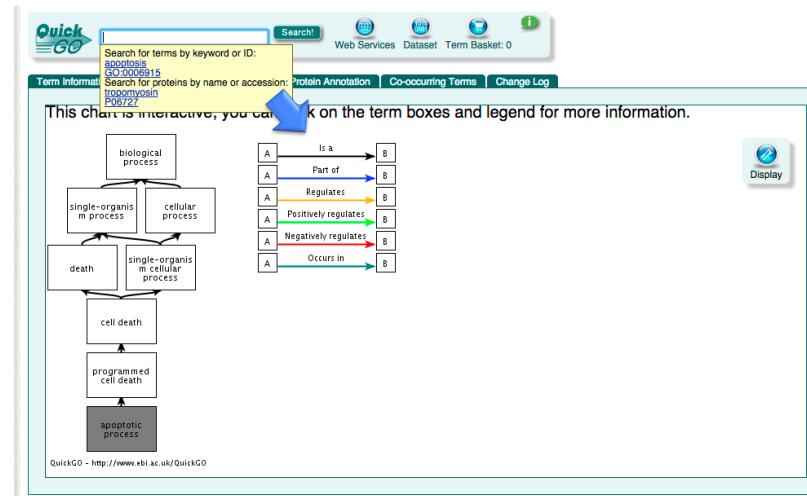
Module 1: Introduction to Pathway and Network Analysis

bioinformatics.ca

GO Software Tools

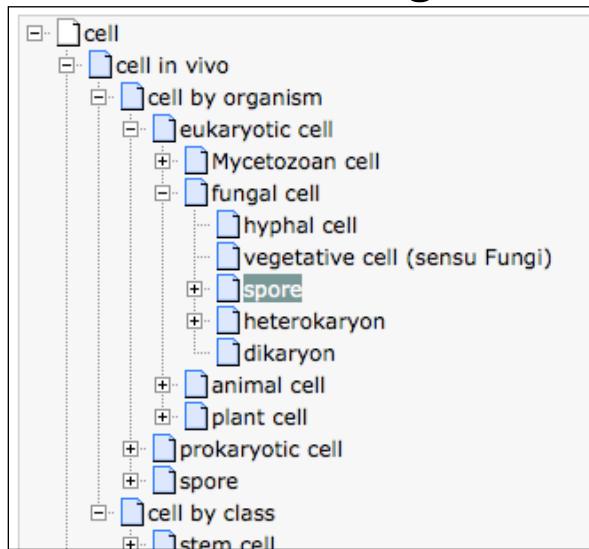
- GO resources are freely available to anyone without restriction
 - ontologies, gene associations and tools developed by GO
- Other groups have used GO to create versatile tools

Accessing GO: QuickGO



<http://www.ebi.ac.uk/QuickGO/>

Other Ontologies



<http://www.ebi.ac.uk/ontology-lookup>

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Pathway Databases

- <http://www.pathguide.org/> lists ~550 pathway related databases
- MSigDB: <http://www.broadinstitute.org/gsea/msigdb/>
- <http://www.pathwaycommons.org/> collects major ones

Pathway Commons

Download FAQ Publications Contact

Pathway Commons

Search and visualize public biological pathway information. Single point of access.

BRCA1, BRCA2, MDM2 Start exploring »

Pathway Commons is a network biology resource and acts as a convenient point of access to biological pathway information collected from public pathway databases, which you can search, visualize and download. All data is freely available, under the license terms of each contributing database.

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca

Pathways and other gene function attributes

- Available in databases
- Pathways
 - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
 - Gene Ontology molecular function, cell location
 - Chromosome position
 - Disease association
 - DNA properties
 - TF binding sites, gene structure (intron/exon), SNPs
 - Transcript properties
 - Splicing, 3' UTR, microRNA binding sites
 - Protein properties
 - Domains, secondary and tertiary structure, PTM sites
 - Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (general)
 - <http://www.ensembl.org>
- Entrez Gene (general)
 - <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>
- Model organism databases
 - E.g. SGD: <http://www.yeastgenome.org/>
- Many others: discuss during lab

Ensembl BioMart

- Convenient access to gene list annotation

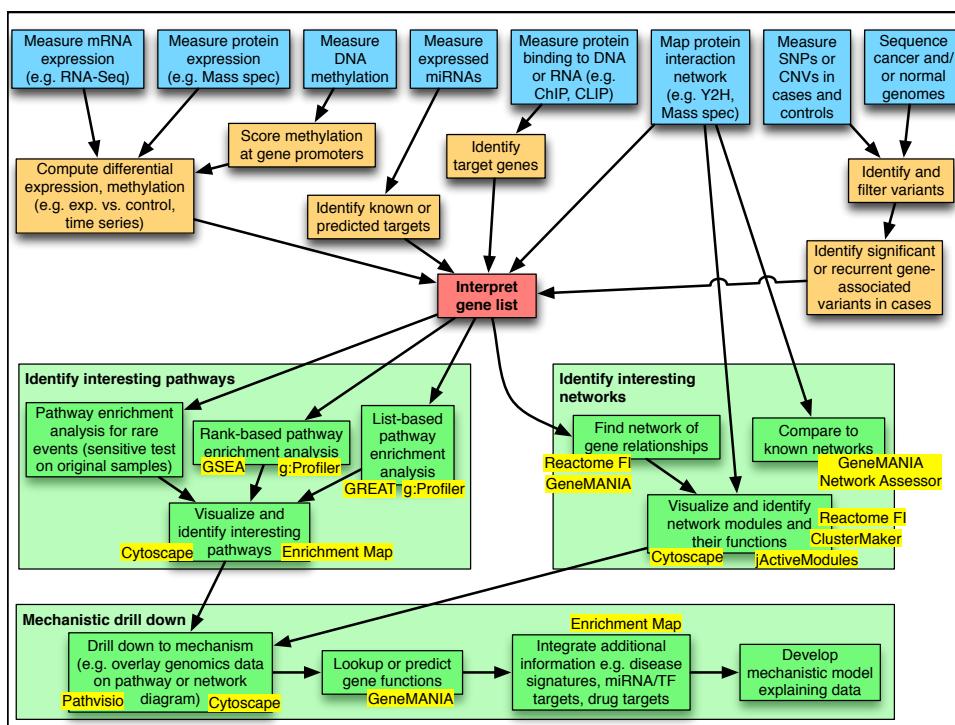
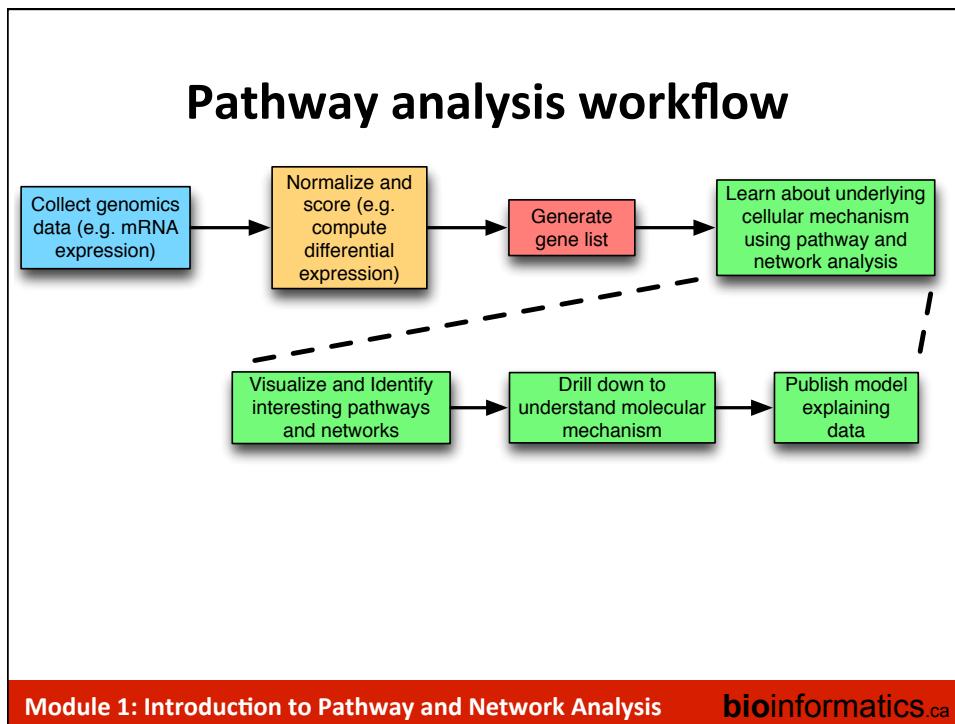
The screenshot shows the Ensembl BioMart interface. At the top left, there's a 'Dataset' dropdown set to 'Homo sapiens genes (GRCh37)' and a 'Filters' dropdown also set to 'Homo sapiens genes (GRCh37)'. Below these are sections for 'Attributes' (with options like 'Ensembl Gene ID' and 'Ensembl Transcript') and 'REGION' (checkboxes for GENE, TRANSCRIPT EVENT, GENE ONTOLOGY, EXPRESSION, MULTI SPECIES COMPARISONS, PROTEIN DOMAINS). There are also checkboxes for 'Limit to genes ...' (with a dropdown for 'Protein feature scanprosite ID(s)'), 'Limit to genes with these family or domain IDs:' (with a dropdown for 'Ensembl Protein Family ID(s)'), 'Transmembrane domains', 'Signal domains', and 'VARIATIONS'. To the right, three arrows point downwards from the text labels to specific parts of the interface: 'Select genome' points to the 'Dataset' dropdown; 'Select filters' points to the 'REGION' section; and 'Select attributes to download' points to the 'Attributes' section.

www.ensembl.org

What Have We Learned?

- Pathways and other gene attributes in databases
 - Pathways from Gene Ontology (GO) and pathway databases
 - Gene Ontology (GO)
 - GO is a classification system and dictionary for biological concepts
 - Annotations are contributed by many groups
 - More than one annotation term allowed per gene
 - Some genomes are annotated more than others
 - Annotation comes from manual and electronic sources
 - GO can be simplified for certain uses (GO Slim)
- Many gene attributes available from genome databases such as Ensembl

Module 1: Introduction to Pathway and Network Analysis bioinformatics.ca



Lab: Gene IDs and Attributes

- Objectives
 - Learn about gene identifiers, Synergizer and BioMart
- Use yeast demo gene list (module1YeastGenes.txt)
- Convert Gene IDs to Entrez Gene: Use g:Profiler
- Get GO annotation + evidence codes
 - Use Ensembl BioMart
 - Summarize terms & evidence codes in a table
- Do it again with your own gene list
 - If compatible with covered tools, run the analysis. If not, instructors will recommend tools for you.