bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

Module 1
Introduction to Pathway and Network Analysis of Gene Lists

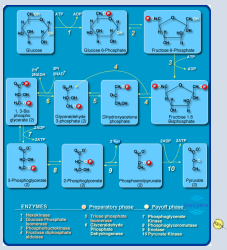
Gary Bader

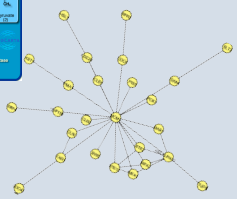
Pathway and Network Analysis of –omic Data

June 1-3, 2015

bioinformatics.ca

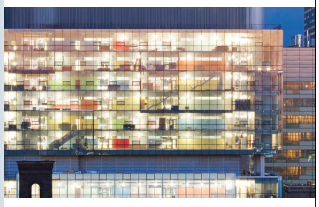
GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2





Donnelly Centre for Cellular + Biomolecular Research

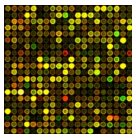
UNIVERSITY OF TORONTO



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

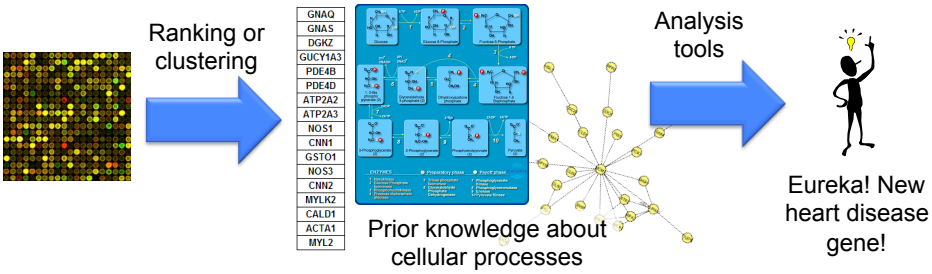
GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

?

GenMAPP.org

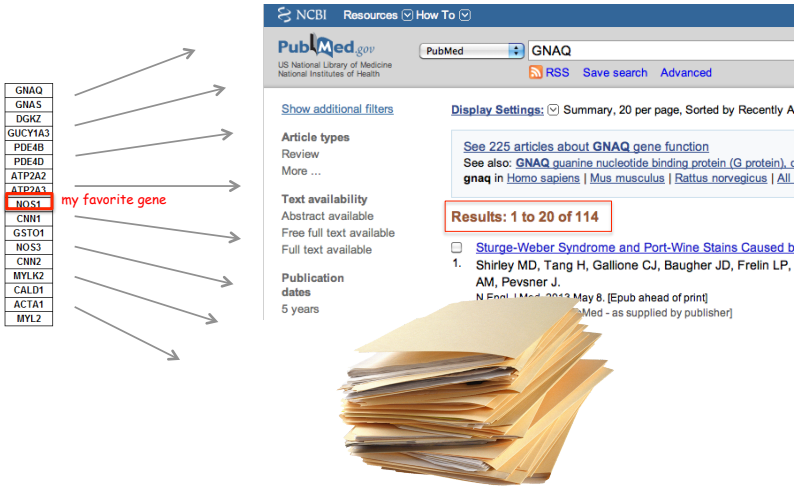
Interpreting Gene Lists

- 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



Pathway and network analysis

- Save time compared to traditional approach



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)

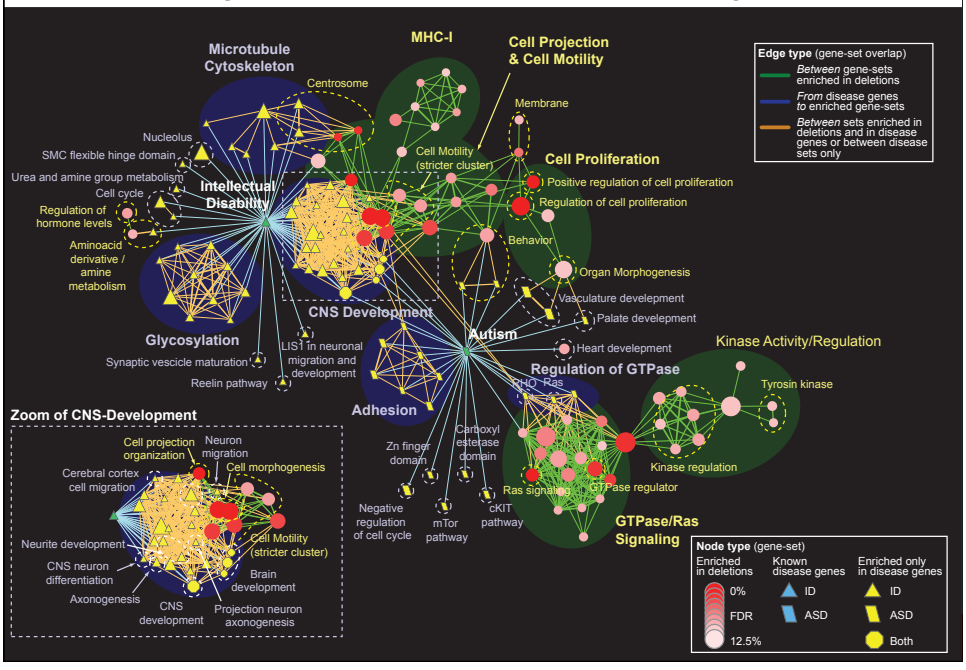
- Genetics
  - highly heritable
    - monozygotic twin concordance 60-90%
    - dizygotic twin concordance 0-10% (depending on the stringency of diagnosis)
  - known genetics:
    - 5-15% rare single-gene disorders and chromosomal re-arrangements
    - de-novo CNV previously reported in 5-10% of ASD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

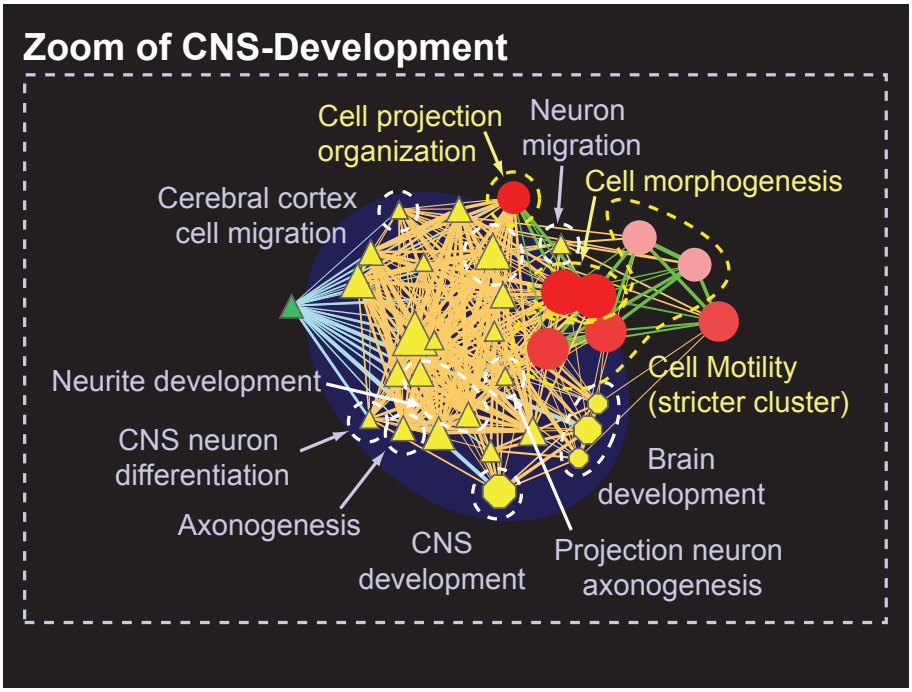
Pinto et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010 Jun 9.

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, iPattern, 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

Pathways Enriched in Autism Spectrum





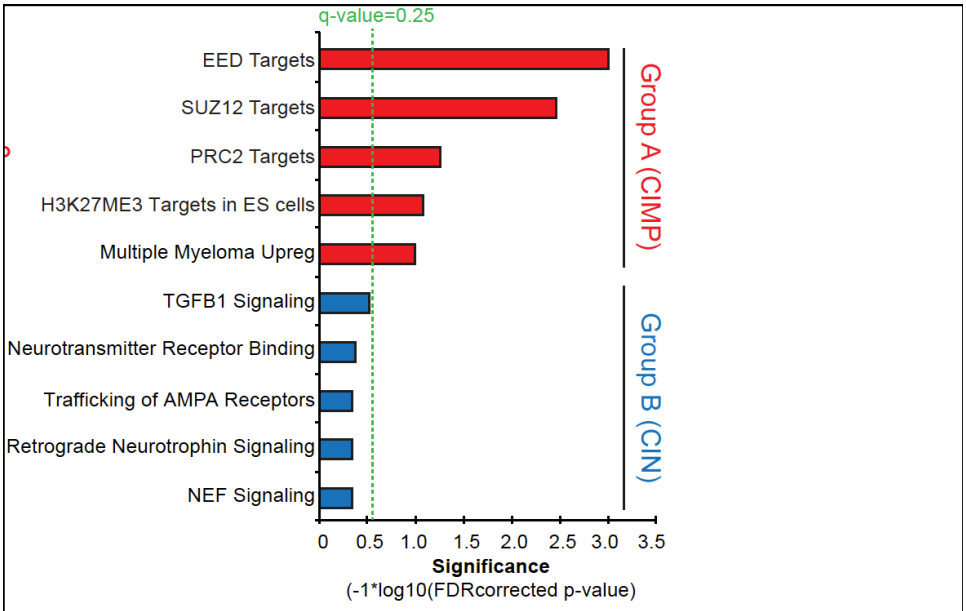
Pathway analysis example 2

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

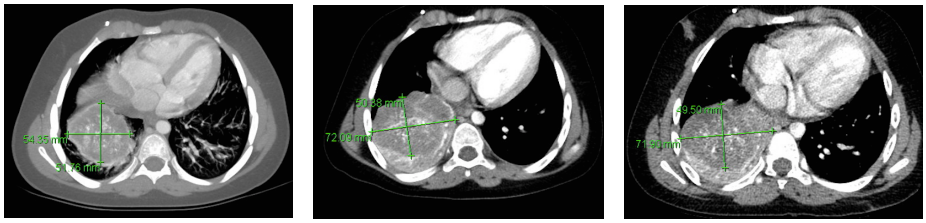
Witt et al., Cancer Cell 2011 **Nature. 2014 Feb 27;506(7489):445-50**

Steve Mack, Michael Taylor, Scott Zuyderduyn



polycomb repressor complex 2 – inhibited by SAHA, DZNep, GSK343 – killed PFA cells
No known treatment, so now going to clinical trial

9 yo with metastatic PF ependymoma to lung treated with azacytidine



2 months

3 months
3 cycles
Vidaza

Effect lasted 15 months

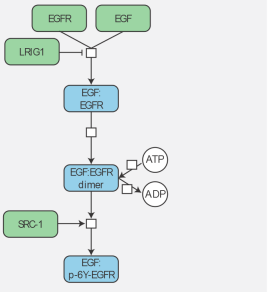
Benefits of Pathway Analysis

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

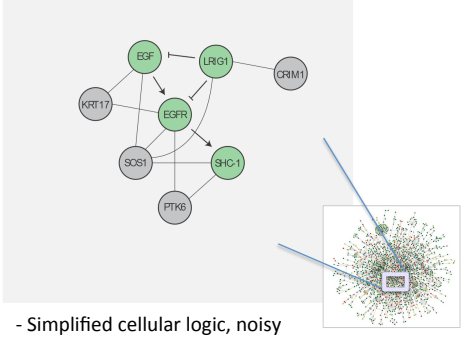
Pathways vs. Networks

EGFR-centered Pathway



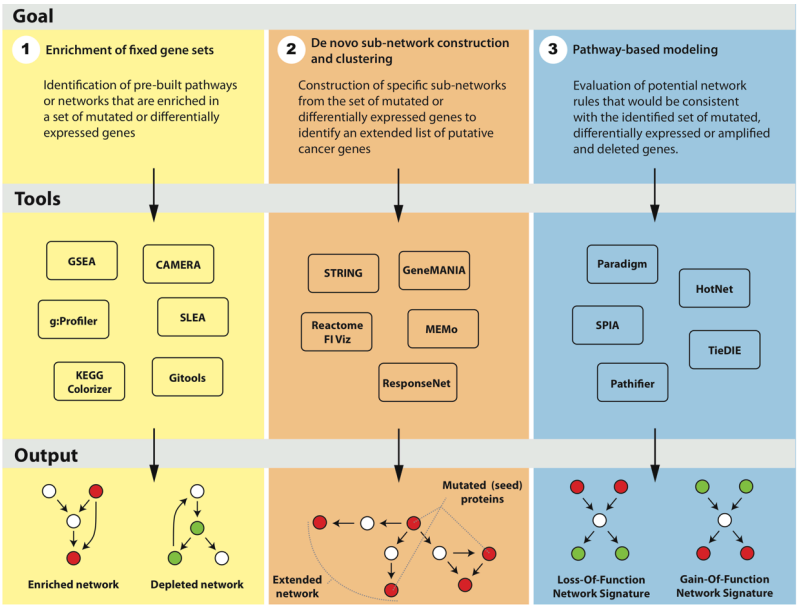
- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

EGFR-centered Network

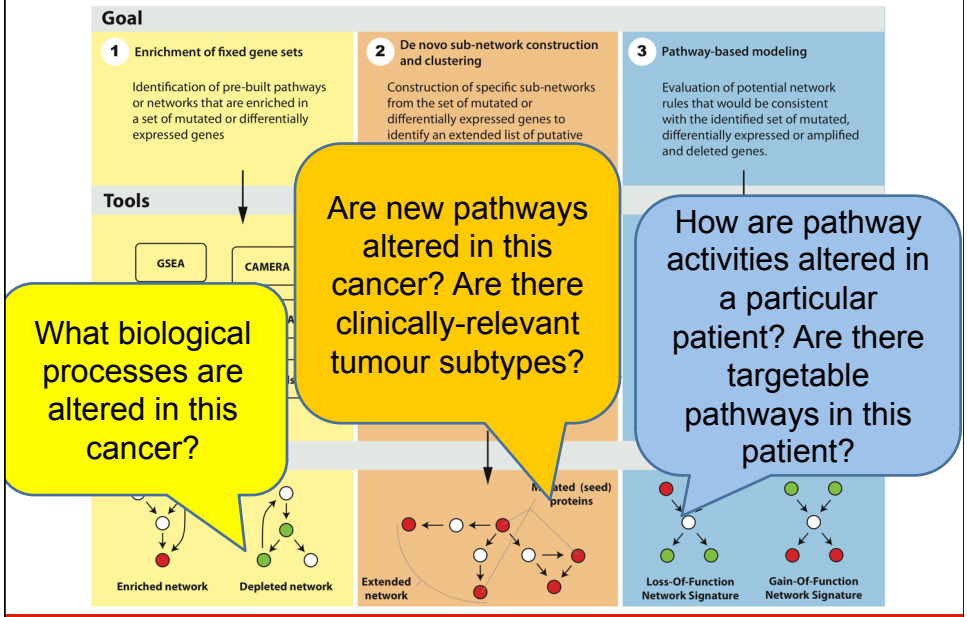


- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration

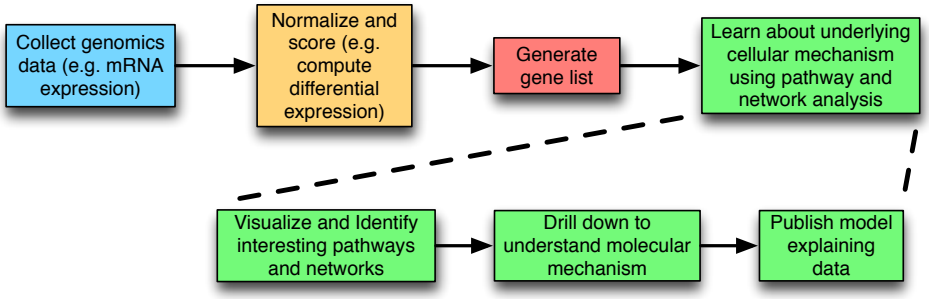
Types of Pathway/Network Analysis

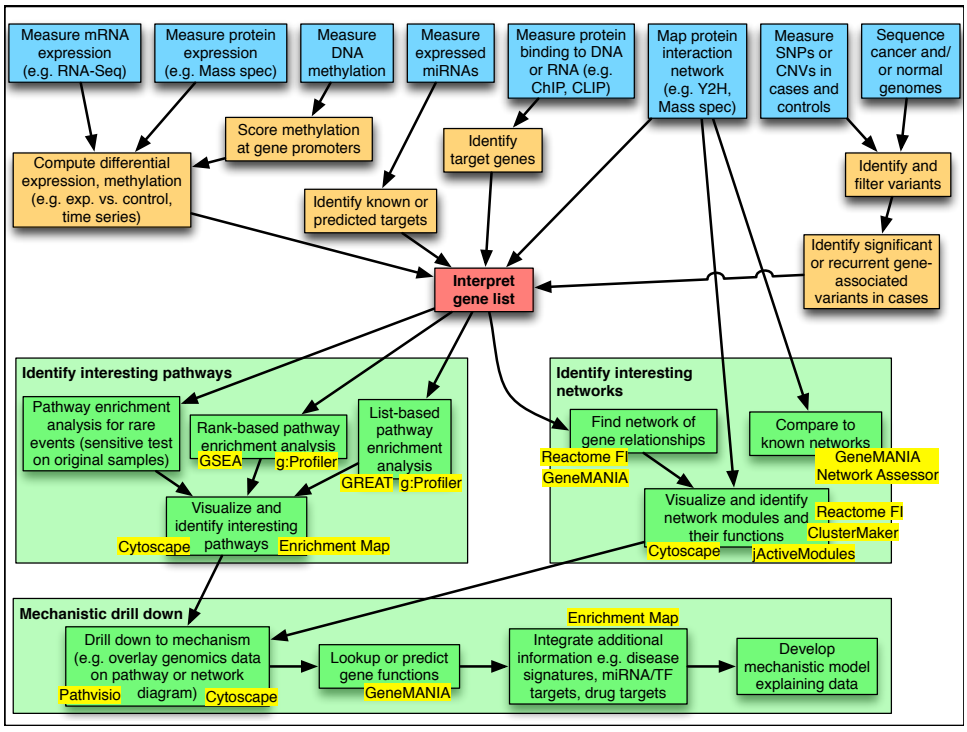


Types of Pathway/Network Analysis



Pathway analysis workflow overview

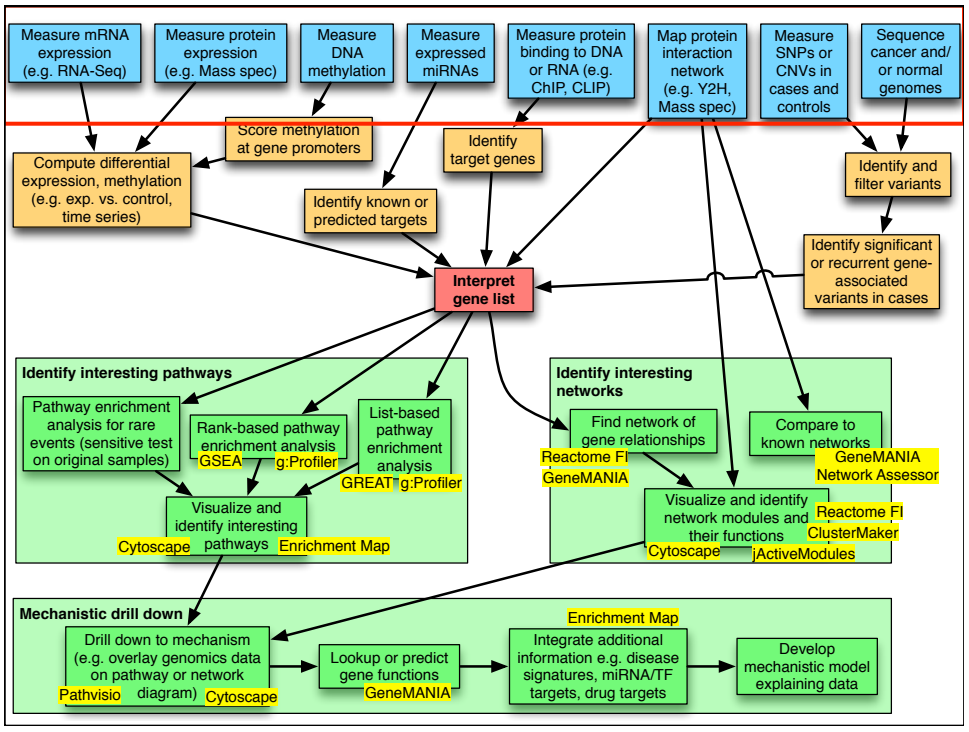




Where Do Gene Lists Come From?

- 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?



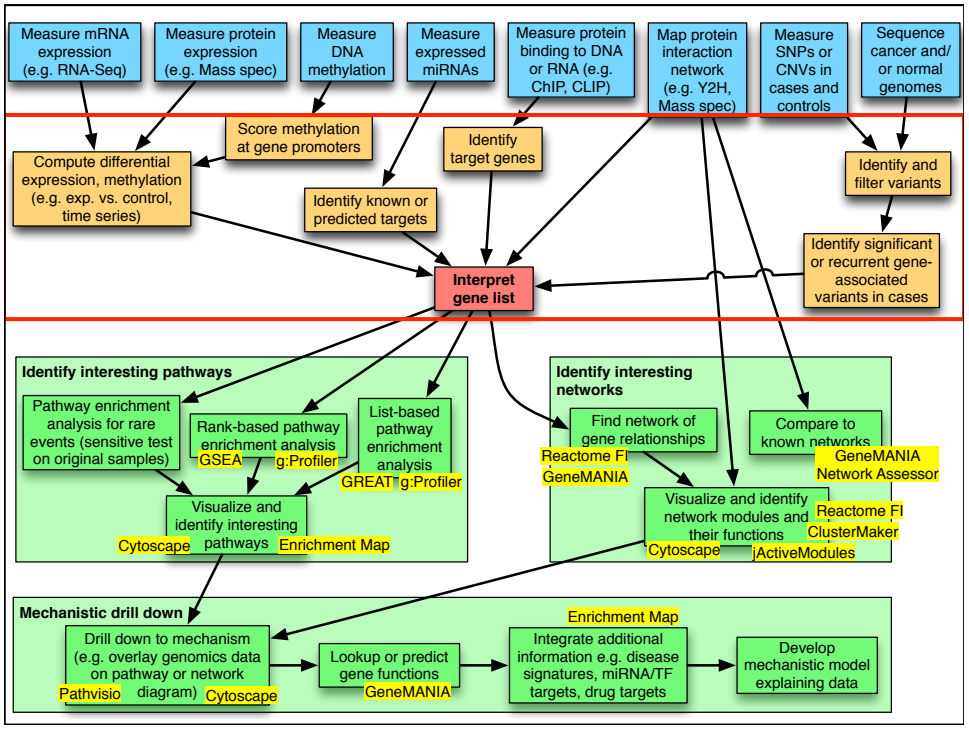
What Do Gene Lists Mean?

- 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

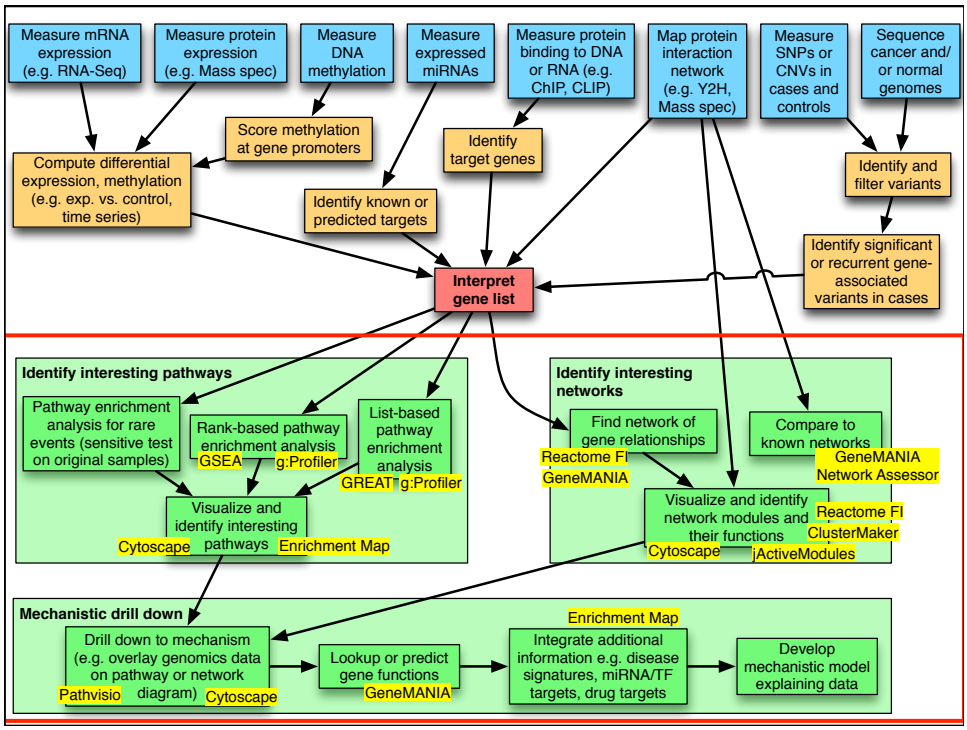


Biological Questions

- 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

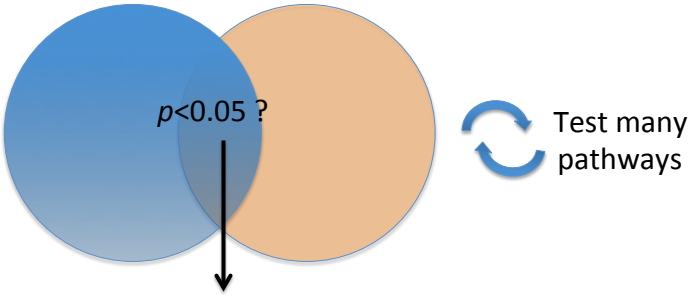
- Computational analysis methods we will cover
  - Day 1: Pathway enrichment analysis: summarize and compare
  - Day 2: Network analysis: predict gene function, find new pathway members, identify functional modules (new pathways)
  - Day 3: Regulatory network analysis: find and analyze controllers



Pathway enrichment analysis

Gene list from experiment:
Genes down-regulated in drug-sensitive brain cancer cell lines

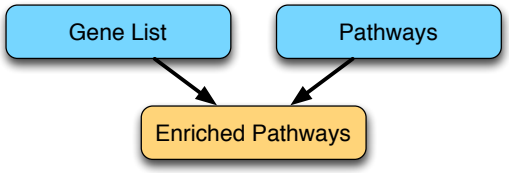
Pathway information:
All genes known to be involved in *Neurotransmitter signaling*



Statistical test: are there more annotations in gene list than expected?

Hypothesis: drug sensitivity in brain cancer is related to reduced neurotransmitter signaling

Pathway Enrichment Analysis

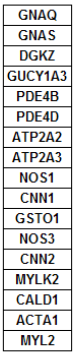


GSEA, g:Profiler

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

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.



Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

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

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

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

Letters to Nature

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

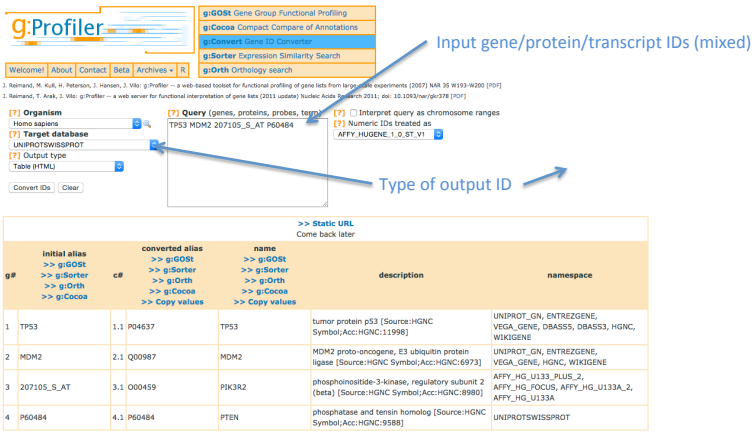
Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

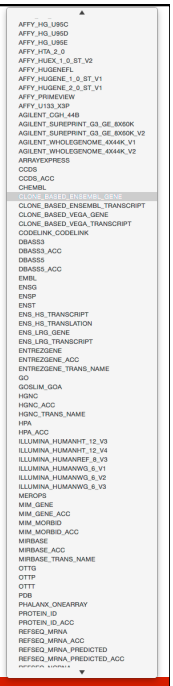
Hiroaki Kawasaki & Kazunari Taira

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.

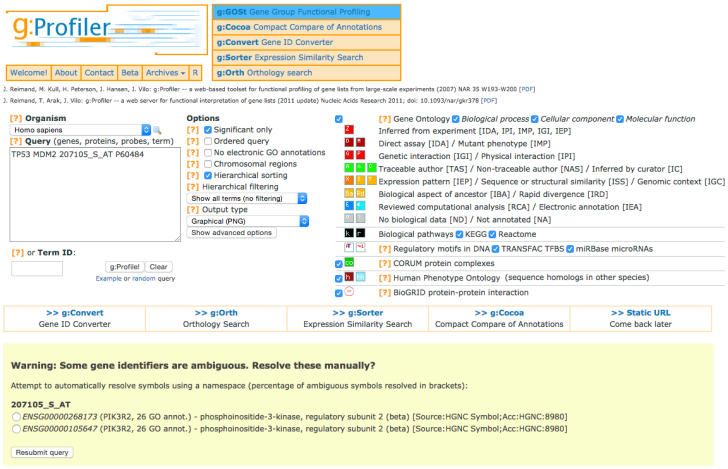
ID Mapping Services





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

Beware of ambiguous ID mappings



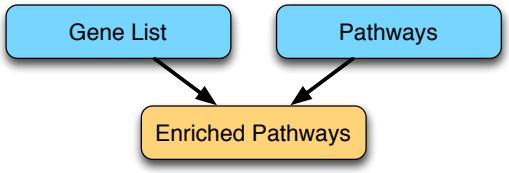
Recommendations

- 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

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

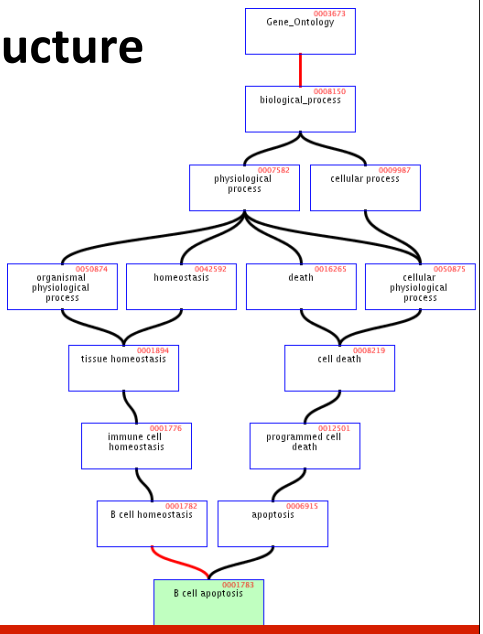
What is the Gene Ontology (GO)?

- 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

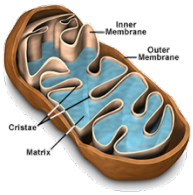
GO Structure

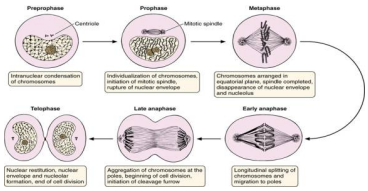
- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



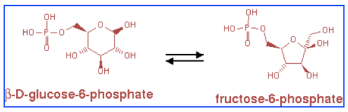
What GO Covers?

- 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

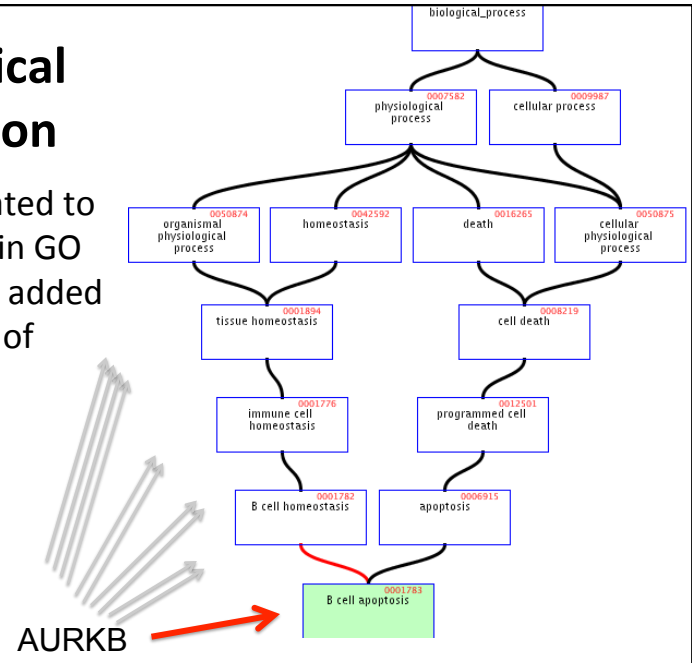
	Jun 2012	Apr 2015	increase
Biological process	23,074	28,158	22%
Molecular function	9,392	10,835	15%
Cellular component	2,994	3,903	30%
total	**37,104**	**42,896**	**16%**

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

- Experimental Evidence Codes
  - 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

- Computational Analysis Evidence Codes
  - 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

- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available

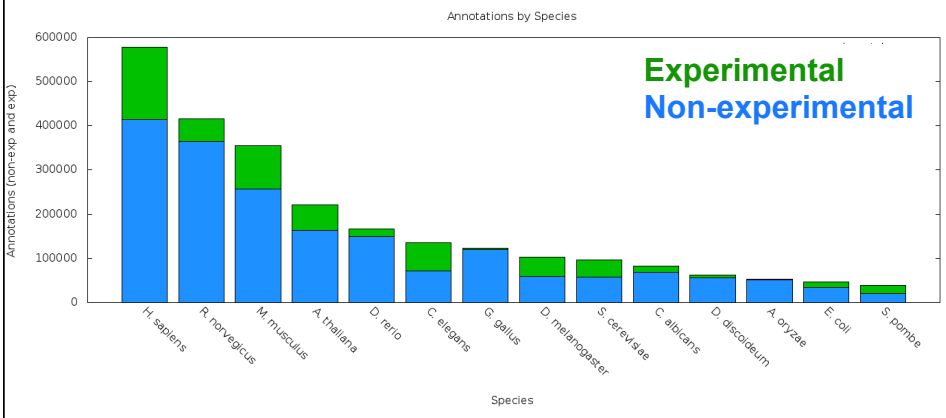
- **IEA:** Inferred from electronic annotation

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://www.geneontology.org/GO.downloads.annotations.shtml

Variable Coverage



www.geneontology.org, Apr 2015

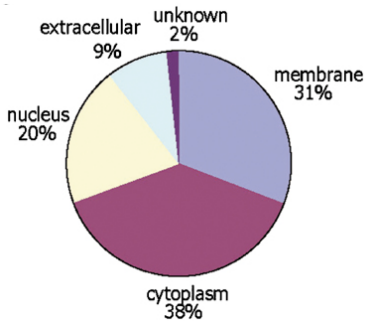
For your information

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)*

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

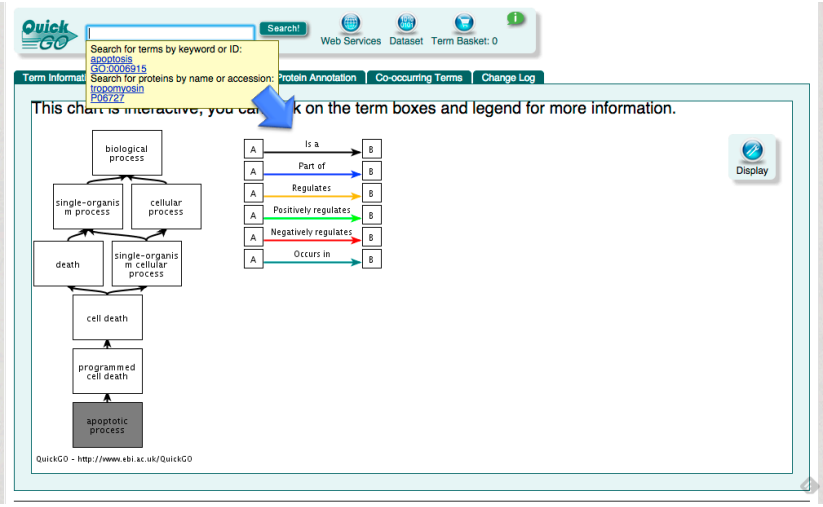


Crockett DK et al. Lab Invest. 2005 Nov; 85(11):1405-15

GO Software Tools

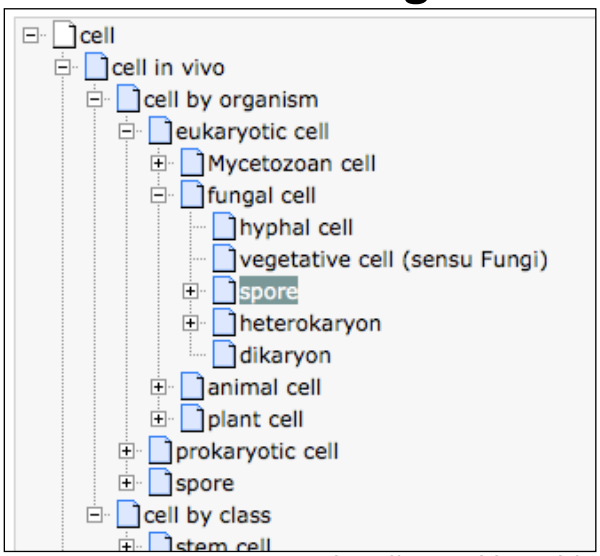
- GO resources are freely available to anyone without restriction
  - ontologies, gene associations and tools developed by GO
- Other groups have used GO to create versatile tools

Accessing GO: QuickGO



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

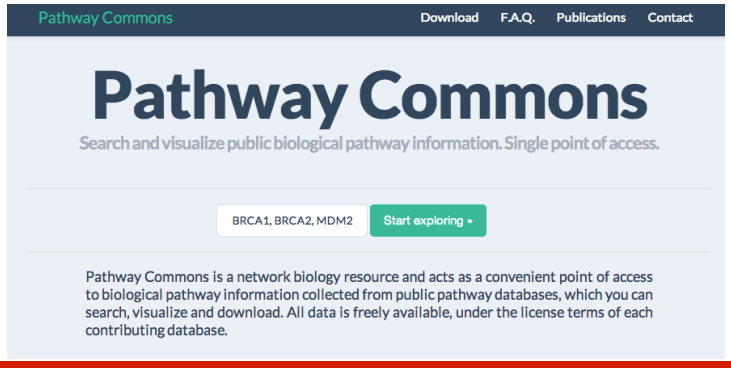
Other Ontologies



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

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



Pathways and other gene function attributes

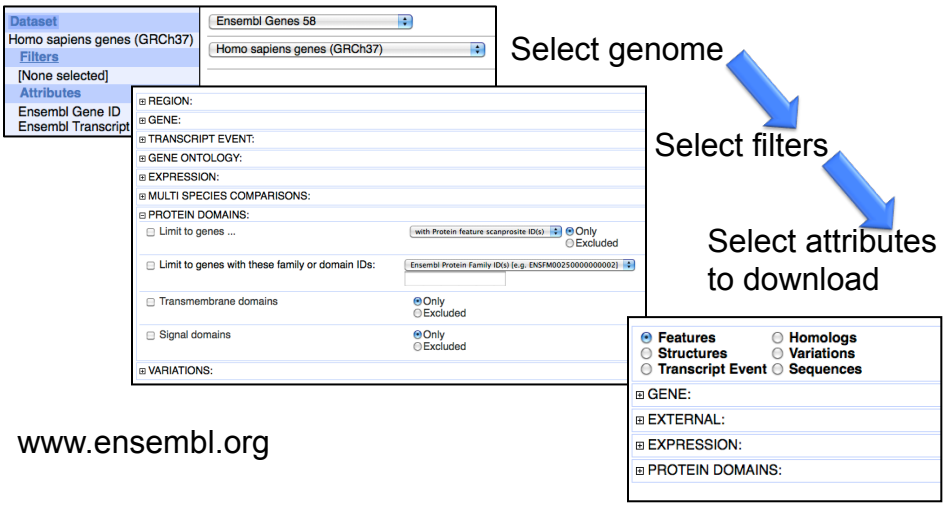
- Available in databases
- Pathways
  - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
  - Gene Ontology molecular function, cell location
  - Chromosome position
  - Disease association
  - DNA properties
    - TF binding sites, gene structure (intron/exon), SNPs
  - 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

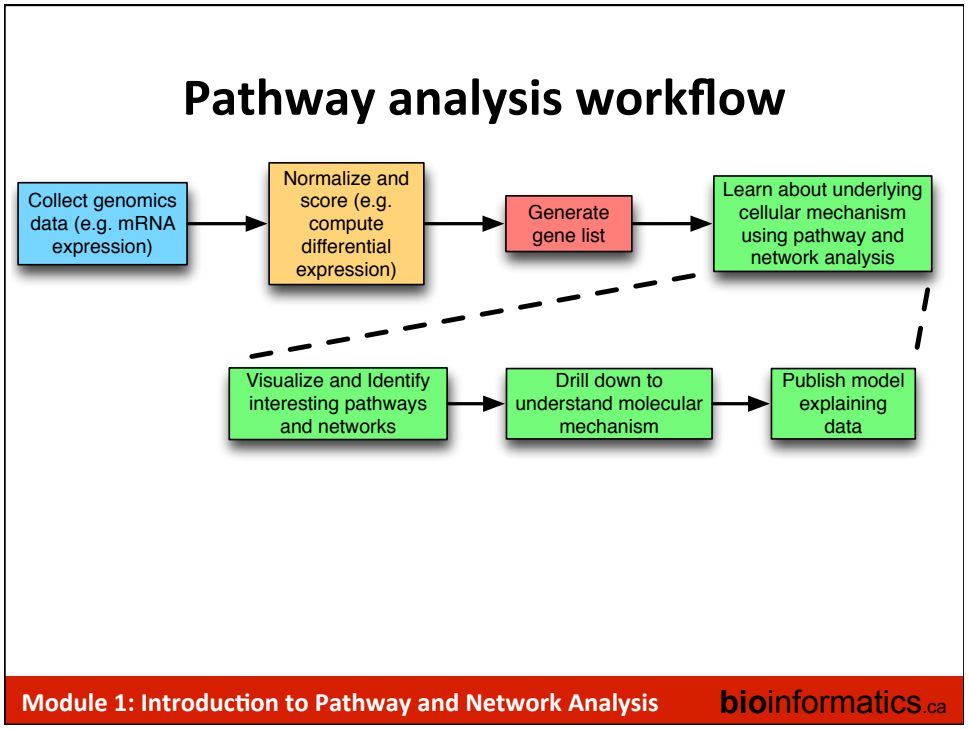
- Convenient access to gene list annotation

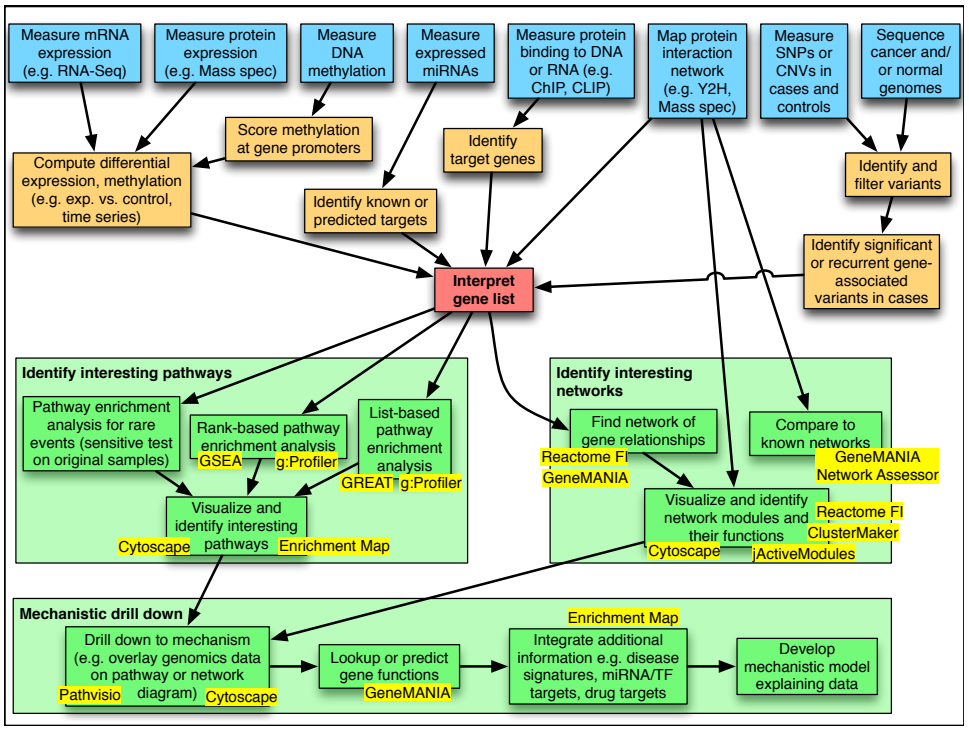


www.ensembl.org

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


Pathway analysis workflow
Collect genomics data (e.g. mRNA expression)
Normalize and score (e.g. compute differential expression)
Generate gene list
Learn about underlying cellular mechanism using pathway and network analysis
Visualize and Identify interesting pathways and networks
Drill down to understand molecular mechanism
Publish model explaining data
Module 1: Introduction to Pathway and Network Analysis
bioinformatics.ca


Measure mRNA expression (e.g. RNA-Seq)
Measure protein expression (e.g. Mass spec)
Measure DNA methylation
Measure expressed miRNAs
Measure protein binding to DNA or RNA (e.g. ChIP, CLIP)
Map protein interaction network (e.g. Y2H, Mass spec)
Measure SNPs or CNVs in cases and controls
Sequence cancer and/ or normal genomes
Score methylation at gene promoters
Identify target genes
Compute differential expression, methylation (e.g. exp. vs. control, time series)
Identify known or predicted targets
Identify and filter variants
Identify significant or recurrent gene-associated variants in cases
Interpret gene list
Identify interesting pathways
Pathway enrichment analysis for rare events (sensitive test on original samples)
Rank-based pathway enrichment analysis
GSEA
g:Profiler
List-based pathway enrichment analysis
GREAT
g:Profiler
Visualize and identify interesting pathways
Cytoscape
Enrichment Map
Identify interesting networks
Find network of gene relationships
Reactome FI
GeneMANIA
Compare to known networks
GeneMANIA
Network Assessor
Visualize and identify network modules and their functions
Reactome FI
ClusterMaker
Cytoscape
jActiveModules
Mechanistic drill down
Drill down to mechanism (e.g. overlay genomics data on pathway or network
Pathvisio
diagram)
Cytoscape
Lookup or predict gene functions
GeneMANIA
Enrichment Map
Integrate additional information e.g. disease signatures, miRNA/TF targets, drug targets
Develop mechanistic model explaining data

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.

We are on a Coffee Break & Networking Session