Gene List Functional Annotation

Dr. Jung Soh
MOL.923 SS 2018

What to do with DE features

- DE features can be
  - Genes
  - Transcripts
  - Proteins
- Analysis starts (not ends!) here
  - *De novo* assembled features
    - Require annotation to know what their functions are
  - Known features
    - Need to know what functions these indicate (as a group)

Why functional annotation

- Get biological understanding from features (genes)
- Validate experiments (RNA-seq expectations)
- Generate new hypotheses (unexpected findings)
- Limitations
  - Only as good as published or known information about genes in known species
  - Reliable annotation difficult to achieve for novel or mixed species

Gene list example

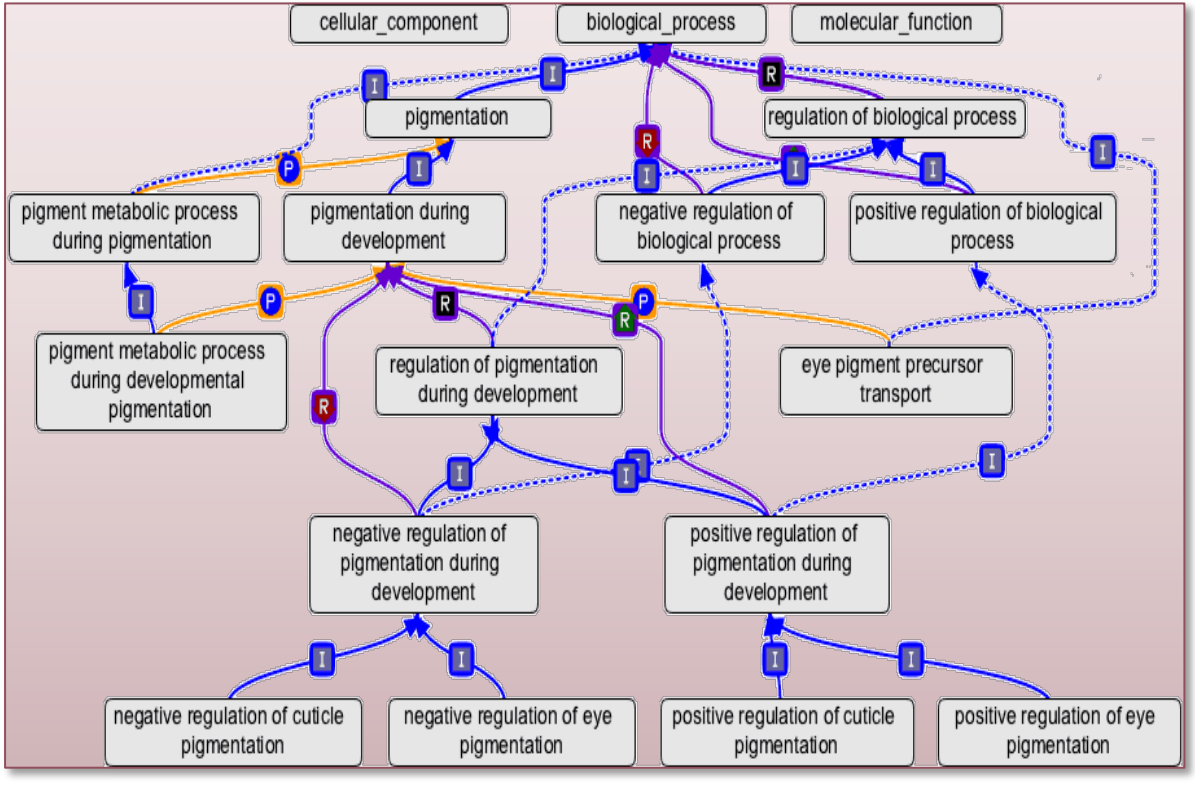
Symbol	Ensembl_ID	Name
AAK1	ENSG00000115977	AP2 associated kinase 1
ACTN1	ENSG00000072110	actinin, alpha 1
ADD3	ENSG00000148700	adducin 3 (gamma)
AEBP1	ENSG00000106624	AE binding protein 1
AIG1	ENSG00000146416	androgen-induced 1
ANKRD30B	ENSG00000224309	ankyrin repeat domain 30B
ANKRD34A	ENSG00000181039	ankyrin repeat domain 34A
ANO3	ENSG00000134343	anoctamin 3
APOE	ENSG00000130203	apolipoprotein E
ATP2B1	ENSG00000070961	ATPase, Ca++ transporting, plasma membrane 1
C14orf132	ENSG00000227051	Uncharacterized protein C14orf132
C1orf115	ENSG00000162817	Uncharacterized protein C1orf115
C1orf54	ENSG00000118292	Uncharacterized protein C1orf54 Precursor
C2orf80	ENSG00000188674	Uncharacterized protein C2orf80
C6orf62	ENSG00000112308	Uncharacterized protein C6orf62 (HBV X-transactivated gene 12 protein)
C7	ENSG00000112936	complement component 7
CACNA1C	ENSG00000151067	calcium channel, voltage-dependent, L type, alpha 1C subunit
CACNA1D	ENSG00000157388	calcium channel, voltage-dependent, L type, alpha 1D subunit
CALM3	ENSG00000160014	calmodulin 3 (phosphorylase kinase, delta)
CAMK2D	ENSG00000145349	calcium/calmodulin-dependent protein kinase II delta
CCNYL1	ENSG00000163249	cyclin Y-like 1

Sources of annotation

- Gene Ontology
  - Most popular, but can be too general
- Pathways
  - KEGG, Reactome, BioCata, PANTHER
- Protein domains
  - Pfam, InterPro
- Other databases
  - Gene families, diseases, co-expression, phenotypes

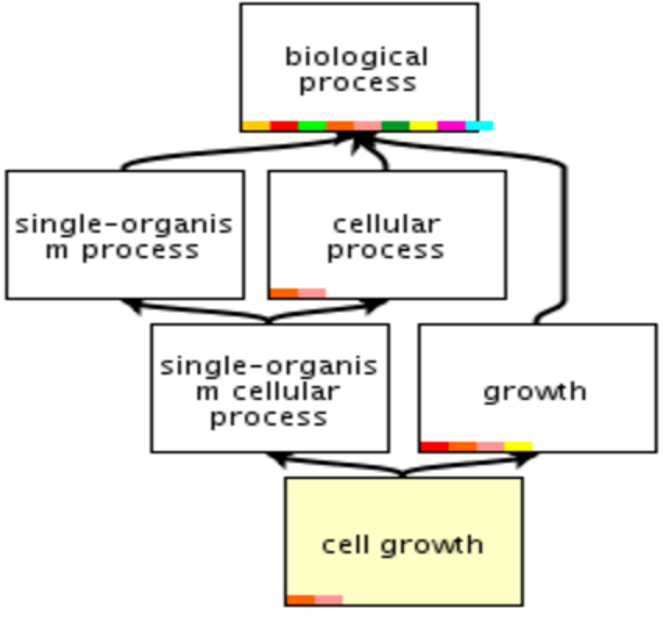
Gene Ontology (GO)

- GO project provides
  - Controlled vocabularies of terms representing gene product properties
- 3 ontologies
  - Cellular Component (CC)
  - Biological Process (BP)
  - Molecular Function (MF)



GO term example

- id: GO:0016049
- name: cell growth
- namespace: biological_process
- def: "The process in which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present." [GOC:ai]
- subset: goslim_generic
- subset: goslim_plant
- subset: gosubset_prok
- synonym: "cell expansion" RELATED []
- synonym: "cellular growth" EXACT []
- synonym: "growth of cell" EXACT []
- is_a: GO:0009987 ! cellular process
- is_a: GO:0040007 ! growth
- relationship: part_of GO:0008361 ! regulation of cell size

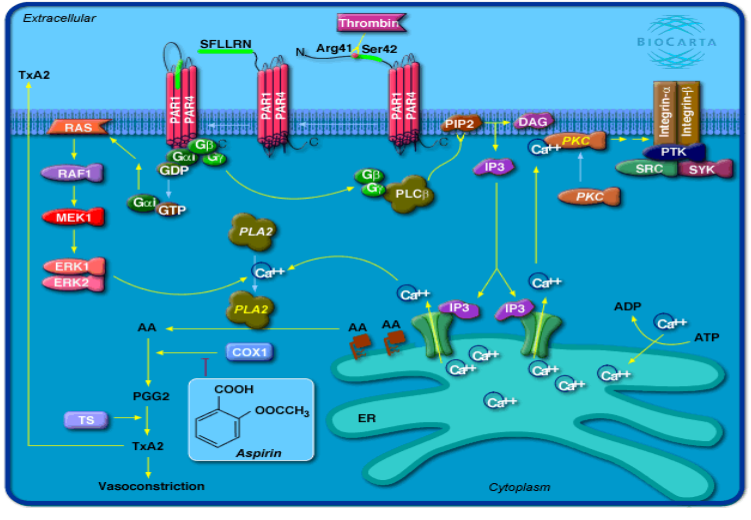


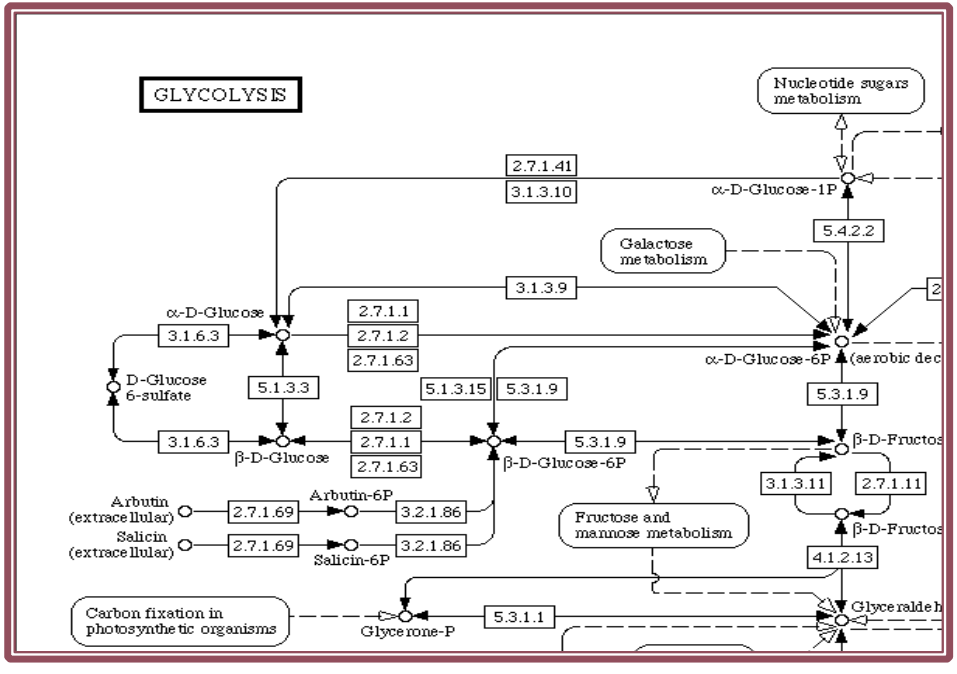
GO annotation

- Which biological processes or molecular functions are enriched in my gene list?
- How much are they enriched?
  - Fold enrichment values
- Results given as a set of enriched or over-represented GO terms
  - Degrees of enrichment usually available
  - Interpretation is important
  - Often need more information from other annotation sources

Pathways

- KEGG pathways most popular
- Many others exist
  - BioCarta
  - MetaCyc
  - Reactome





Pathway annotation

- Are there specific pathways enriched in my gene list?
- What are other genes involved in those pathways?
- Results given as a set of pathways
  - More than simple terms (enzymes, involved genes, reactions)
  - Interpretation and cross-check with other sources of annotation necessary

Gene list annotation tools

- Mostly Web-based
  - Enter your gene list (with gene symbols or other IDs)
  - Select species
  - Start analysis
  - Retrieve lots of information
- Popular websites for gene list annotation
  - PANTHER
  - DAVID
  - ToppGene
  - GOrilla

PANTHER

- One of the most widely used gene list functional annotation tools
- Annotations include
  - GO terms (mostly)
  - Also
    - PANTHER protein classes
    - PANTHER pathways

Protein ANnotation THrough Evolutionary Relationship

DAVID

- Uses several annotation sources
  - GO, pathways, and many others
- Annotation presented in several ways
  - Functional annotation chart
  - Functional annotation table
  - Annotation clusters
- Lots of information provided by hyperlinks
  - Mostly useful, sometimes redundant
  - Finding what you want is not too easy!

Database for Annotation, Visualization and Integrated Discovery

ToppGene

- Uses several annotation sources with focus on disease information
  - GO, pathways, gene families, phenotypes, diseases
- Annotation provided on a single long page
  - Easy to find annotation terms (it's either there or not there)
  - Limited visualization

Lab overview