Introduction to functional enrichment analysis

Monash Genomics & Bioinformatics Platform

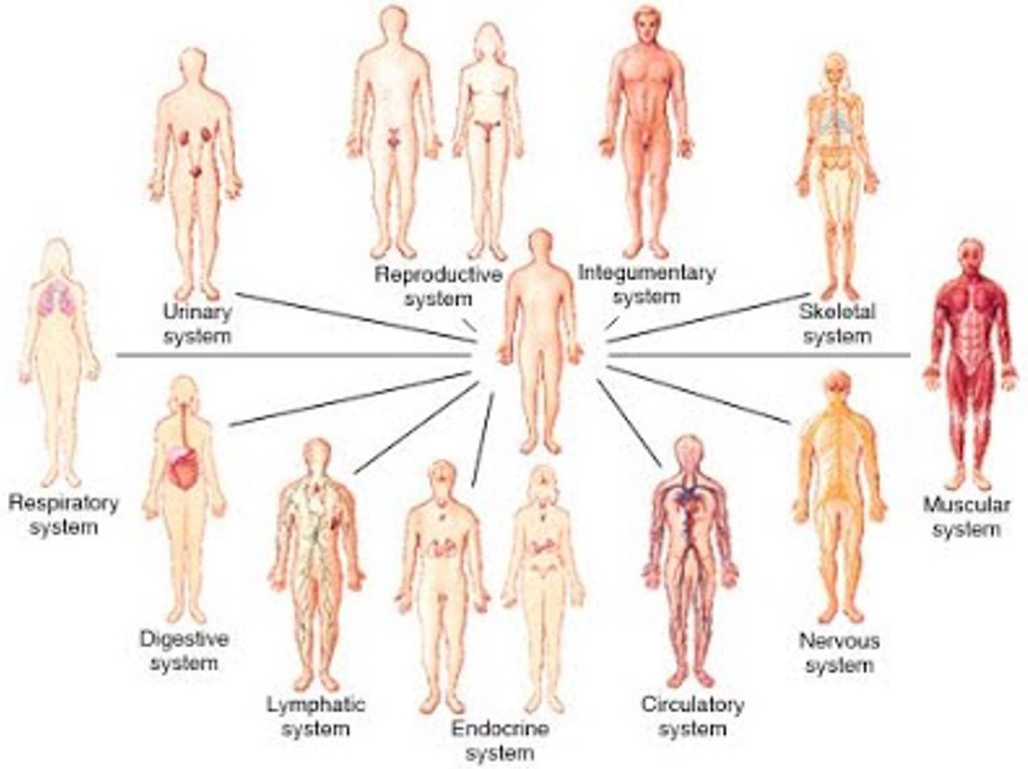
https://tinyurl.com/omicsMar24

Overview

- Omics Experiment & Terminologies
- Functional Classification
  - Input list
  - Gene Ontology
  - Pathways
- Enrichment Analysis Overview
- Statistical Concepts
- Gene-list based enrichment analysis + GSEA
- Limitations and pitfalls

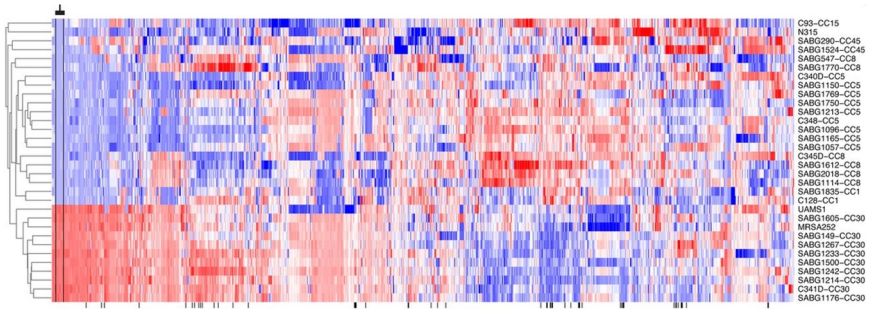
One Genome

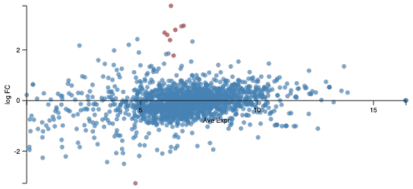
- Epigenome
- Transcriptome
- Proteome
- Kinome
- Metabolome



Many phenotypic states

High-throughput Experiment





- Monitor systems by observing the behaviour of in the order of 100s and 10^6 molecules per experiment
- Results in the order of 10^2- 10^4 features, as a list

How do we extract the meaning and mechanistic insight from long list of molecules?

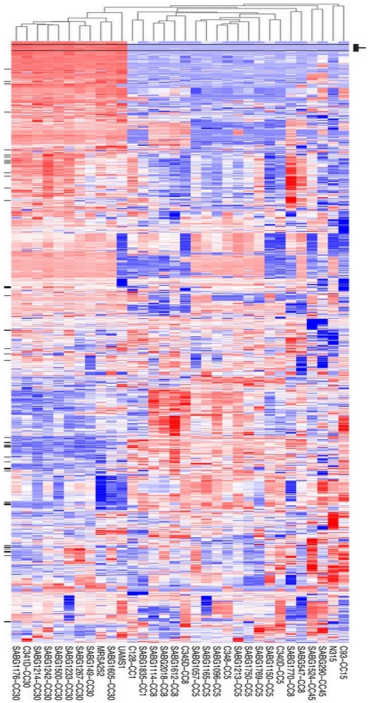
Example - RNA-Seq (transcriptomics)

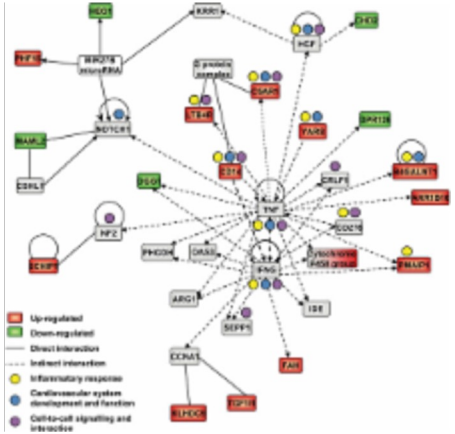
Gene	FC	p-value
A1BG	1.34	5.15E-05
A2M	1.35	0.00283
AAK1	-1.09	0.0238
ABAT	-2.15	0.00403
ABCC4	-2.17	0.00249
ABCF1	-1.28	0.00147
ABHD10	-1.77	0.00182
ABHD11	-3.66	3.81E-06
ABHD14B	-1.97	0.024
ACAA1	-1.84	0.00414
ACAA2	1.11	0.0156
ACACA	-2.92	0.000124
ACAD8	-1.69	0.0116
ACAD9	-1.88	8.50E-06
ACADSB	-2.53	0.000634
ACBD3	-2.12	0.000444
ACE	1.82	0.025
ACIN1	-1.64	2.57E-05
ACLY	-2.02	1.98E-05
ACOX3	-1.41	0.0197
ACP1	-1.55	0.00273
ACP2	-1.94	3.75E-05
ACSL1	-2.68	0.000584
ACSL3	-1.91	0.00166
ACSM3	-2.28	0.000703
ACSS1	-3.04	2.16E-05
ACTG2	1.85	0.00507
ACY1	-2.72	2.61E-05
ADAR	-2.57	2.34E-06
ADH1B	2.48	0.00574
ADPRHL2	-1.79	0.00156
ADSL	-1.43	0.000453
AEBP1	1.28	0.002
AFM	1.98	0.00602
AGA	-2.3	0.00226
AGK	-1.26	0.00693
AGT	1.27	0.000142
AHSG	1.24	0.000589
AIFM1	-1.03	0.000553
AIM1	-2.18	0.001
AIMP1	-2.47	0.000113
AIMP2	-1.26	0.0023
AK4	-1.55	0.00644
ALAD	1.71	1.90E-05
ALCAM	-2.39	0.000149
ALDH18A1	-3.04	2.17E-07
ALDH5A1	-1.51	0.0172
ALDH6A1	-1.28	5.97E-06
AMACR	-3.46	1.23E-05
AMBP	1.24	0.000694

Downstream analysis- what do the results mean?

- What genes behave differently between two conditions, and what does this tell us about the condition of interest?

Gene	FC	p-value
A1BG	1.34	5.15E-05
A2M	1.35	0.00283
AAK1	-1.09	0.0238
ABAT	-2.15	0.00403
ABCC4	-2.17	0.00249
ABCF1	-1.28	0.00147
ABHD10	-1.77	0.00182
ABHD11	-3.66	3.81E-06
ABHD14B	-1.97	0.024
ACAA1	-1.84	0.00414
ACAA2	1.11	0.0156
ACACA	-2.92	0.000124
ACAD8	-1.69	0.0116
ACAD9	-1.88	8.50E-06
ACADSB	-2.53	0.000634
ACBD3	-2.12	0.000444
ACE	1.82	0.025
ACIN1	-1.64	2.57E-05
ACLY	-2.02	1.98E-05
ACOX3	-1.41	0.0197
ACP1	-1.55	0.00273
ACP2	-1.94	3.75E-05
ACSL1	-2.68	0.000584
ACSL3	-1.91	0.00166
ACSM3	-2.28	0.000703
ACSS1	-3.04	2.16E-05
ACTG2	1.85	0.00507
ACY1	-2.72	2.61E-05
ADAR	-2.57	2.34E-06
ADH1B	2.48	0.00574
ADPRHL2	-1.79	0.00156
ADSL	-1.43	0.000453
AEBP1	1.28	0.002



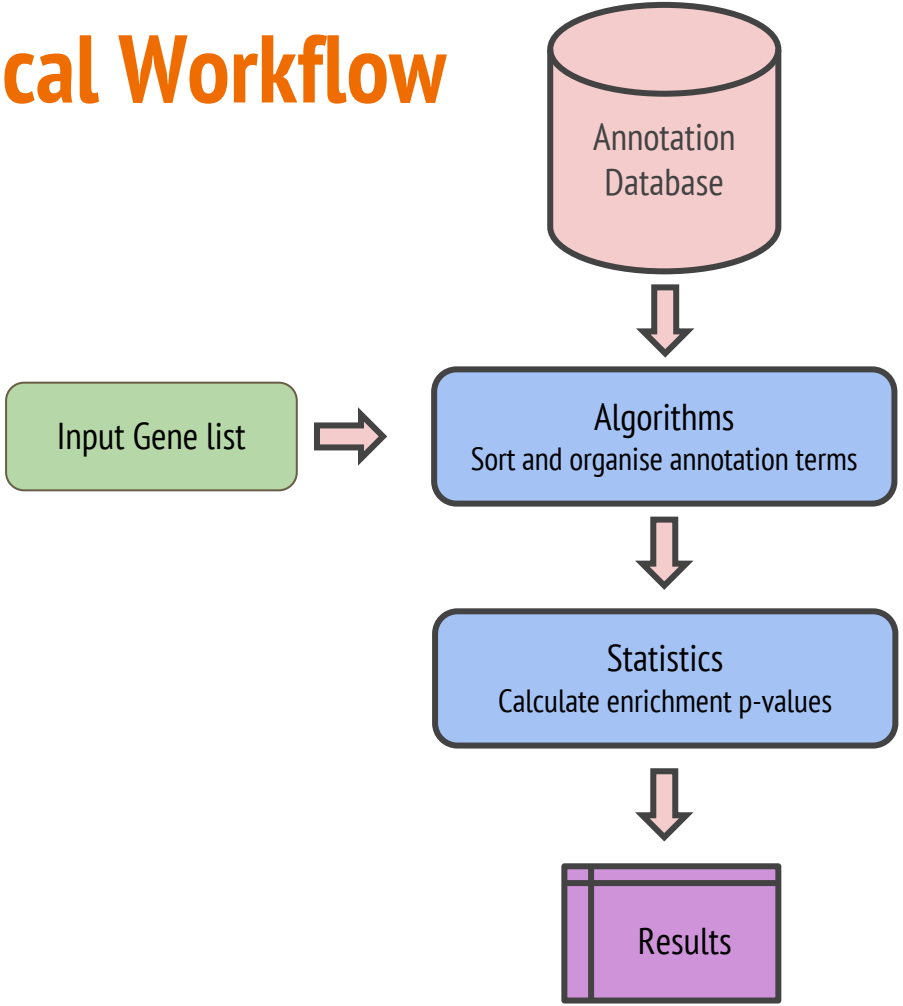


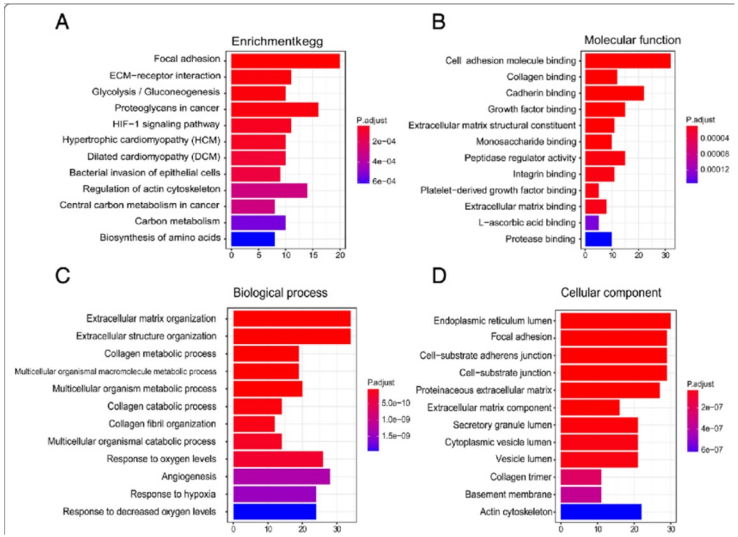
Functional analysis of genes/proteins

- Given a set (potenti : do they mean?
- Knowledge-driven
- We try and partitio low us to identify trends in th est in our experiment
- Can be based on prior knowledge or be derived from experimental data
- **The aim is to give a number (score or p-value) to function/pathway.**

Gene	FC	p-value
A1BG	1.34	5.15E-05
A2M	1.35	0.00283
AAK1	-1.09	0.0238
ABAT	-2.15	0.00403
ABCC4	-2.17	0.00249
ABCF1	-1.28	0.00147
ABHD10	-1.77	0.00182
ABHD11	-3.66	3.81E-06
ABHD14B	-1.97	0.024
ACAA1	-1.84	0.00414
ACAA2	1.11	0.0156
ACACA	-2.92	0.000124
ACAD8	-1.69	0.0116
ACAD9	-1.88	8.50E-06
ACADSB	-2.53	0.000634
ACBD3	-2.12	0.000444
ACE	1.82	0.025
ACIN1	-1.64	2.57E-05
ACLY	-2.02	1.98E-05
ACOX3	-1.41	0.0197
ACP1	-1.55	0.00273
ACP2	-1.94	3.75E-05
ACSL1	-2.68	0.000584
ACSL3	-1.91	0.00166
ACSM3	-2.28	0.000703
ACSS1	-3.04	2.16E-05
ACTG2	1.85	0.00507
ACY1	-2.72	2.61E-05
ADAR	-2.57	2.34E-06
ADH1B	2.48	0.00574
ADPRHL2	-1.79	0.00156
ADSL	-1.43	0.000453
AEBP1	1.28	0.002
AFM	1.98	0.00602
AGA	-2.3	0.00226
AGK	-1.26	0.00693
AGT	1.27	0.000142
AHSG	1.24	0.000589
AIFM1	-1.03	0.000553
AIM1	-2.18	0.001
AIMP1	-2.47	0.000113
AIMP2	-1.26	0.0023
AK4	-1.55	0.00644
ALAD	1.71	1.90E-05
ALCAM	-2.39	0.000149
ALDH18A1	-3.04	2.17E-07
ALDH5A1	-1.51	0.0172
ALDH6A1	-1.28	5.97E-06
AMACR	-3.46	1.23E-05
AMBP	1.24	0.000694

Typical Workflow





Visualisation

Input list

Gene	FC	p-value
A1BG	1.34	5.15E-05
A2M	1.35	0.00283
AAK1	-1.09	0.0238
ABAT	-2.15	0.00403
ABCC4	-2.17	0.00249
ABCF1	-1.28	0.00147
ABHD10	-1.77	0.00182
ABHD11	-3.66	3.81E-06
ABHD14B	-1.97	0.024
ACAA1	-1.84	0.00414
ACAA2	1.11	0.0156
ACACA	-2.92	0.000124
ACAD8	-1.69	0.0116
ACAD9	-1.88	8.50E-06
ACADSB	-2.53	0.000634
ACBD3	-2.12	0.000444
ACE	1.82	0.025
ACIN1	-1.64	2.57E-05
ACLY	-2.02	1.98E-05
ACOX3	-1.41	0.0197
ACP1	-1.55	0.00273
ACP2	-1.94	3.75E-05
ACSL1	-2.68	0.000584
ACSL3	-1.91	0.00166
ACSM3	-2.28	0.000703
ACSS1	-3.04	2.16E-05
ACTG2	1.85	0.00507
ACY1	-2.72	2.61E-05
ADAR	-2.57	2.34E-06
ADH1B	2.48	0.00574
ADPRHL2	-1.79	0.00156
ADSL	-1.43	0.000453
AEBP1	1.28	0.002
AFM	1.98	0.00602
AGA	-2.3	0.00226
AGK	-1.26	0.00693
AGT	1.27	0.000142
AHSG	1.24	0.000589
AIFM1	-1.03	0.000553
AIM1	-2.18	0.001
AIMP1	-2.47	0.000113
AIMP2	-1.26	0.0023
AK4	-1.55	0.00644
ALAD	1.71	1.90E-05
ALCAM	-2.39	0.000149
ALDH18A1	-3.04	2.17E-07
ALDH5A1	-1.51	0.0172
ALDH6A1	-1.28	5.97E-06
AMACR	-3.46	1.23E-05
AMBP	1.24	0.000694

Terminologies

- How to define a gene-list?
- Test set / Differentially expressed Biomolecules
  - Whole list
  - Up-regulated molecules
  - Down-regulated molecules
  - Individual cluster members
- Background set
- ID mapping (non-model organism)
  - Finding orthologs

Annotation Databases

Functional Classification of Biomolecules

- **Gene Ontology**
  - Controlled vocabulary that describes the function of gene products, the process in which they participate and the locations they are found in the cell.
- **Pathways**
  - Databases of “canonical” pathways capturing what we know about signalling, metabolism, DNA repair and other cellular processes
- **Empirical gene sets**
  - Those are derived from carefully controlled experiments that are thought to capture the molecular phenotype of given system

Gene Ontology

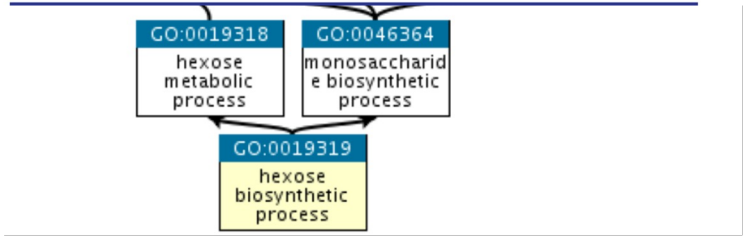
Gene Ontology

- Controlled and structured hierarchical vocabulary for describing the properties and functions of "**gene products**"
- Three categories of attributes
  - Biological Processes
  - Molecular Functions
  - Cellular Components
- **Hierarchical**- parents and child terms establish more-general and more-specific descriptors of function

Gene Ontology: tool for the unification of biology

The Gene Ontology Consortium*

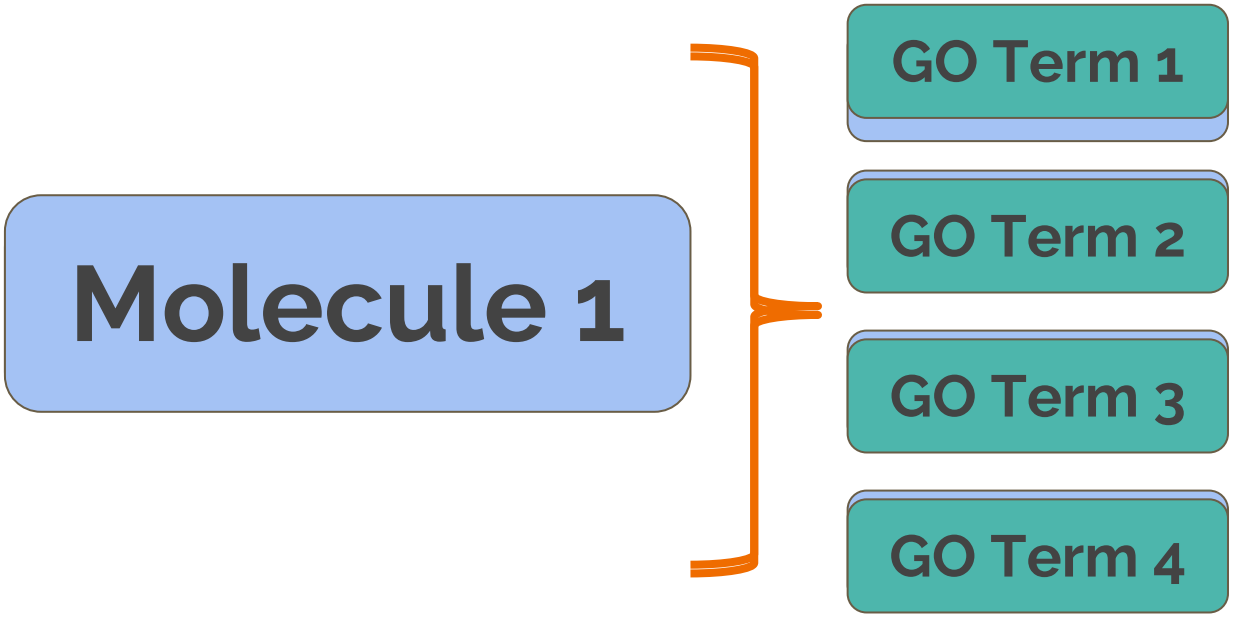
Genomic sequencing has made it clear that a large fraction of the genes specifying the core biological functions are shared by all eukaryotes. Knowledge of the biological role of such shared proteins in one organism can often be transferred to other organisms. The goal of the Gene Ontology Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing. To this end, three independent ontologies accessible on the World-Wide Web (http://www.geneontology.org) are being constructed: biological process, molecular function and cellular component.



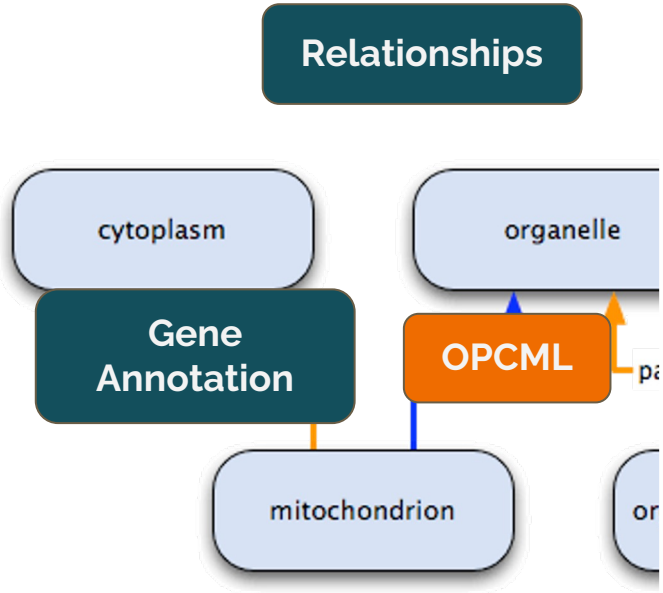
This reflect the fact that:

- *biosynthetic process* is a subtype of *metabolic process* and
- a *hexose* is a subtype of *monosaccharide*

GO Term vs. Gene Annotation



Example



Gene Product	Symbol	Qualifier	GO Term	Evidence
UniProtKB:Q14982	OPCML	involved_in	GO:0008038 (P) neuron recognition	ECO:0000304 TAS
UniProtKB:Q14982	OPCML	involved_in	GO:0007155 (P) cell adhesion	ECO:0000304 TAS
UniProtKB:Q14982	OPCML	part_of	GO:0005886 (C) plasma membrane	ECO:0000304 TAS
UniProtKB:Q14982	OPCML	part_of	GO:0005886 (C) plasma membrane	ECO:0000304 TAS
UniProtKB:Q14982	OPCML	part_of	GO:0005576 (C) extracellular region	ECO:0000304 TAS
UniProtKB:Q14982	OPCML	part_of	GO:0031225 (C) anchored component of membrane	ECO:0000322 IEA
UniProtKB:Q14982	OPCML	part_of	GO:0016020 (C) membrane	ECO:0000322 IEA
UniProtKB:Q14982	OPCML	part_of	GO:0005886 (C) plasma membrane	ECO:0000322 IEA
UniProtKB:Q14982	OPCML	involved_in	GO:0007155 (P) cell adhesion	ECO:0000322 IEA

https://geneontology.org/docs/ontology-relations/

Gene Annotation

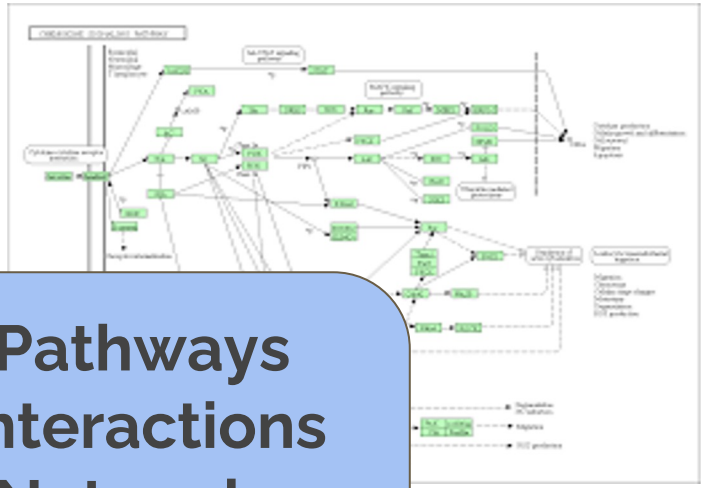
Evidence Type	Evidence Code
Experimental	Inferred from Experiment (EXP) Inferred from Direct Assay (IDA) Inferred from Physical Interaction (IPI) Inferred from Mutant Phenotype (IMP) Inferred from Genetic Interaction (IGI) Inferred from Expression Pattern (IEP)Inferred from Experiment (EXP)
Phylogenetically-inferred	Inferred from Biological aspect of Ancestor (IBA) Inferred from Biological aspect of Descendant (IBD) Inferred from Key Residues (IKR) Inferred from Rapid Divergence (IRD)
Computational analysis	Inferred from Sequence or structural Similarity (ISS) Inferred from Sequence Orthology (ISO) Inferred from Sequence Alignment (ISA) Inferred from Sequence Model (ISM) Inferred from Genomic Context (IGC) Inferred from Reviewed Computational Analysis (RCA)
Author statement	Traceable Author Statement (TAS) Non-traceable Author Statement (NAS)
Curator statement	Inferred by Curator (IC) No biological Data available (ND)
Electronic annotation	Inferred from Electronic Annotation (IEA)

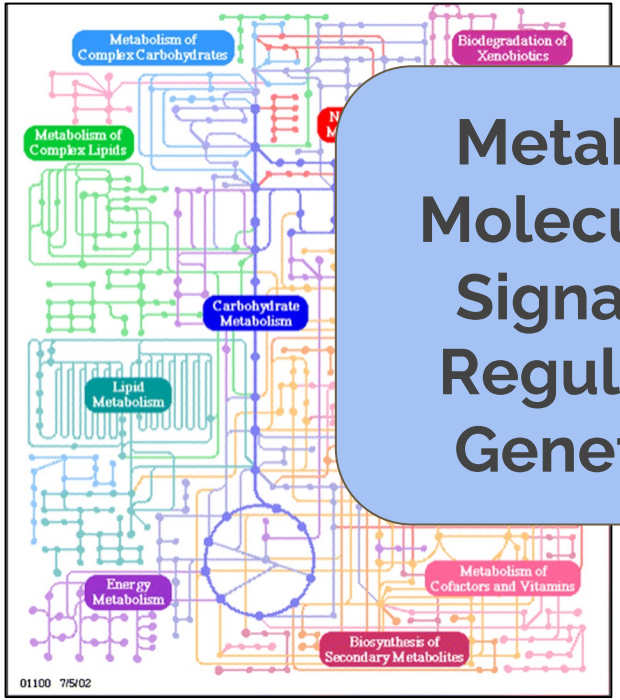
http://geneontology.org/docs/guide-go-evidence-codes/

Pathway Annotation

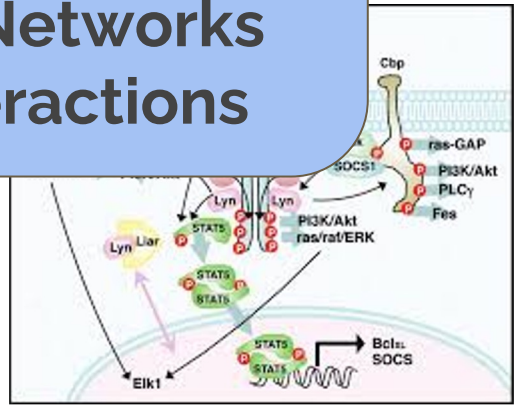
Pathway Commons

Pathways in Biology

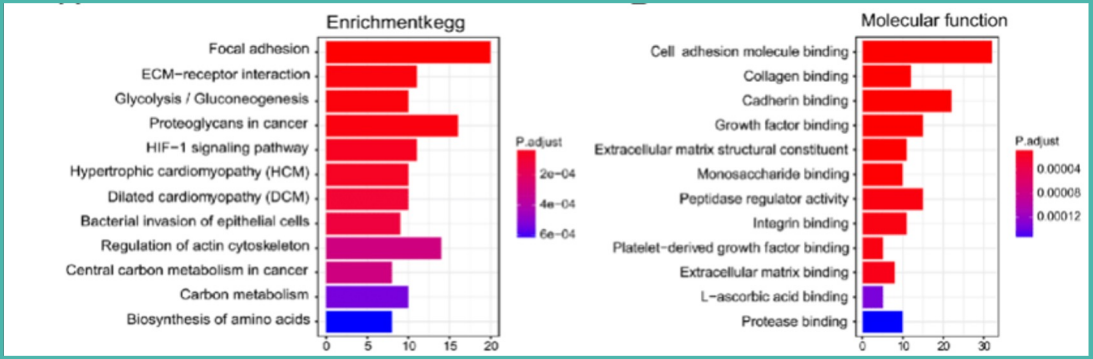




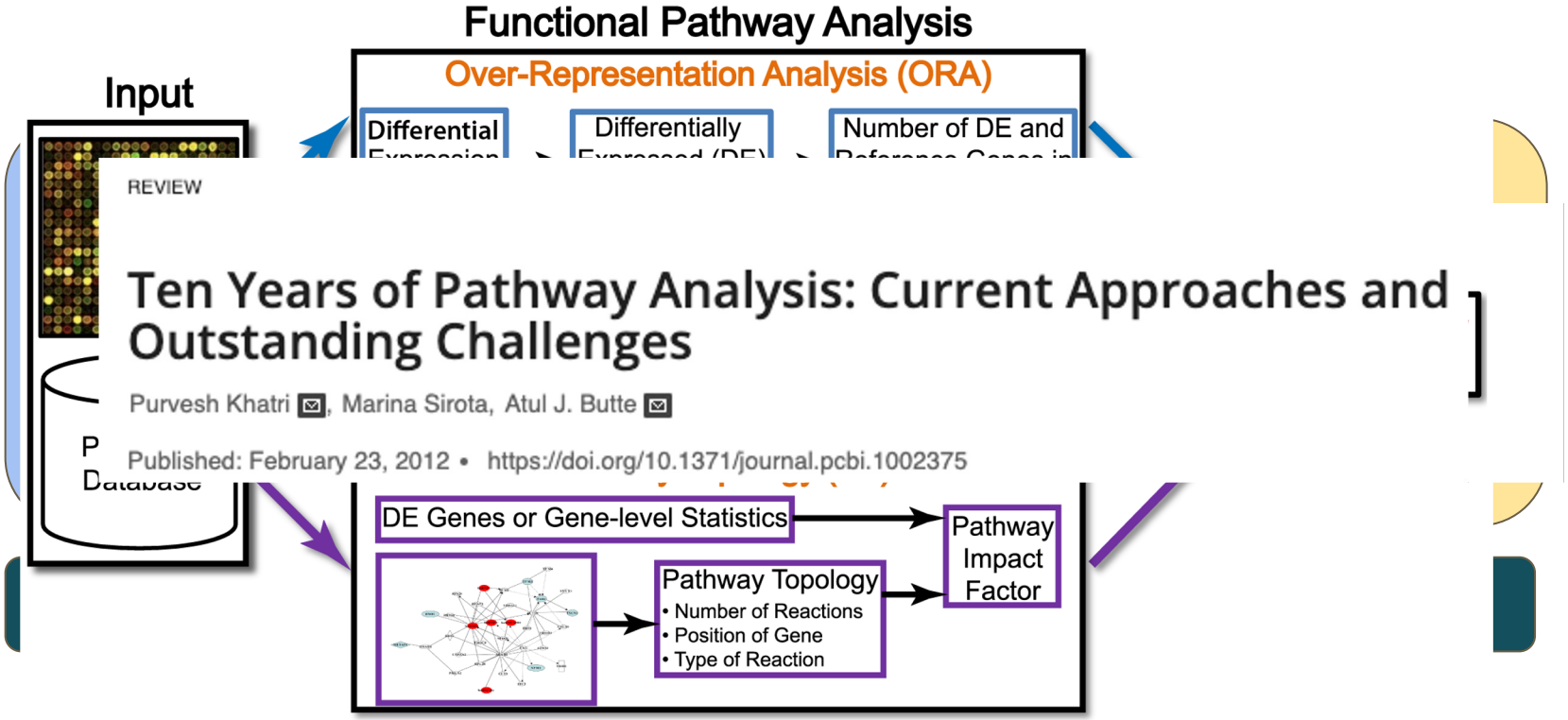
Metabolic Pathways
Molecular Interactions
Signalling Networks
Regulatory Networks
Genetic Interactions



Functional analysis Overview



Enrichment analysis



REVIEW

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri, Marina Sirota, Atul J. Butte

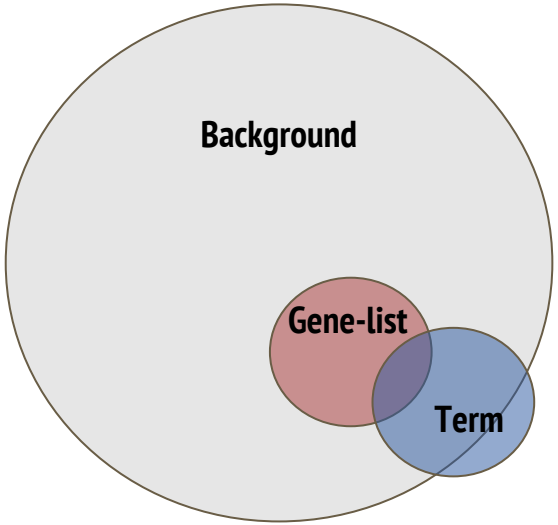
Published: February 23, 2012 • https://doi.org/10.1371/journal.pcbi.1002375

Gene-list based enrichment analysis

Statistical Concepts

Fisher's Exact test

- Test for overlap



The 'lady tasting tea' experiment

Sir Ronald Fisher

Gene	.. in term	.. not in term	Total
.. in gene-list	50	100	150
.. not in gene-list	200	15900	16100
Total	250	16000	16250

```
fisher.test(x, alternative='greater')
phyper(x, lower.tail=TRUE)
```

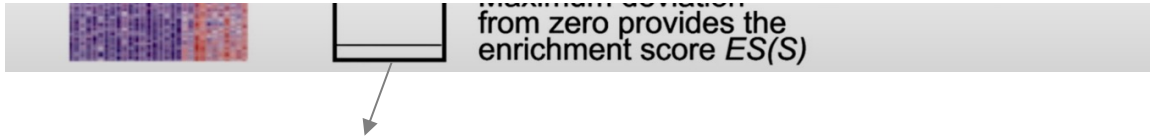
Ranked based enrichment analysis

GSEA

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [c]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [d]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [e]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [f]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [g]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [i]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [j]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142



- Mitochondria
- MAP kinase (Mitogen-activated protein Kinase) signalling
- Cell cycle control
- .

https://www.pnas.org/doi/epdf/10.1073/pnas.0506580102

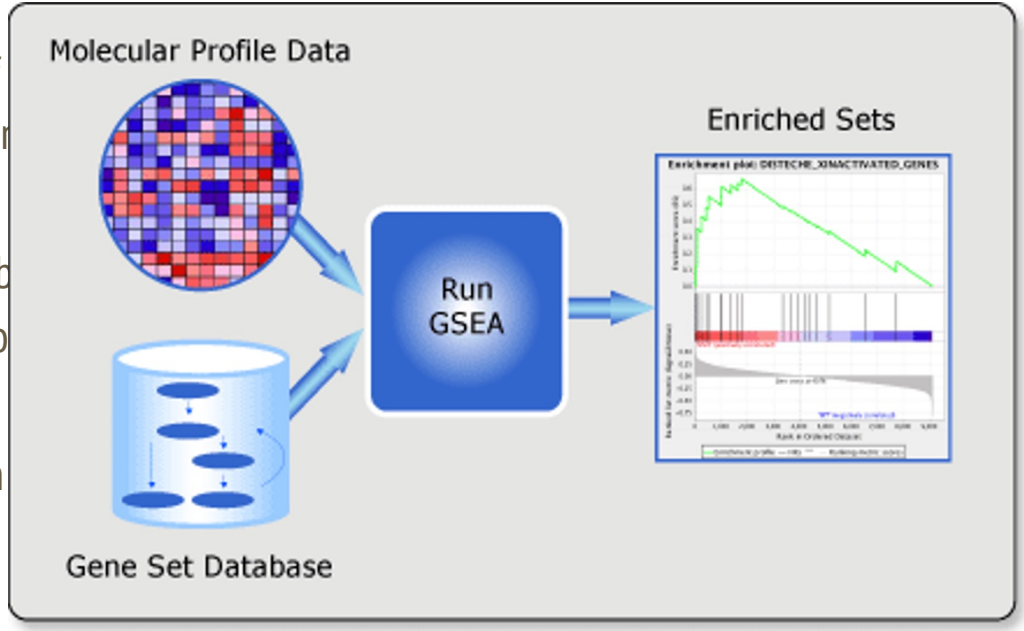
Enrichment Statistics Hands-on

Over-representation analysis

- Input
  - Gene list (Expression dataset), typically gene symbols or identifiers.
  - Background
- Pro's
  - Simple interpretation: Identifies over-represented gene sets or pathways.
  - Easy to implement and understand.
- Con's
  - Effect size and statistical significance of biomolecules are ignored (Log fold change or p-value or rank)
  - Uses only most significant genes and discards others (P-value 0.06)
  - Each gene is independent of each other

Rank-Based analysis

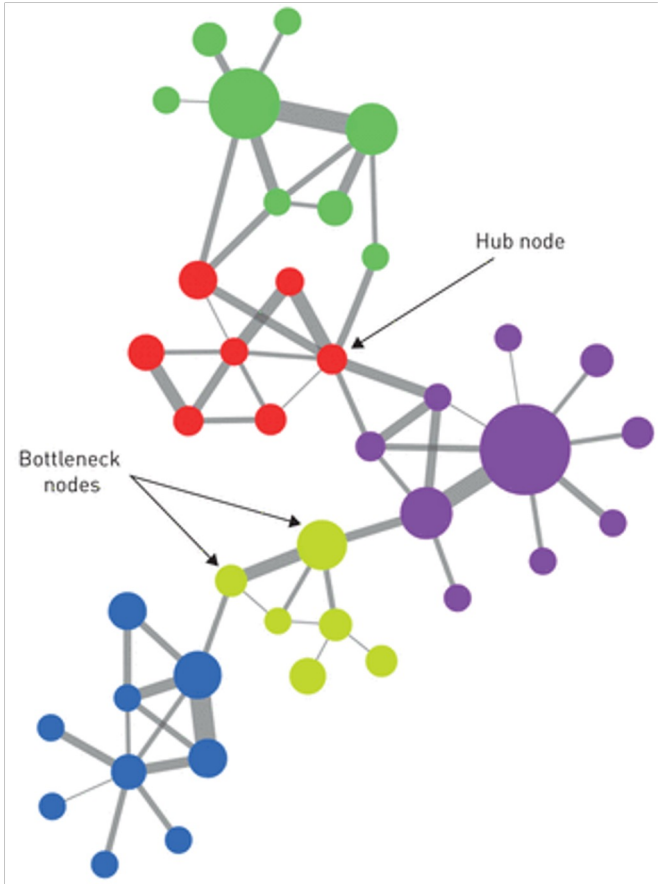
- Input
  - Ranked list of
  - Typically a ger
- Pro's
  - Use all availab
  - Dependence o
- Con's
  - Analyses each



s in gene

Topology based enrichment

- Considers position of biomolecules in the pathway
- Limitations
  - True pathway topology is dependent on the type of cell due to cell-specific gene expression profiles and condition being studied. However, this information is rarely available and is fragmented in literature



Some considerations

- Enrichment vs. Depletion
- Multiple hypothesis correction
- Importance of background, what is it.
- How often a annotation database being updated?
- Effect of gene list size? What size do you need?

Summary

- Terminologies
- Annotation Databases
- Enrichment Analysis Overview
- Statistical Concepts
- Gene-list based enrichment analysis
- Limitations and considerations

R packages and online tools

- enrichR
- ReactomePA
- fgsea
- gprofiler2
- STRINGdb
- clusterProfiler
- WebGestaltR

- WebGestalt: https://www.webgestalt.org/
- ShinyGO: http://bioinformatics.sdstate.edu/go/
- https://rokai.io/ & https://rokai.io/explorer/ (Robust Inference of Kinase Activity - RoKAI)

ǫ

GO Categories

Molecular Function (MF)	Molecular-level activities performed by gene products. *catalytic activity* and *transporter activity*; *adenylate cyclase activity* or *Toll-like receptor binding*. GO molecular functions are often appended with the word "activity" (a *protein kinase* would have the GO molecular function *protein kinase activity*).
Cellular Component (CC)	A location, relative to cellular compartments and structures. *cellular anatomical entities*, includes cellular structures such as the *plasma membrane* and the *cytoskeleton*, as well as membrane-enclosed cellular compartments such as the *mitochondrion*
Biological Process (BP)	The larger processes, or 'biological programs' accomplished by multiple molecular activities. *DNA repair* or *signal transduction*. *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*.

An example of GO annotation: human "cytochrome c":
molecular function *oxidoreductase activity*,
the **biological process** *oxidative phosphorylation*, and
the **cellular component** *mitochondrial intermembrane space*.

Note: a biological process is not equivalent to a pathway.

https://geneontology.org/docs/ontology-documentation