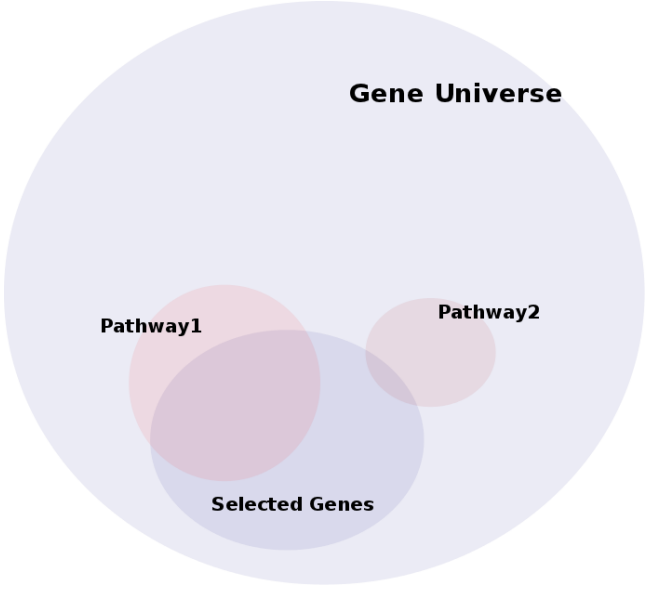
Functional Enrichment Analysis

Falko Noé

University of Zurich[UZH]
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Explorative Functional Analysis

- Challenge
- Functional databases
- Functional enrichment
  - Overrepresentation analysis
  - Gene set enrichment analysis
- Tools

Data Interpretation - Challenge

Differential Gene Expression

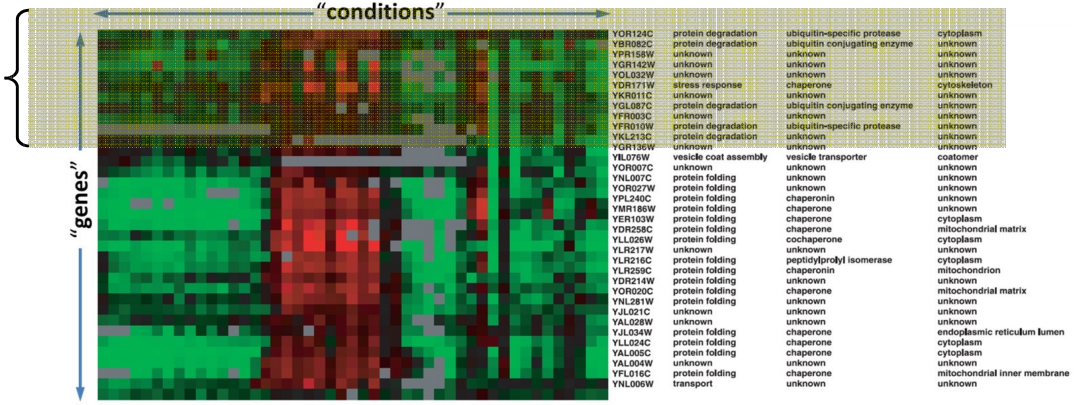
- Hundreds/thousands of candidate genes
- Gene-by-gene interpretation difficult
- Linking different experiments

- Filter results (Venn)
- Identify:
  - key molecules (TFs, miRNAs, common effectors, master regulators)
  - Enriched Biological Processes/Pathways
  - Networks (links across candidates)

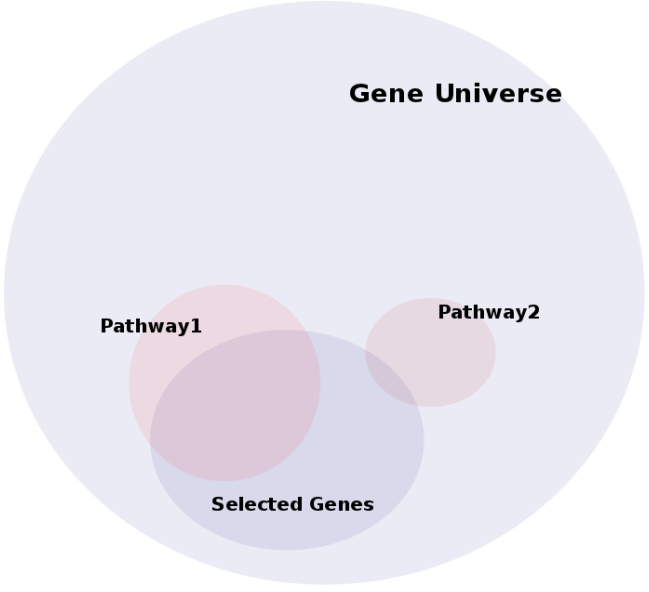
Filtering list of DEGs

- FDR cutoff < 0.05-0.1 → p-value cutoff < 0.01
- Fold change cutoff (e.g., > 2 fold change)
- Rank list (top 10% based on p-value / fold change)
- Clustered genes with similar expression patterns
- Candidate gene list

Filtered/selected gene set



Explorative Functional Analysis



- Challenge
- Functional databases
- Functional enrichment
  - Overrepresentation analysis
  - Gene set enrichment analysis
- Tools

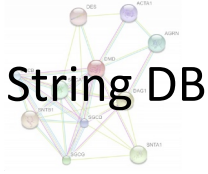
Which functional databases can be interrogated?

- **Gene Ontology**

http://amigo1.geneontology.org/cgi-bin/amigo/go.cgi

- **Pathways**

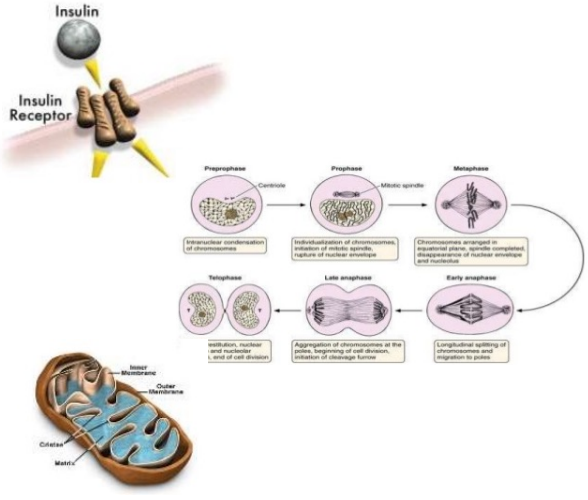
- **Protein class**



These databases are typically constructed based on protein-protein interaction experiments, signaling pathway disruption experiment, literature screening (and combinations of the above)

Gene Ontology (GO) terms

- Three ontology domains:
  1. Molecular function (MF): basic activity or task
     e.g. catalytic activity, calcium ion binding
  2. Biological process (BP): broad objective or goal
     e.g. signal transduction, immune response
  3. Cellular component (CC): location or complex
     e.g. nucleus, mitochondrion



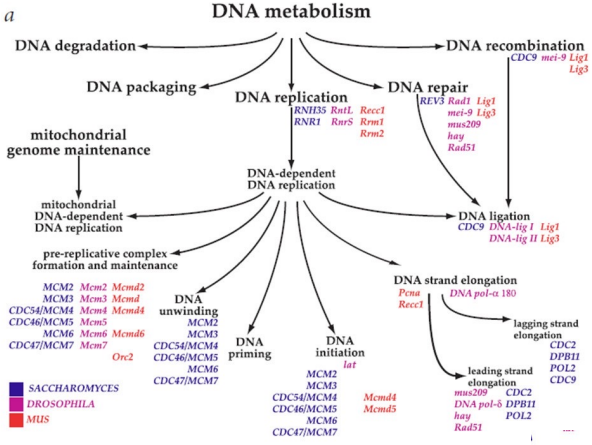
- Genes can have multiple annotations:

Cytochrome gene

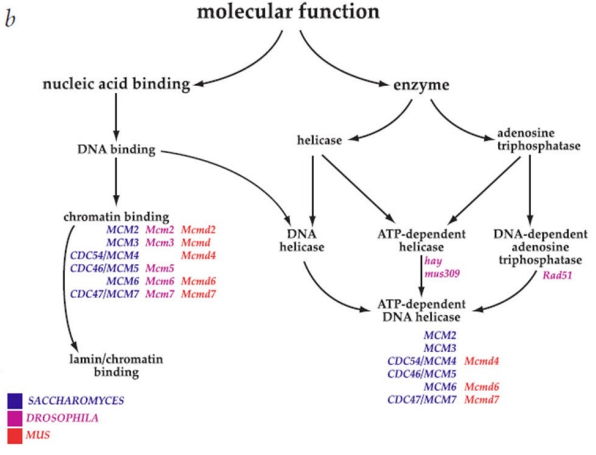
- Molecular function (MF): oxidoreductase activity
- Biological process (BP): oxidative phosphorylation , induction of cell death
- Cellular component (CC): mitochondrial matrix, mitochondrial inner membrane

GO terms are hierarchical

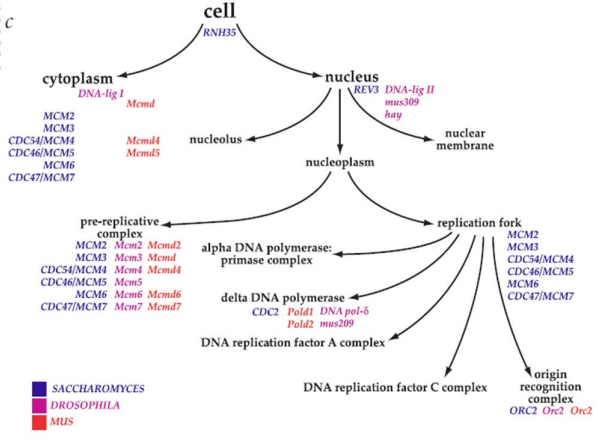
Biological process



Molecular function



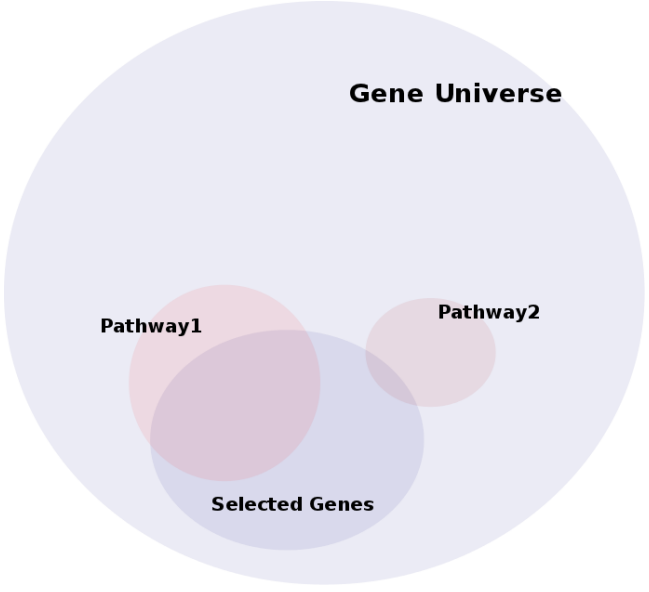
Cellular component



Functional databases: limitations

- Accurate functional annotation is only available for model organisms
- Functional terms are abundant and highly correlated
  - Multiple testing of interdependent terms (FDR)

Explorative Functional Analysis

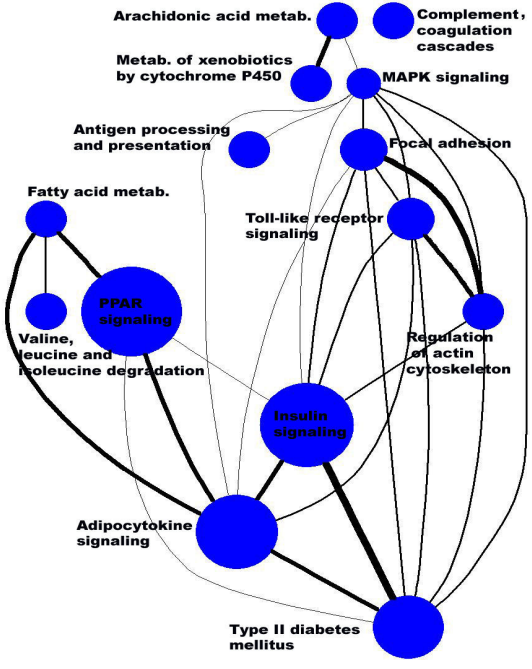


- Challenge
- Functional databases
- Functional enrichment
  - Overrepresentation analysis
  - Gene set enrichment analysis
- Tools

Overrepresentation analysis

Case study:

- With a meta analysis Rasche et al. identified 213 genes associated with Type II Diabetes
- For a given KEGG pathway, they counted how many of these 213 genes are members of this pathway
- If this number is more than expected by chance, then this pathway could potentially be relevant to Type II Diabetes
  - Fisher's exact test



BMC Genomics. 2008 Jun 30;9:310. doi: 10.1186/1471-2164-9-310.

Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus.

Rasche A[1], Al-Hasani H, Herwig R.

Fisher's exact test

- Statistical test used to determine if there are non-random associations between two categorical variables.
- Fisher devised the test following a comment from a colleague, who claimed to be able to detect whether the tea or the milk was added first to her cup.

Contingency table:

	Present in gene list	Not present in gene list	Total
Member of KEGG pathway	107	637	780
Not member of KEGG pathway	452	8673	9125
Total	559	9346	9905

- Free tools
  - **WebGestalt**, R-packages, Cytoscape
- Commercial tools
  - MetaCore/GeneGo, Ingenuity (IPA)

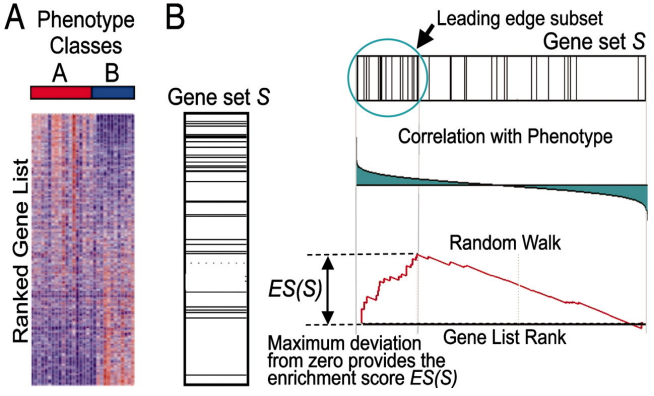
Observed: **107**

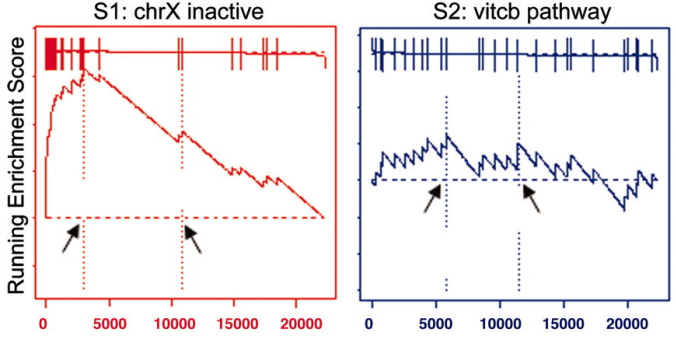
Expected: $\frac{780 * 559}{9905} = \mathbf{44}$

Enrichment: **2.43x,** p-val < 0.00001

Gene set enrichment analysis (GSEA)

- **Input**: Unfiltered genes ordered in a ranked list L, according to their fold change or p-value
- **Goal**: Determine whether members of a gene set S are randomly distributed throughout the list L or tend to occur toward the top (or bottom) of L, as measured by Enrichment score (ES)
- **Enrichment score (ES):** The degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L
- Random walk down the list L, increasing the running-sum when encounter a gene in S and decreasing it when encounter a gene not in S.
- ES is the max deviation from zero encountered in the random walk (Kolmogorov–Smirnov test)





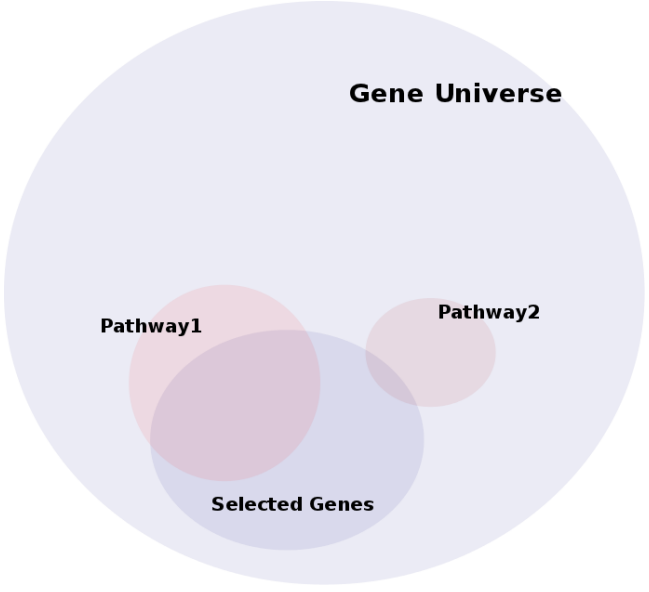
Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30.

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

Subramanian A[1], Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.

Gene set enrichment analysis (GSEA): advantages

- Does not require a priori selection (e.g., p-value thresholds, clusters)
  - Not risking to generate too short lists
  - Meta-analyses are not limited by overlapping factors
- Aggregated small effects better captured
  - (would potentially miss with ORA)



Explorative Functional Analysis

- Challenge
- Functional databases
- Functional enrichment
  - Overrepresentation analysis
  - Gene set enrichment analysis
- Tools

Tools for functional enrichment analysis

- Free tools
  - WebGestalt (http://www.webgestalt.org/)
  - Panther (http://pantherdb.org/)
  - David (https://david.ncifcrf.gov/)
  - R Packages (topGO, GSEABase, **clusterProfiler**, ...)
  - GSEA (http://software.broadinstitute.org/gsea/index.jsp)
  - Cytoscape
- Commercial tools
  - MetaCore/GeneGO (https://portal.genego.com/)
  - Ingenuity Pathway Analysis (https://www.qiagenbioinformatics.com/products/ingenuity- pathway-analysis/)

GO Enrichment - Example from FGCZ Reports

GO-Term	P-Value	Ratio
immune response	2.77e-14	50/289
. . . .T-helper 17 cell differentiati	5.45e-05	3/3
. . . . .positive regulation of natural	4.24e-06	7/16
.regulation of immune response	7.26e-07	16/81
inflammatory response	3.11e-09	39/252
cellular defense response	6.71e-07	13/49
G-protein coupled receptor sig	7.12e-07	28/161
cytokine-mediated signaling pa	2.39e-06	23/196
.chemokine-mediated signaling p	1.66e-05	8/22
negative regulation of viral g	3.33e-06	8/30
positive regulation of natural	2.58e-05	4/5
response to virus	5.88e-05	13/96
.defense response to virus	4.86e-08	20/120
cytolysis	8.07e-05	5/13
positive regulation of cell ad	9.08e-05	8/29

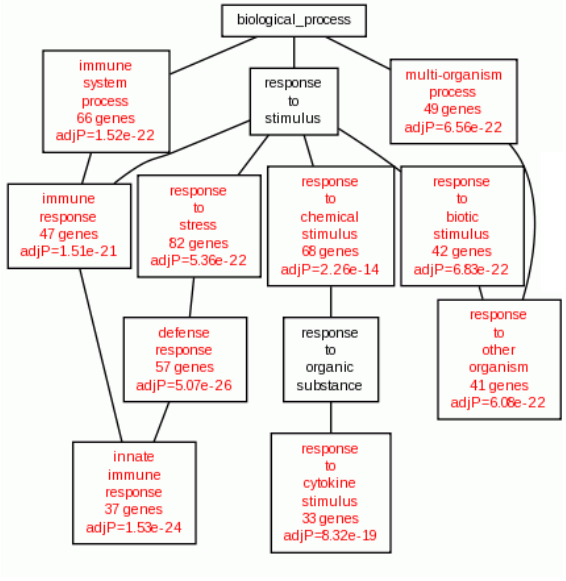
289 Genes in the Gene Universe are annotated with 'Immune Response'

50 Genes in Candidates List are belonging to 'Immune Response'

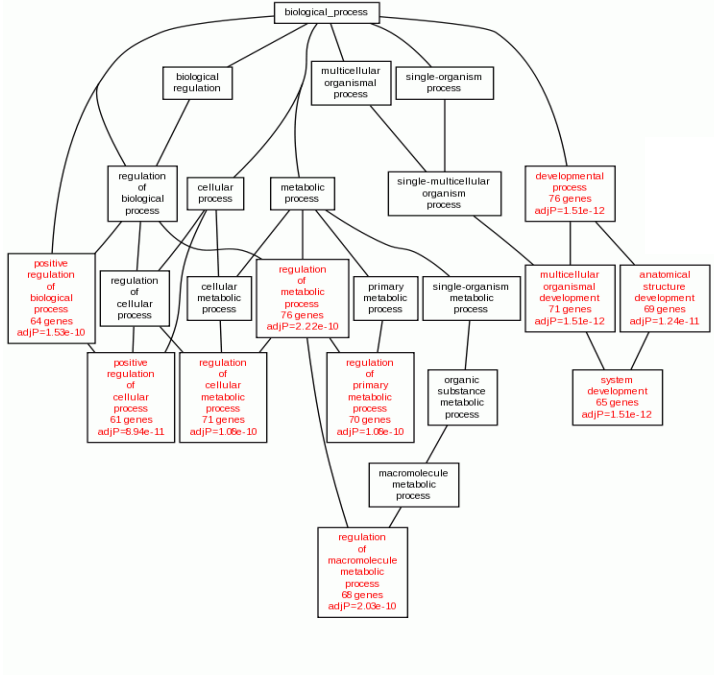
GO Enrichment - Example with WebGestalt

http://webgestalt.org

Up regulated genes

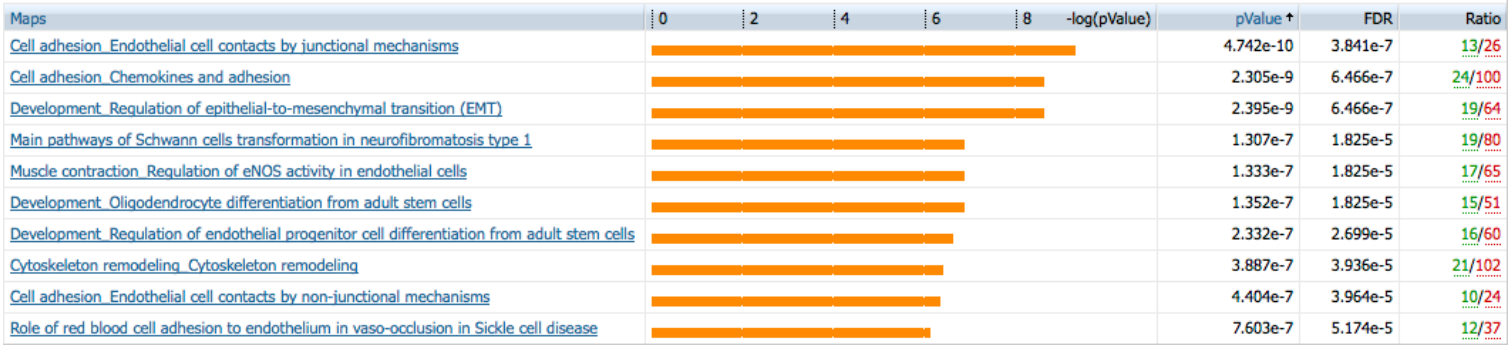


Down regulated genes



Pathway Enrichment - Example with Metacore

- https://portal.genego.com, commercial (FGCZ has a license)
- Upload a file with list of selected genes
- One click enrichment analysis (eg. pathway enrichment analysis)

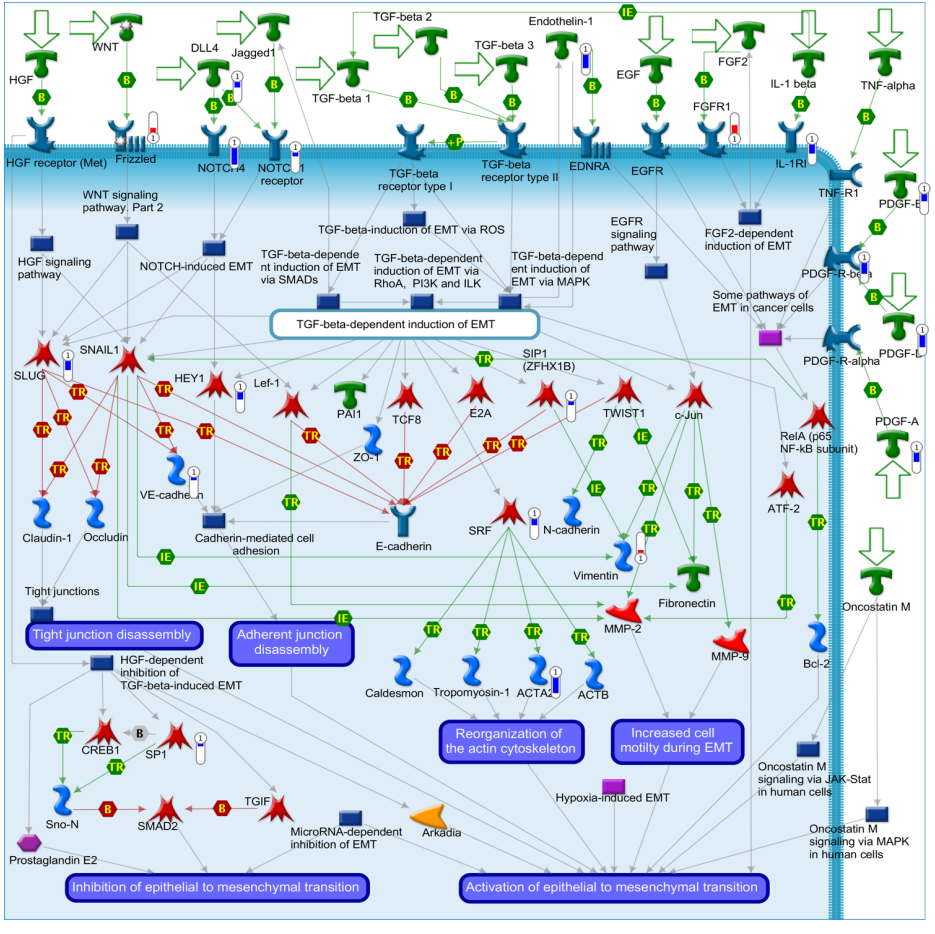


Maps	0 2 4 6 8 -log(pValue)	pValue	FDR	Ratio
Cell adhesion_Endothelial cell contacts by junctional mechanisms		4.742e-10	3.841e-7	13/26
Cell adhesion_Chemokines and adhesion		2.305e-9	6.466e-7	24/100
Development_Regulation of epithelial-to-mesenchymal transition (EMT)		2.395e-9	6.466e-7	19/64
Main pathways of Schwann cells transformation in neurofibromatosis type 1		1.307e-7	1.825e-5	19/80
Muscle contraction_Regulation of eNOS activity in endothelial cells		1.333e-7	1.825e-5	17/65
Development_Oligodendrocyte differentiation from adult stem cells		1.352e-7	1.825e-5	15/51
Development_Regulation of endothelial progenitor cell differentiation from adult stem cells		2.332e-7	2.699e-5	16/60
Cytoskeleton remodeling_Cytoskeleton remodeling		3.887e-7	3.936e-5	21/102
Cell adhesion_Endothelial cell contacts by non-junctional mechanisms		4.404e-7	3.964e-5	10/24
Role of red blood cell adhesion to endothelium in vaso-occlusion in Sickle cell disease		7.603e-7	5.174e-5	12/37

- By clicking on the pathway name, one can get a full picture of the genes involved in that pathway, with genes from the uploaded list specifically marked

(example on the next slide: Development regulation of EMT)

Pathway Enrichment - Example with Metacore



Summary

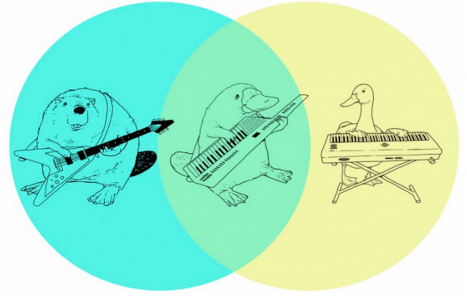
- Besides p value and/or fold change thresholds, clustering and venn diagram analysis can be used to filter DEGs
- GO terms are highly correlated, distributed across > 1,000 categories
- Functional analysis is only reliable for organisms with good annotation
  - GSEA of the full set of genes is more robust and versatile than functional overrepresentation analysis in selected/filtered set of genes

- Input for overrepresentation analysis
  - **Filtered** gene list based on p-value and/or fold change threshold
- Input for gene set enrichment analysis
  - **Unfiltered** gene list ranked based on p-value /or fold change

Additional Slides

Filtering list of DEGs: Venn diagram

- Idea
  - consider gene lists as sets and calculate intersection ($A \cap B$), union ($A \cup B$) and complement ($A \setminus B$, $B \setminus A$)
- Typical cases
  - Compare gene list to
    - data with similar context
    - public data
    - data from a different platform

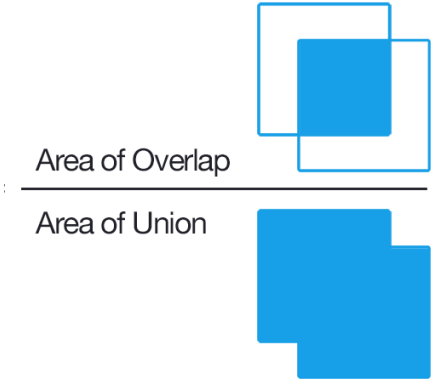


Jaccard index

- Jaccard index is a measure of the similarity between sets
- Given set A and set B, the Jaccard index is given by:

$$\frac{|A \cap B|}{|A \cup B|}$$

- Takes the values of 1 when sets are identical, and 0 when there is no overlap



FGCZ - Interactive Shiny Apps

http://fgcz-shiny.uzh.ch

ShinyApp	Description
Explore RNA-seq counts	Perform clustering and MDS plots; identify effect sizes and potential outliers
Explore RNA-seq differential expression	Filter and visualize your differential expression result; inspect individual genes; identify functional categories associated with gene lists
pcaExplorer	Visualization of RNA-seq data based on Pricipal Component analysis using pcaExplorer package
Heatmap	create an interactive gene expression heatmap based on a differential expression result generated by SUSHI
2-way-Venn	compare two sets of differential expression results to create a VennDiagram
3-way-Venn	compare three sets of differential expression results to create a VennDiagram
Correlation	interactive scatterplot for two sets of differential expression results

Which functional categories are really enriched for my set of genes?

Bonferroni count: 241

Export results View: Overlaid Area Chart of Difference

Displaying only results with P<0.05; click here to display all results

	Mus musculus (REF)	testForPanther.txt (Δ Hierarchy NEW!)				
PANTHER GO-Slim Biological Process	#	#	expected	Fold Enrichment	+/-	P value
sensory perception	1283	8	27.76	.29	-	1.66E-03
Unclassified	9053	146	195.89	.75	-	0.00E00
transport	1996	72	43.19	1.67	+	3.48E-03
↳localization	2177	80	47.11	1.70	+	5.36E-04
nervous system development	668	32	14.45	2.21	+	8.05E-03
↳system development	1084	50	23.46	2.13	+	1.41E-04
↳developmental process	2027	79	43.86	1.80	+	6.54E-05
cellular component movement	403	22	8.72	2.52	+	2.28E-02
↳cellular process	8700	230	188.25	1.22	+	1.66E-02
cellular component morphogenesis	529	29	11.45	2.53	+	1.70E-03
cell adhesion	486	30	10.52	2.85	+	1.13E-04
↳biological adhesion	486	30	10.52	2.85	+	1.13E-04
protein glycosylation	175	13	3.79	3.43	+	3.61E-02
muscle contraction	151	13	3.27	3.98	+	8.44E-03
polysaccharide metabolic process	167	15	3.61	4.15	+	1.32E-03

Note hierarchy to avoid redundancy

Promoter Analysis

- **Idea**
  - Identification and quantification of TFBS in promoter region of candidate genes
  - Compare results with background model (typically 200-500 housekeeping or randomly selected genes)
  - Compute ranking of TFs based on TFBS enrichment
- **Typical case**
  - Identify master regulatory transcription factors (potential marker discovery…)

Software tool

free tools (eg. Clover)
GeneXplain (uses *TRANSFAC*®-, *TRANSPATH*®-DB from BIOBASE)

Analog type of analysis

Compute a ranking of microRNAs
(based upon target gene expression; e.g. R-package 'MiRAGE')