STAT718/BIOL703: Genomic Data Science
Functional Analysis

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Objectives

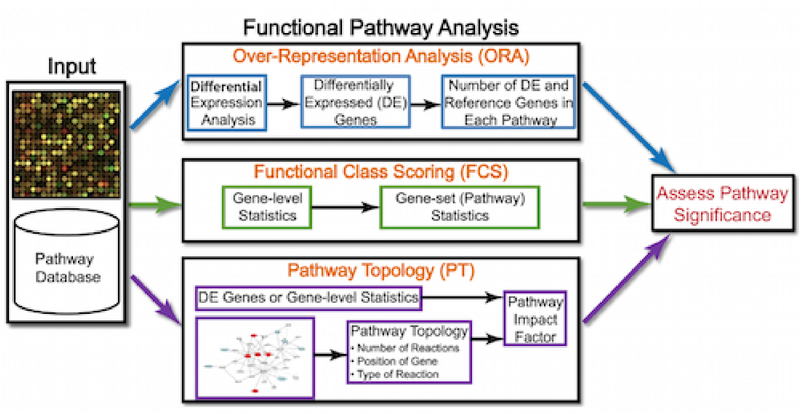
- Determine how functions are attributed to genes using Gene Ontology terms
- Understand the theory of how functional enrichment tools yield statistically enriched functions or interactions
- Discuss functional analysis using over-representation analysis, functional class scoring, and pathway topology methods
- Explore functional analysis tools

Functional Analysis

The output of RNA-seq differential expression analysis is a list of significant differentially expressed genes (DEGs). To gain greater biological insight on the differentially expressed genes there are various analyses that can be done:

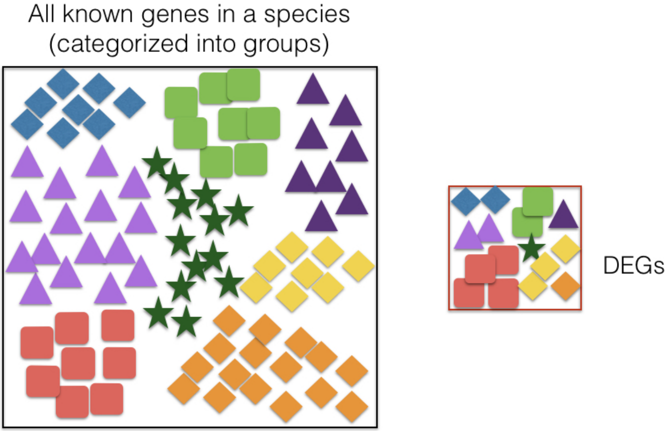
- Determine whether there is enrichment of known biological functions, interactions, or pathways
- Identify genes' involvement in novel pathways or networks by grouping genes together based on similar trends
- Use global changes in gene expression by visualizing all genes being significantly up- or down-regulated in the context of external interaction data

Pathway Analysis Tools


Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Over-representation Analysis

To determine whether any categories are over-represented, one can determine the probability of having the observed proportion of genes associated with a specific category in the gene list vs. the background (gene categorizations for the appropriate organism).



Example: Top Table

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
156	ABL1	1.10	9.20	9.03	4.88e-14	1.23e-10	21.29
1915	ABL1	1.15	9.00	8.59	3.88e-13	4.89e-10	19.34
155	ABL1	1.20	7.90	7.34	1.23e-10	1.03e-07	13.91
163	YES1	1.43	5.00	7.05	4.55e-10	2.87e-07	12.67
2066	PON2	1.18	4.24	6.66	2.57e-09	1.30e-06	11.03
2014	KLF9	1.78	8.62	6.39	8.62e-09	3.63e-06	9.89
1262	ALDH1A1	1.03	4.33	6.24	1.66e-08	6.00e-06	9.27
437	MARCKS	1.68	4.47	5.97	5.38e-08	1.70e-05	8.16
1269	AHNAK	1.35	8.44	5.81	1.10e-07	3.08e-05	7.49
1366	ANXA1	1.12	5.09	5.48	4.27e-07	1.08e-04	6.21

Enrichment Analysis

Is the selected set of genes enriched in the genes in the cell cycle pathway?

	Related to	Annotated but not	Not	Total
	Cell Cycle	Related to Cell Cycle	Annotated	
DE gene	100	691	9	800
Non DE gene	285	5012	65	5362
All gene	385	5703	74	6162

Enrichment Analysis

	Related to Cell Cycle	Annotated but not Related to Cell Cycle	Total
DE gene	100	691	791
Non DE gene	285	5012	5297
All gene	385	5703	6162

$$\frac{100}{791} = 12.64\% \qquad \frac{285}{5297} = 5.38\%$$

We can use the Hypergeometric test.

Hypergeometric Test

	Related to	Annotated but not	Total
	Cell Cycle	Related to Cell Cycle	
DE gene	100	691	791
Non DE gene	285	5012	5297
All gene	385	5703	6162

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{385}{100}\binom{5703}{691}}{\binom{6162}{791}}$$

This test will result in an adjusted p-value (after multiple test correction) for each category tested.

Biological Databases

- Gene Ontology (GO)
- Kyoto Encyclopedia of Genes and Genomes (KEGG)
- Ingenuity pathway analysis

GO Ontologies

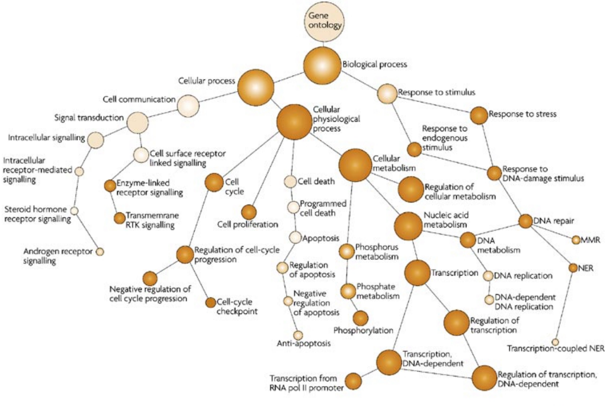
To describe the roles of genes and gene products, GO terms are organized into three independent controlled vocabularies (ontologies) in a species-independent manner:

- **Biological process**: refers to the biological role involving the gene or gene product, and could include "transcription", "signal transduction", and "apoptosis". A biological process generally involves a chemical or physical change of the starting material or input.
- **Molecular function**: represents the biochemical activity of the gene product, such activities could include "ligand", "GTPase", and "transporter".
- **Cellular component**: refers to the location in the cell of the gene product. Cellular components could include "nucleus", "lysosome", and "plasma membrane".

Each GO term has a term name (e.g. DNA repair) and a unique term accession number (GO:0005125).

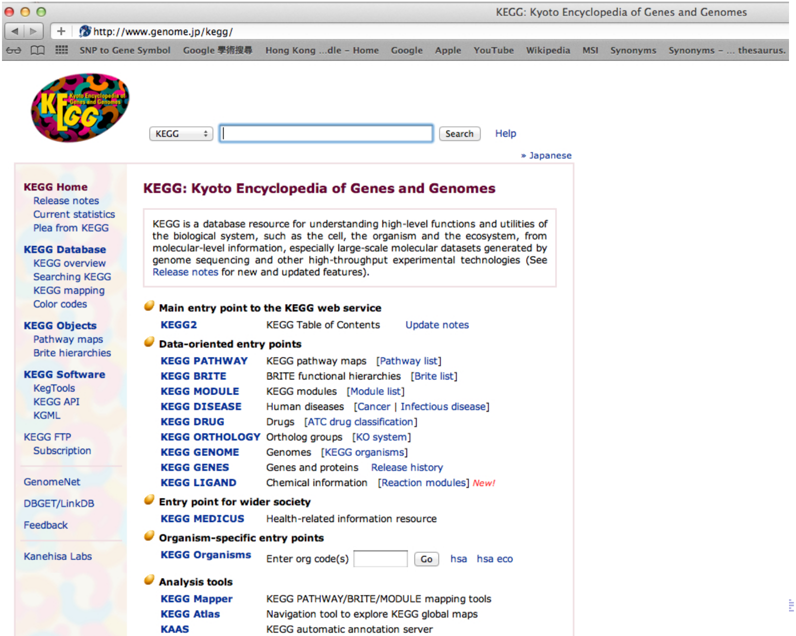
Go term hierarchy

To do this, GO ontologies are hierarchical, ranging from general, ‘parent’, terms to more specific, ‘child’ terms.


Gene ontology
Biological process
Cellular process
Response to stimulus
Cell communication
Response to stress
Signal transduction
Cellular physiological process
Response to endogenous stimulus
Intracellular signalling
Cell surface receptor linked signalling
Cellular metabolism
Response to DNA-damage stimulus
Intracellular receptor-mediated signalling
Enzyme-linked receptor signalling
Cell cycle
Cell death
Regulation of cellular metabolism
Programmed cell death
Transmemrane RTK signalling
Steroid hormone receptor signalling
Cell proliferation
Nucleic acid metabolism
DNA repair
Apoptosis
Androgen receptor signalling
Regulation of cell-cycle progression
Phosphorus metabolism
DNA metabolism
MMR
Regulation of apoptosis
Transcription
DNA replication
NER
Negative regulation of cell cycle progression
Phosphate metabolism
DNA-dependent DNA replication
Cell-cycle checkpoint
Negative regulation of apoptosis
Regulation of transcription
Phosphorylation
Anti-apoptosis
Transcription-coupled NER
Transcription, DNA-dependent
Transcription from RNA pol II promoter
Regulation of transcription, DNA-dependent

Nature Reviews | Cancer

KEGG pathway


KEGG: Kyoto Encyclopedia of Genes and Genomes
http://www.genome.jp/kegg/
SNP to Gene Symbol
Google 學術搜尋
Hong Kong ...dle – Home
Google
Apple
YouTube
Wikipedia
MSI
Synonyms
Synonyms – ... thesaurus.
Kyoto Encyclopedia of Genes and Genomes
KEGG
KEGG
Search
Help
» Japanese

KEGG Home
- Release notes
- Current statistics
- Plea from KEGG

KEGG Database
- KEGG overview
- Searching KEGG
- KEGG mapping
- Color codes

KEGG Objects
- Pathway maps
- Brite hierarchies

KEGG Software
- KegTools
- KEGG API
- KGML

KEGG FTP
- Subscription

GenomeNet

DBGET/LinkDB

Feedback

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See Release notes for new and updated features).

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps [Pathway list]
KEGG BRITE	BRITE functional hierarchies [Brite list]
KEGG MODULE	KEGG modules [Module list]
KEGG DISEASE	Human diseases [Cancer | Infectious disease]
KEGG DRUG	Drugs [ATC drug classification]
KEGG ORTHOLOGY	Ortholog groups [KO system]
KEGG GENOME	Genomes [KEGG organisms]
KEGG GENES	Genes and proteins Release history
KEGG LIGAND	Chemical information [Reaction modules] *New!*

Entry point for wider society

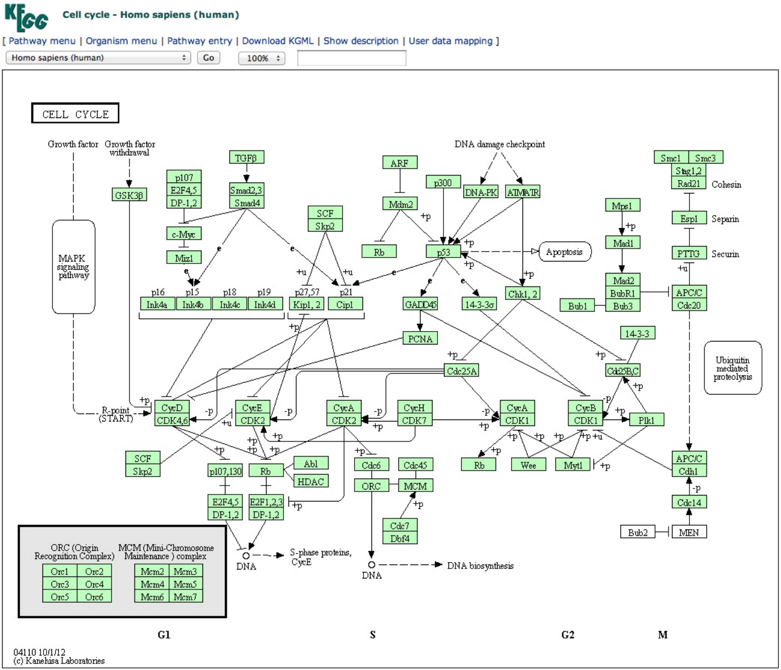
KEGG MEDICUS	Health-related information resource

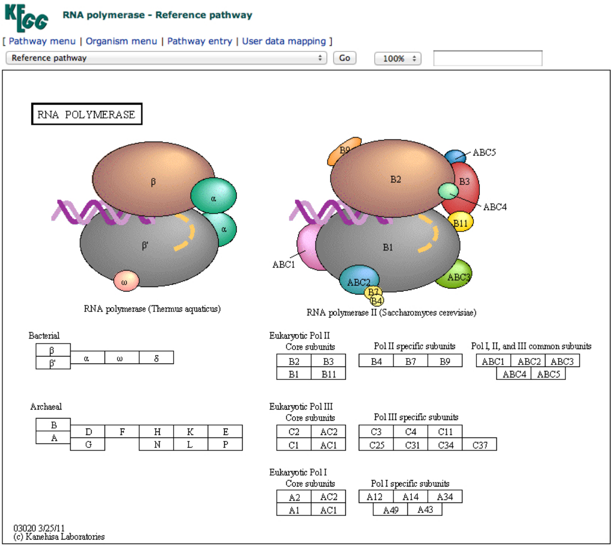
Organism-specific entry points

KEGG Organisms	Enter org code(s) Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
KEGG Atlas	Navigation tool to explore KEGG global maps
KAAS	KEGG automatic annotation server


Cell cycle - Homo sapiens (human)
[Pathway menu | Organism menu | Pathway entry | Download KGML | Show description | User data mapping]
Homo sapiens (human)
Go
100%
CELL CYCLE
Growth factor
Growth factor withdrawal
MAPK signaling pathway
GSK3β
p107
E2F4,5
DP-1,2
c-Myc
Miz1
TGFβ
Smad2,3
Smad4
SCF
Skp2
ARF
Mdm2
Rb
p300
DNA damage checkpoint
DNA-PK
ATM/ATR
p53
Apoptosis
GADD45
14-3-3σ
Chk1, 2
PCNA
Cdc25A
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
R-point (START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc6
ORC
Cdc45
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
Smc1
Smc3
Stag1,2
Rad21
Cohesin
Espl
Separin
PTTG
Securin
Mps1
Mad1
Mad2
BubR1
Bub3
Bub1
APC/C
Cdc20
14-3-3
Cdc25B,C
Ubiquitin mediated proteolysis
APC/C
Cdh1
Cdc14
Bub2
MEN
ORC (Origin Recognition Complex)
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
+p
-p
+u
e
G1
S
G2
M
04110 10/1/12
(c) Kanehisa Laboratories



RNA polymerase - Reference pathway

[Pathway menu | Organism menu | Pathway entry | User data mapping]

Reference pathway | Go | 100%

RNA POLYMERASE

RNA polymerase (Thermus aquaticus)

RNA polymerase II (Saccharomyces cerevisiae)

Bacterial

β	α	ω	δ
β'			

Archaeal

B	D	F	H	K	E
A	G		N	L	P

Eukaryotic Pol II

Core subunits

B2	B3
B1	B11

Pol II specific subunits

B4	B7	B9

Pol I, II, and III common subunits

ABC1	ABC2	ABC3
ABC4	ABC5	

Eukaryotic Pol III

Core subunits

C2	AC2
C1	AC1

Pol III specific subunits

C3	C4	C11	
C25	C31	C34	C37

Eukaryotic Pol I

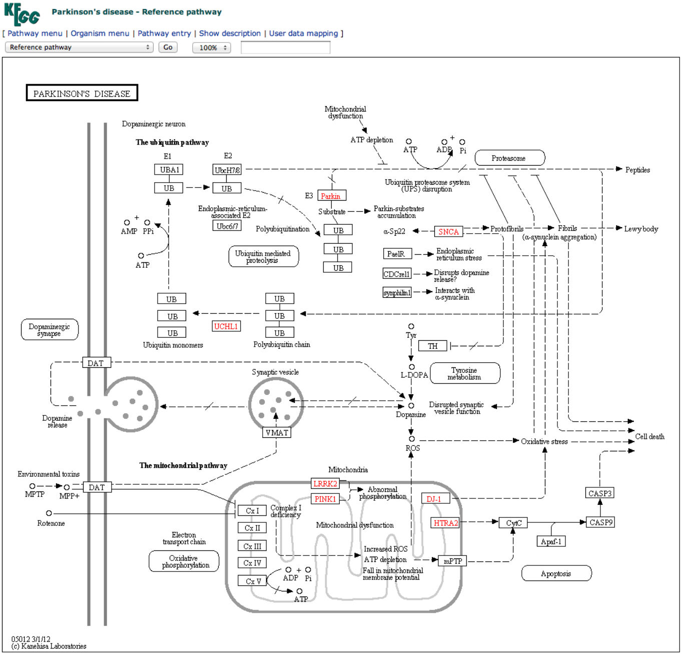
Core subunits

A2	AC2
A1	AC1

Pol I specific subunits

A12	A14	A34
A49	A43	

03020 3/25/11
(c) Kanehisa Laboratories



Parkinson's disease - Reference pathway
[Pathway menu | Organism menu | Pathway entry | Show description | User data mapping]
Reference pathway
Go
100%
PARKINSON'S DISEASE
Dopaminergic neuron
The ubiquitin pathway
E1
E2
UBA1
UbcH7/8
UB
UB
AMP
PPi
ATP
Endoplasmic-reticulum-associated E2
Ubc6/7
Polyubiquitination
Ubiquitin mediated proteolysis
E3
Parkin
Substrate
Parkin-substrates accumulation
UB
UB
UB
Mitochondrial dysfunction
ATP depletion
ATP
ADP
Pi
Proteasome
Peptides
Ubiquitin proteasome system (UPS) disruption
α-Sp22
SNCA
Protofibrils
Fibrils
(α-synuclein aggregation)
Lewy body
PaelR
Endoplasmic reticulum stress
CDCrel1
Disrupts dopamine release?
synphilin1
Interacts with α-synuclein
UB
UB
UB
Ubiquitin monomers
UCHL1
UB
UB
UB
Polyubiquitin chain
Tyr
TH
L-DOPA
Tyrosine metabolism
Dopaminergic synapse
DAT
Synaptic vesicle
Dopamine release
VMAT
Dopamine
Disrupted synaptic vesicle function
ROS
Oxidative stress
Cell death
The mitochondrial pathway
Mitochondria
Environmental toxins
MPTP
MPP+
DAT
Rotenone
LRRK2
PINK1
Abnormal phosphorylation
DJ-1
Cx I
Complex I deficiency
Cx II
Cx III
Cx IV
Cx V
ADP
Pi
ATP
Electron transport chain
Oxidative phosphorylation
Mitochondrial dysfunction
Increased ROS
ATP depletion
Fall in mitochondrial membrane potential
HTRA2
mPTP
CytC
Apaf-1
CASP9
CASP3
Apoptosis
05012 3/1/12
(c) Kanehisa Laboratories

KEGG database

- kg.hsa$sigmet.idx: signaling and metabolism gene sets
- kg.hsa$dise.idx: disease gene sets

gage

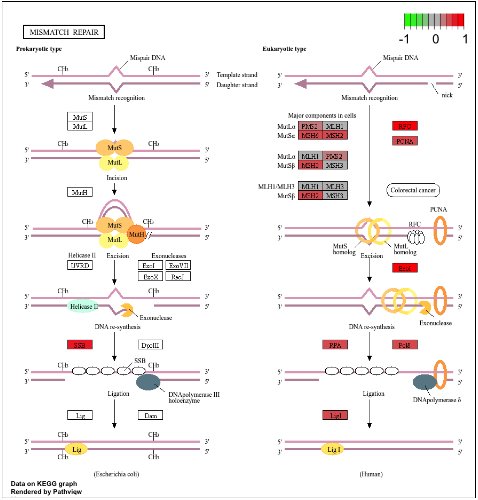
Generally applicable gene-set enrichment (gage) is a popular bioconductor package for performing gene-set and pathway analysis.

gage Results

	p.geomean	stat.mean	p.val	q.val	set.size
hsa04110 Cell cycle	1.05e-08	5.77	1.05e-08	2.52e-06	158
hsa03013 Nucleocytoplasmic transport	4.67e-07	5.14	4.67e-07	5.61e-05	107
hsa04657 IL-17 signaling pathway	8.85e-06	4.44	8.85e-06	6.00e-04	91
hsa03008 Ribosome biogenesis in eukaryotes	1.00e-05	4.44	1.00e-05	6.00e-04	83
hsa03010 Ribosome	2.29e-05	4.18	2.29e-05	1.10e-03	134

Code for analysis can be found in FunctionalAnalysis.R on the course website.

Visualizing gage Results: KEGG



Visualizing gage Results

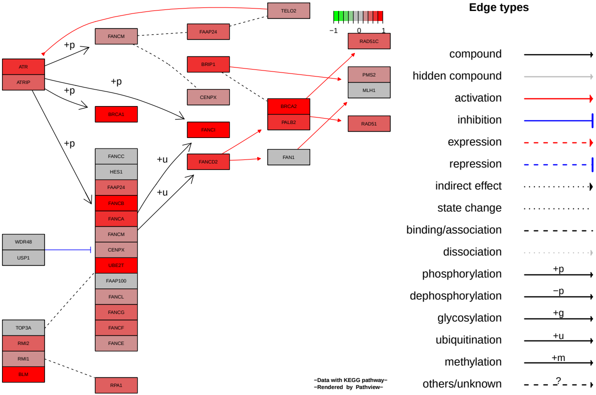


Figure: Fanconi anemia pathway

Gene Set Enrichment Analysis (GSEA)

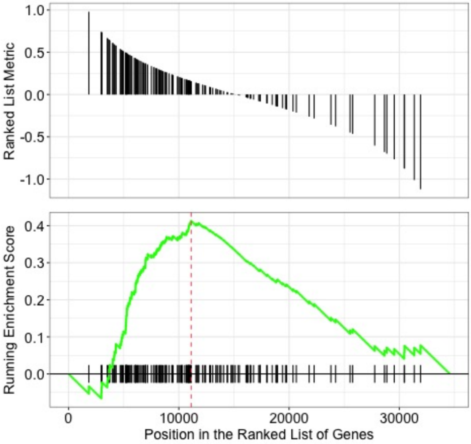


Figure: GSEA plot for the Spliceosome pathway: hsa03040

GSEA output

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adj
1	hsa04657	IL-17 signaling pathway	91	0.710	2.771	0.004	0.0
2	hsa03013	Nucleocytoplasmic transport	107	0.582	2.380	0.004	0.0
3	hsa03030	DNA replication	36	0.684	2.289	0.003	0.0
4	hsa03460	Fanconi anemia pathway	54	0.628	2.255	0.004	0.0
5	hsa03008	Ribosome biogenesis in eukaryotes	83	0.559	2.178	0.004	0.0
6	hsa05204	Chemical carcinogenesis - DNA adducts	63	-0.652	-2.175	0.001	0.0
7	hsa05323	Rheumatoid arthritis	86	0.549	2.157	0.004	0.0
8	hsa00982	Drug metabolism - cytochrome P450	64	-0.634	-2.123	0.001	0.0
9	hsa04976	Bile secretion	85	-0.565	-1.999	0.001	0.0
10	hsa05340	Primary immunodeficiency	37	-0.660	-1.997	0.001	0.0

Gene Set Enrichment Analysis (GSEA)

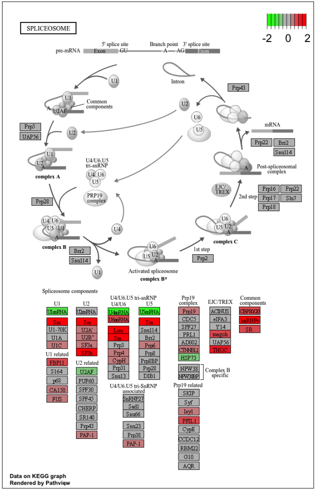
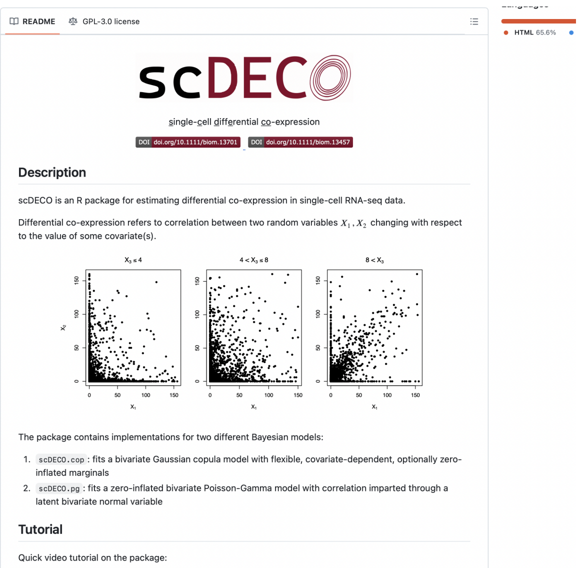


Figure: Pathview plot for the Spliceosome pathway: hsa03040

Differential Co-Expression Analysis



single-cell differential co-expression

Description

scDECO is an R package for estimating differential co-expression in single-cell RNA-seq data.

Differential co-expression refers to correlation between two random variables X_1, X_2 changing with respect to the value of some covariate(s).

The package contains implementations for two different Bayesian models:

1. `scDECO.cop` : fits a bivariate Gaussian copula model with flexible, covariate-dependent, optionally zero-inflated marginals
2. `scDECO.pg` : fits a zero-inflated bivariate Poisson-Gamma model with correlation imparted through a latent bivariate normal variable

Tutorial

Quick video tutorial on the package:

Developed by Yen-Yi Ho Lab

Resources for functional analysis

- g: Profiler - `http://biit.cs.ut.ee/gprofiler/index.cgi`
- David - `http://david.abcc.ncifcrf.gov/tools.jsp`
- clusterProfiler - `http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html`
- GENEMANIA - `http://www.genemania.org/`
- GenePattern - `http://www.broadinstitute.org/cancer/software/genepattern/` (need to register)
- WebGestalt - `http://bioinfo.vanderbilt.edu/webgestalt/` (need to register)
- AmiGO - `http://amigo.geneontology.org/amigo`
- REviGo (visualizing GO analysis, input is GO terms) - `http://revigo.irb.hr/`
- scDECO - `https://github.com/YenYiHo-Lab/scDECO`
- GSEA - `http://software.broadinstitute.org/gsea/index.jsp`
- SPIA - `https://www.bioconductor.org/packages/release/bioc/html/SPIA.html`
- GAGE/Pathview - `http://www.bioconductor.org/packages/release/bioc/html/gage.html`