

Gene Set Enrichment Analysis

Typical Analysis

- Compare two different conditions
- Produce lists with up and down regulated genes
- Use Panther or GO annotation to look for over-representation of functional groups within the lists

Problem

- Which cut-off should we use?
 - Fold change
 - P-values
- If small changes are observed, few single genes may come out with significant change using SAM or t-test.

Gene Set Enrichment Analysis

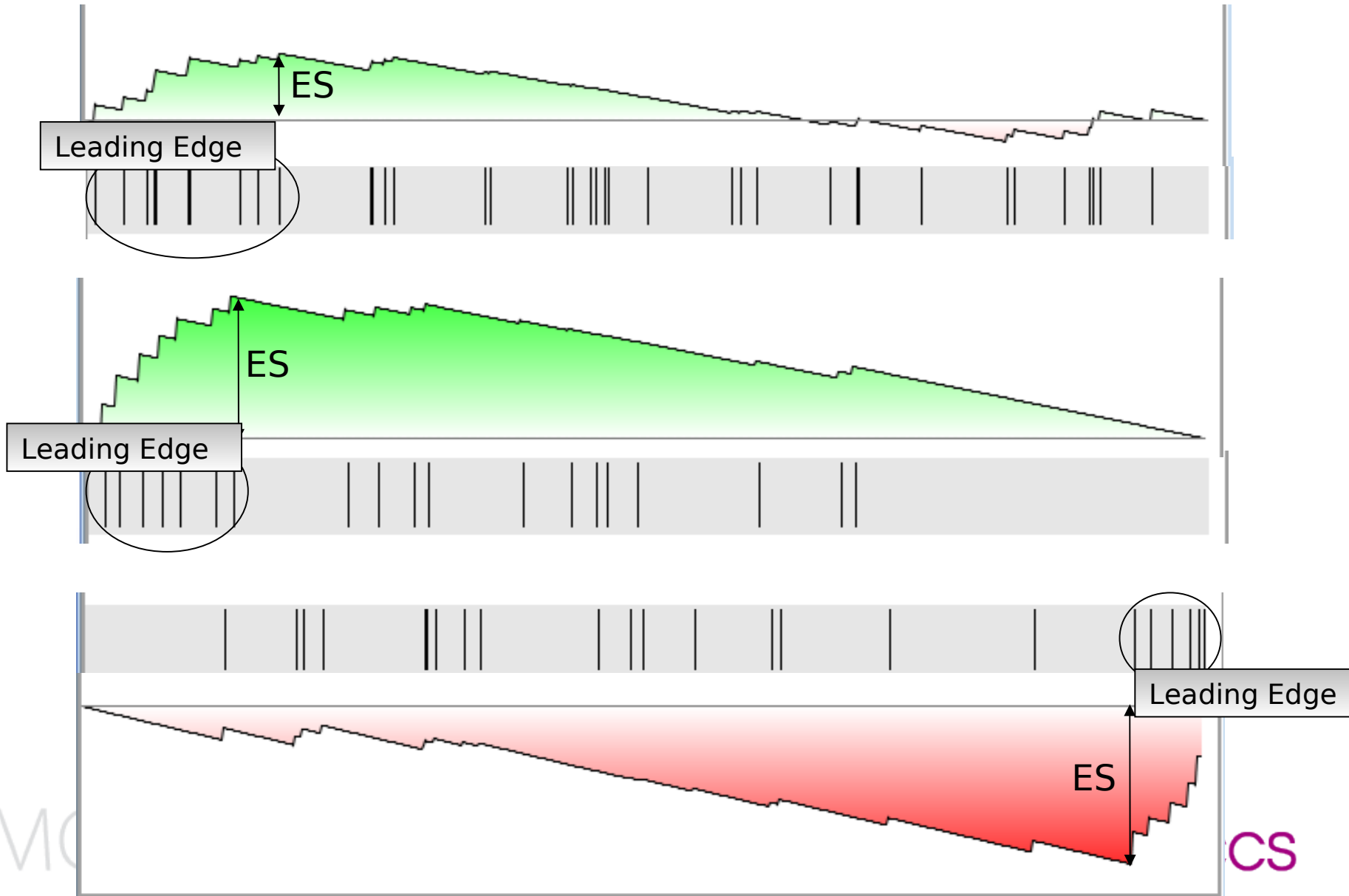
(Subramanian et al, Broad Institute)

- Genes are related and definitely not acting independantly of eachother.
 - Use known relations in analysis of experimental results
 - More true to biologic reality
 - Easier to interpret / show trends
 - Gives stronger statistics
- Gene Sets
 - Apriori knowledge
 - Use existing resources: Publications, candidate lists, public experimental data etc...

Example of .gmt file

CHESLER_D6MIT150_TR	na	BACH1	MSX3	GATA6	IMMT	CDX4	DLX4	HOXB6	TIMD2
DEATHPATHWAY	na	BID	TRAF2	TNFRSF25	NFKBIA	NFKB1	TNFSF12	CASP6	CASP3
MRNA_PROCESSING	na	FUSIP1	U2AF2	SNRPB2	HNRPR	SF3B4	SF3B3	SMNDC1	NONO
LE_MYELIN_DN	na	SNCG	IL16	HMGCR	UTRN	SNCA	NDRL	FDFT1	OGN
TCAPOPTOSISPATHWAY	na	TNFSF6	CD3G	CD3D	CCR5	CD3E	CD4	TNFRSF6	TRA@
MRNA_PROCESSING_RE	na	HNRPAB	NCBP2	NCBP1	PRPF4B	RNMT	U2AF2	SNRPD3	HNRPD
ATP_SYNTHESIS	na	ATP5E	ATP6V0C_///	ATP6AP1	ATP6V1H	ATP6V1G2	ATP6V1G1	ATP6V1B2	ATP6V1B1
ROSS_CBF	na	ADCY7	CD52	PMAIP1	HLA-DMB	SYNGR1	CBFB	ISG20	RCBTB1
AGUIRRE_PANCREAS_C	na	HNRPA1 /// L	TBK1	PRKAG1	LEMD3	CCT2	TEGT	OR7E47P	ATP2B1
CCR3PATHWAY	na	PRKCA	HRAS	PIK3C2G	MAP2K1	MYL2	ARHA	LIMK1	ROCK2
NEUTROPHILPATHWAY	na	ITGAL	ICAM1	CD44	SELL	PECAM1	ITGB2	SELE	ITGAM
INOSITOL_METABOLISM	na	ALDOA	ALDH6A1	TPI1	ALDOC	ALDOB			
TRYPTOPHAN_METABO	na	CYP3A4	CYP3A5	KYNU	CYP3A7	CYP2J2	CYP2C19	CYP2C18	EHHADH
ELECTRON_TRANSPORT	na	CYP24A1	PGD	ADH1C	ADH1B	ADH1A	COX5A	HNRPM	INDO
ACETYLCHOLINE_SYNT	na	CHKA	ACHE	PEMT	SLC18A3	PDHA2	PCYT1A	PDHA1	CHAT
ALTERNATIVEPATHWAY	na	BF	C8A	C7	C9	PFC	C3	C6	C5
BYSTRYKH_SCP2_CIS_T	na	KIF1C	PSMB6	6330403K07F	DYNLL2	CCL9	LIG3	MPO	A1451896
PGC1APATHWAY	na	MEF2C	SYT1	PPARA	MEF2B	ESRRA	MEF2A	CAMK1G	CAMK2G
ST_INTERFERON_GAMM	na	STATIP1	REG1A	IFNG	PLA2G2A	JAK1	JAK2	PTPRU	STAT1
HUMAN_MITODB_6_2002	na	HCCS	OXA1L	SLC9A6	RPLP2	OGDH	PDHA1	OKL38	ACAA2
CITRATE_CYCLE_TCA_C	na	ACO2	ACO1	SUCLG2	SUCLG1	DLST /// DLS	CS	IDH3B	IDH3A
ST_WNT_CA2_CYCLIC_G	na	TF	CAMK2G	SLC6A13	DAG1	ITPKB	ITPR3	PDE6H	PDE6G
BIOSYNTHESIS_OF_STE	na	MVD	HMGCR	FDPS /// LOC	FDPS	LSS	PMVK	FDFT1	SQLE
KENNY_WNT_UP	na	SEPHS2	S100A8	GNA11	BTRC	PTPN22	MMP2	SFRS7	EBNA1BP2
TYPE_III_SECRETION_S	na	ATP5E	ATP6AP1	ATP6V1H	ATP6V1G2	ATP6V1G1	ATP6V1B2	ATP6V1B1	ATP6V1D
TGF_BETA_SIGNALING	na	HRAS	NOG	LTBP1	FST	NFKB1	TGFB1	CTNNA3	FOS
PMLPATHWAY	na	TNFSF6	HRAS	TNF	SP100	CREBBP	PML	TP53	PRAM-1

Distribution of a gene set

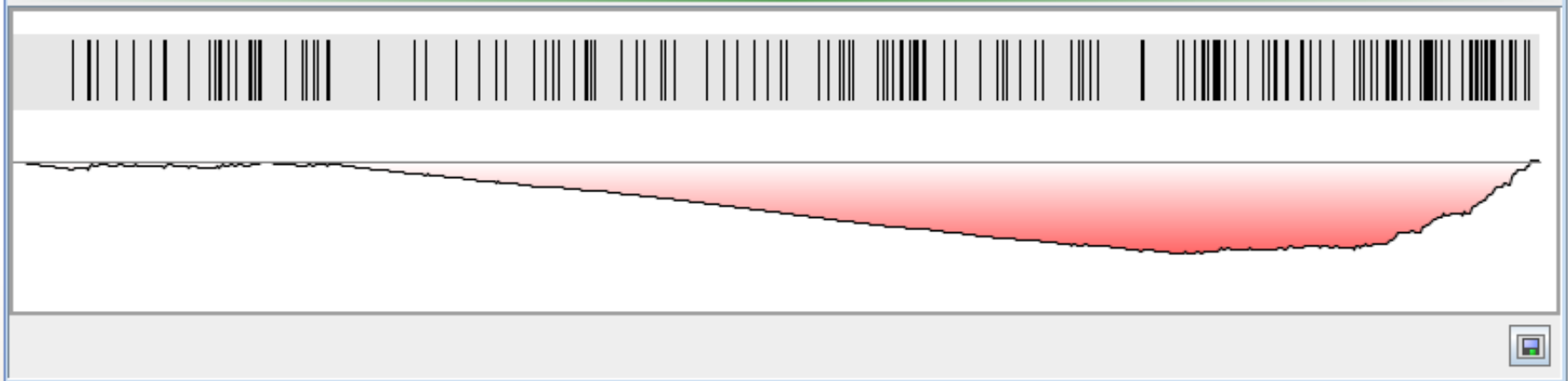


File

Enriched in Cortex

Enriched in Hippocampus

Rank	Gene Set	Size	ES	NES	Nom P-value	FDR (%)
1	Heart development	47	-0.55	-2.01	0.0	0.34
2	Amino acid catabolism	40	-0.59	-1.91	0.0	1.61
3	Fatty acid metabolism	172	-0.45	-1.87	0.0	1.99
4	Other neuronal activity	125	-0.4	-1.84	0.0	1.84
5	Acyl-CoA metabolism	18	-0.7	-1.81	0.0	2.46
6	Amino acid metabolism	212	-0.4	-1.73	0.0	6.12
7	Lysosome transport	12	-0.61	-1.72	0.01	5.64
8	Coenzyme metabolism	53	-0.48	-1.72	0.0	5.09



Make Selection

Leading Edge

All

Result

 Store result in project tree

 Branch selected

Back

Chart

BW

LEA

Leading-Edge Subset

- Examination of the leading-edge subset can reveal a biologically important subset within a gene set
- High scoring gene sets can be grouped on the basis of leading edge subsets of genes that they share.
 - This can reveal which gene sets belong to the same biological processes and which represent distinct processes

Statistical significance

- T-test – use theoretical t-distribution
- SAM – sample permutations
- Rank Product – gene permutations
- GSEA – sample permutations or gene permutations

How do we get significance values?

- If the distribution of scores is known we can read p-values from a precalculated table
- If the distribution is unknown we must estimate it
 - Randomise data (referred to as permutation or resampling), repeat test and compare result to the original one.
 - Repeat permutation a number of times and count the number of times we get a better result than the original one.

Permutations

- Sample permutations means that we randomly swap samples between the different sample groups
- Gene permutations means that we randomly swap the gene profiles between the gene sets

Gene permutations

- Destroys gene – gene correlations
- Underestimate score for a gene set if there is a positive correlation between the genes in the gene set

Sample permutation

- Keeps gene gene correlations
- Few samples means that we underestimate the variance of the scores calculated on the permuted data

Sources of gene sets

- Molecular Signature Database
 - Cytogenetic sets
 - Functional sets
 - Regulatory motif sets
 - Neighbourhood sets
- Gene Ontology
- PantherDB
- Define your own

Gene set considerations I

- Gene sets need to be a priori defined
 - Not derived from current data set
 - Not gradually adjusted/tested to give nice results

Gene set considerations II

- The null hypothesis for each gene set is that it is not differentially expressed
- The number of gene sets tested will affect the FDR values
- It is important to decide before looking at the data which sets of gene sets should be tested, e.g all GO terms, KEGG etc, or only testing certain gene sets that you believe are important for the study

Example 1 – GO terms

The screenshot displays the GeneSet Enrichment Analysis (GSEA) interface. The main window shows a list of enriched GO terms, with 'melanosome' selected. The 'Enriched Up-regulated in ALL' tab is active, showing a table of results. Below the table is a plot showing the enrichment score (ES) curve for the selected term. The 'Result' section includes buttons for 'Store result in project tree', 'Branch selected', and 'Back'. The 'Chart' section has buttons for 'BW' and 'LEA'. The 'Make Selection' section has buttons for 'All' and 'Leading Edge'. The 'Description' section provides a detailed description of the selected GO term.

GO DAG PR log 2 ratios (per pair) uniq proteins

- GARP complex 0 0 0
- pigment granule 0 25 25
 - melanosome 25 25 0**
 - cyanosome 0 0 0
 - pterinosome 0 0 0
 - leucosome 0 0 0
 - iridosome 0 0 0
 - axoneme 0 0 0
 - attachment organelle 0 0 0

Name: melanosome
GO: GO:0042470

Selection:

- Automatic Selection Update
- Recursive Selection
- Maximum Members: 100
- Use Synonyms

Rank	Gene Set	Size	ES	NES	Nom P-...	FDR (%)
1	melanosome	25	-0.72	-2.08	0.0	0.42
2	pigment granule	25	-0.72	-2.08	0.0	0.21
3	calcium-dependent phosph...	10	-0.8	-1.86	0.0	11.87
4	endoplasmic reticulum lumen	15	-0.73	-1.85	0.0	9.54

Result:

- Store result in project tree
- Branch selected
- Back

Chart:

- BW
- LEA

Make Selection:

- All
- Leading Edge

Description:

"A tissue-specific, membrane bounded cytoplasmic organelle within which melanin pigments are synthesized and stored. Melanosomes are synthesized in melanocyte cells." [GOC:j], PMID:11584301]

Example 2 – public data sets

Combined_GEO_gene_sets_010409.gmt - OpenOffice.org Calc

File Edit View Insert Format Tools Data Window Help

GeneSet Enrichment Analysis

Enriched in Group 1 Enriched in Group 2

Rank	Gene Set	Size	ES	NES	No...	FDR...
1	Inflammation_Bronch Epithelium IL13_up	51	-0.52	-2.06	0.0	0.37
2	Hypoxia_HeLa_up	37	-0.49	-1.84	0.0	4.15
3	Angiogenesis_HMEC_tube formation_down	31	-0.5	-1.82	0.01	3.35
4	Epigenetics_Hepatoma TSA up	35	-0.49	-1.8	0.01	2.88

Result

Store result in project tree

Branch selected

Back

Chart

Make Selection

BW

LEA

All

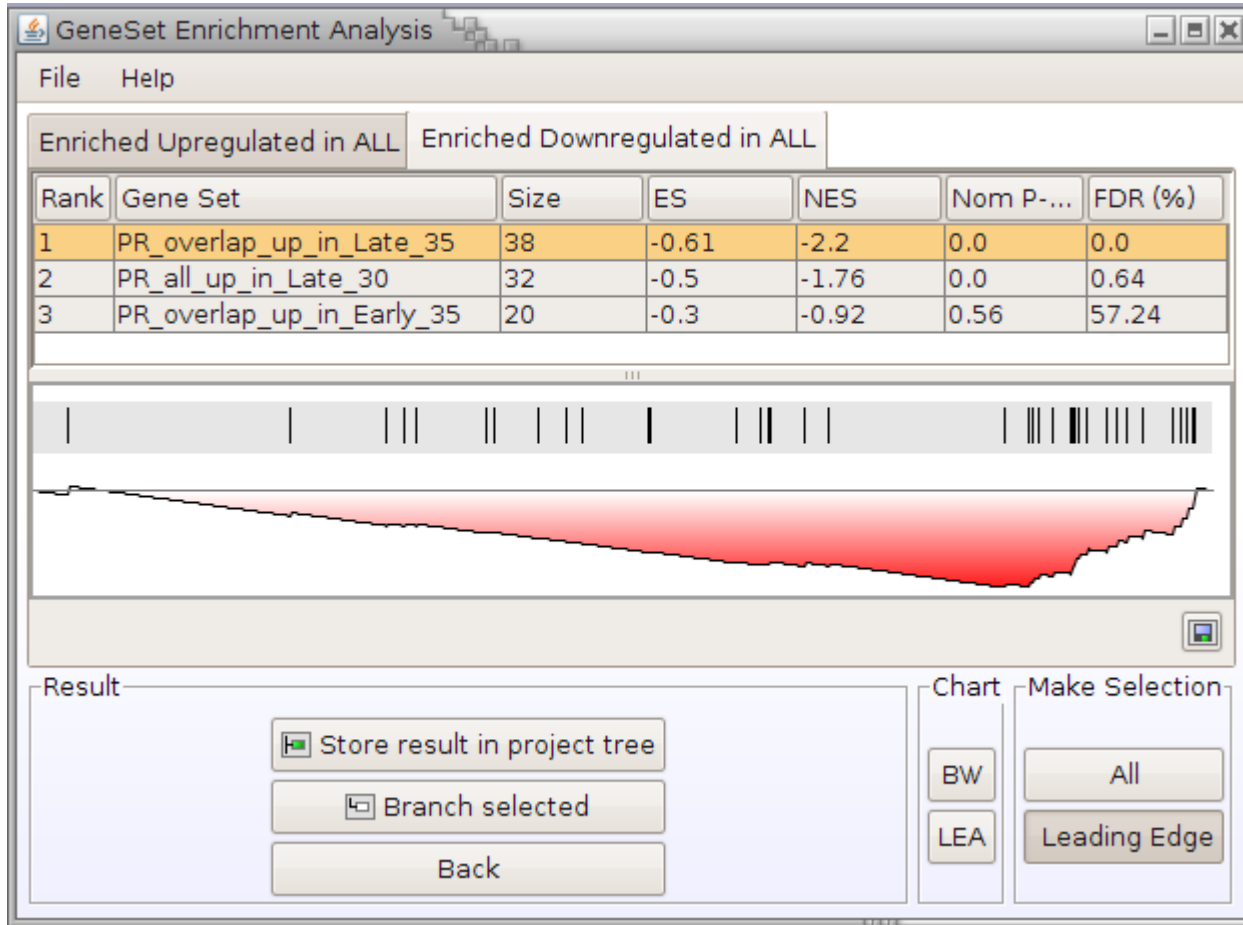
Leading Edge

24	Cytokine_HU								
25	Cytokine_HM								
26	Cytokine_HM								
27	Epigenetics_C								
28	Epigenetics_C								
29	Epigenetics_H								
30	Epigenetics_H								
31	Epigenetics_H								
32	Epigenetics_H								
33	Epigenetics_H								
34	Epigenetics_H								
35	Epigenetics_M								
36	Epigenetics_M								
37	Epigenetics_H								
38	Epigenetics_H								
39	Glucose_Skel								
40	Glucose_Skel								
41	Glucose_myo								
42	Glucose_myo								
43	Hypoxia_mac								
44	Hypoxia_macrophage_down	CCT6A	PLAU	PLEKHO2	TFRC				
45	Hypoxia_lymphaticEC_up	AA825563	AI300520	AI333596	AI655611	AI739132	AI873273	AK3L1	

Sheet1

Sheet 1 / 1 Default 100% STD * Sum=0

Example 3 – MA and PR data



Questions?