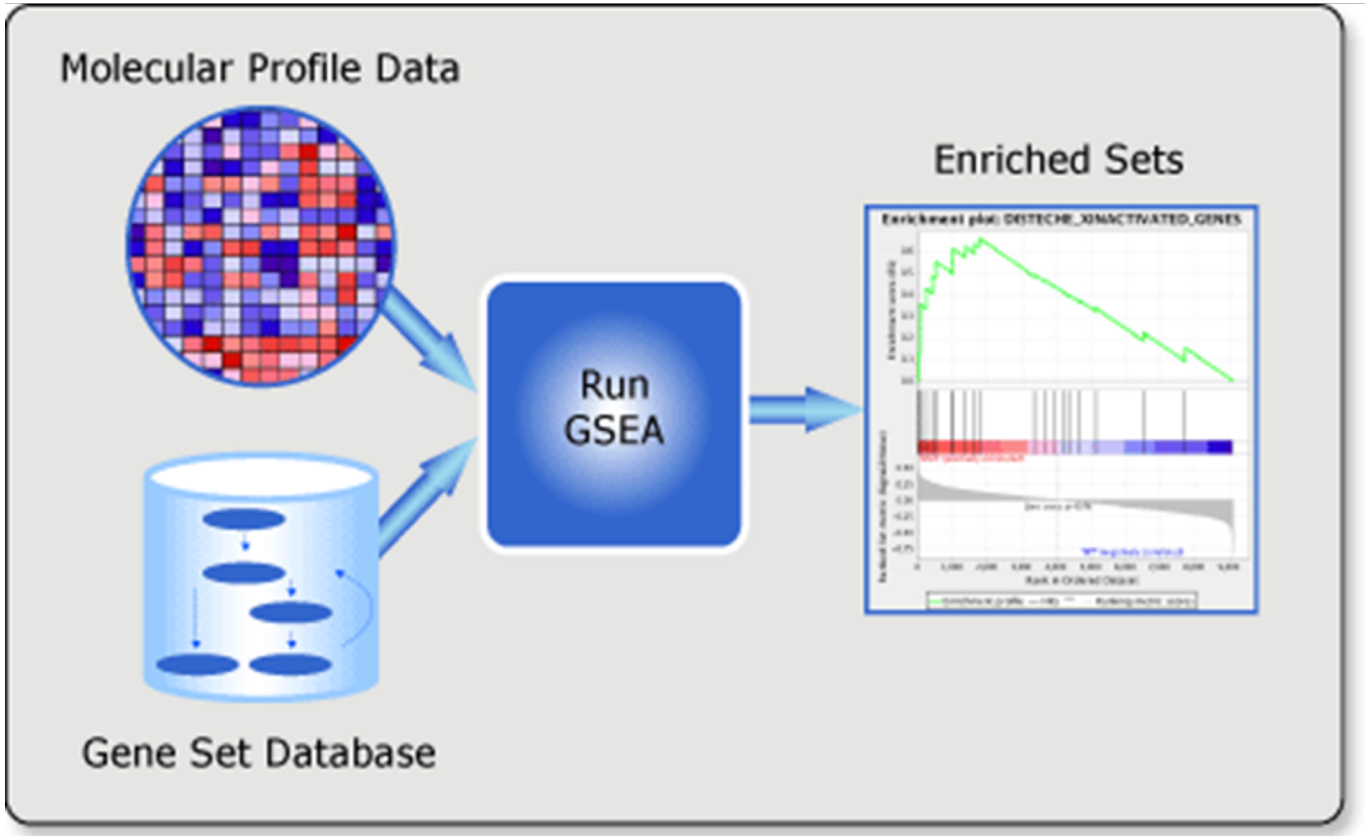
A tutorial of GSEA

NTU Center of Genomic and Precision Medicine

Yi-Wen Hsiao

Date: 2017.09.22

The Concept of GSEA



Gene Set:
(1) Molecular Signatures Database (MSigDB)
(2) Manually created gene set

Step 1: Understanding the input formats

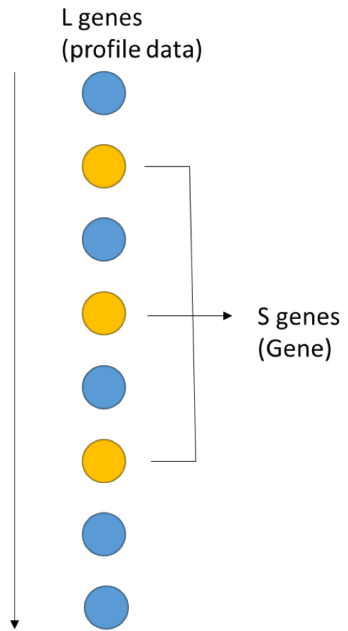
Step 2: Gene Set selection

Step 3: Sorting dataset

Step 4: Calculation of Enrichment Score (ES)

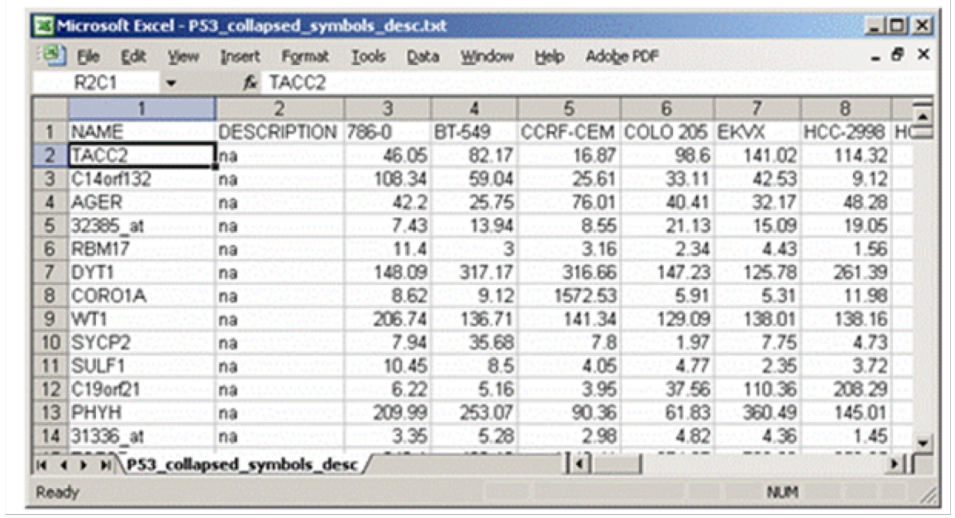
Step 5: Calculation of significant difference

Step 6: Calculation of multiple comparisons



Step 1: Understanding the input formats

Input1_Expression Dataset(*.txt)



	1	2	3	4	5	6	7	8
1	NAME	DESCRIPTION	786-0	BT-549	CCRF-CEM	COLO 205	EKVX	HCC-2998
2	TACC2	na	46.05	82.17	16.87	98.6	141.02	114.32
3	C14orf132	na	108.34	59.04	25.61	33.11	42.53	9.12
4	AGER	na	42.2	25.75	76.01	40.41	32.17	48.28
5	32385_at	na	7.43	13.94	8.55	21.13	15.09	19.05
6	RBM17	na	11.4	3	3.16	2.34	4.43	1.56
7	DYT1	na	148.09	317.17	316.66	147.23	125.78	261.39
8	CORO1A	na	8.62	9.12	1572.53	5.91	5.31	11.98
9	WT1	na	206.74	136.71	141.34	129.09	138.01	138.16
10	SYCP2	na	7.94	35.68	7.8	1.97	7.75	4.73
11	SULF1	na	10.45	8.5	4.05	4.77	2.35	3.72
12	C19orf21	na	6.22	5.16	3.95	37.56	110.36	208.29
13	PHYH	na	209.99	253.07	90.36	61.83	360.49	145.01
14	31336_at	na	3.35	5.28	2.98	4.82	4.36	1.45

Input2_Geneset (*.gmt/*.gmx)

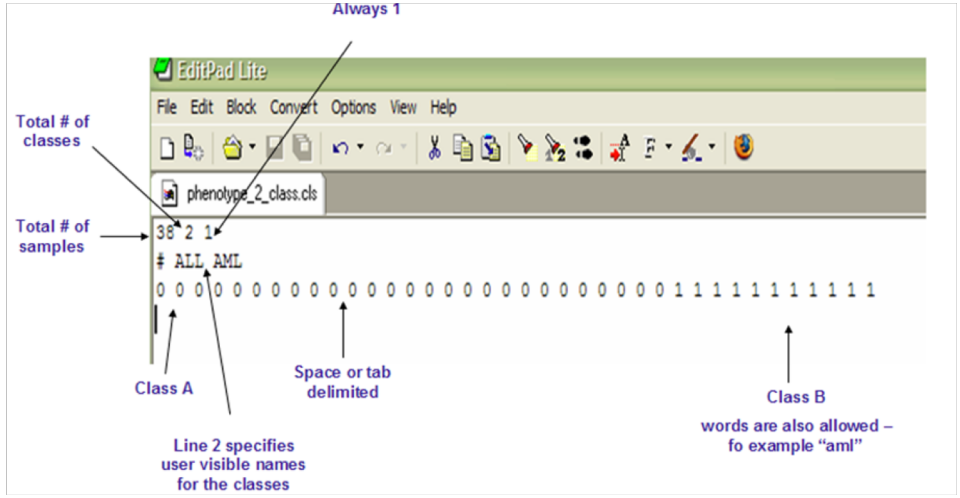
(*.gmt)

	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUFB8
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	IRF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	RECQL4
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	COTL1
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf9
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOSTDC1

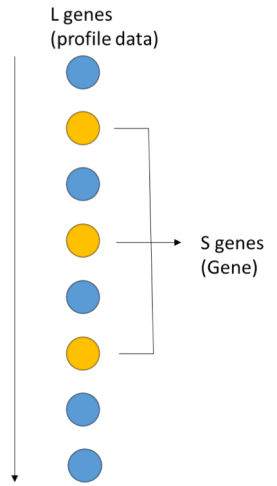
(*.gmx)

	A	B	C
1	chr10q24	chr5q23	chr8q24
2	na	na	na
3	PITX3	ALDH7A1	HAS2
4	SPFH1	IL13	LRRC14
5	NEURL	8-Sep	TSTA3
6	C10orf12	IRF1	DGAT1
7	NDUFB8	ACSL6	RECQL4
8	C10orf95	IL4	GPR172A
9	DNTT	SLC12A2	COL14A1
10	USMG5	PPIC	EXT1
11	CWF19L1	CSF2	RAD21
12	SUFU	SLC22A5	SLA
13	OBFC1	CSNK1G3	
14	PEO1	DMXL1	
15	PIK3AP1	P4HA2	
16	UBTD1	ZNF608	
17	CUTC	LOX	
18	SEC31L2	FTMT	
19	AS3MT	ADAMTS19	
20	MGEA5	IL5	
21	NPM3	IL3	

Input3_Phenotype Labels (.cls)



Step 2: Gene Set selection



(1) Molecular Signatures Database (MSigDB; 8 major collections)

(2) Manually created gene set

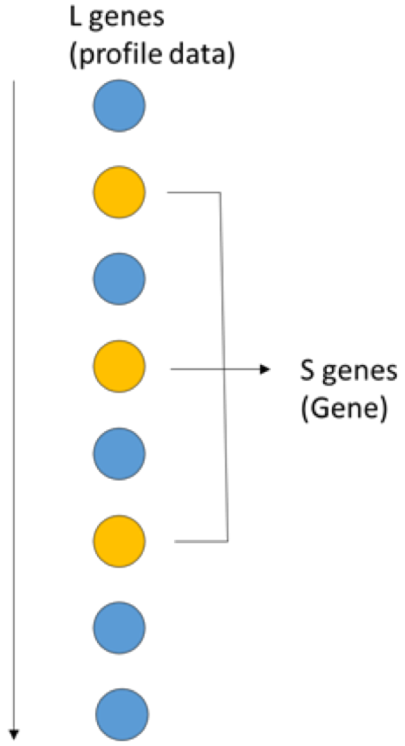
Collection	# of gene sets	Features
H: hallmark gene sets	Non-redundant from C1 to C7 (50)	Aggregating gene sets from many MSigDB
C1: positional gene sets	326	Sorting by chromosome and finding chromosome deletion and sequence amplification
C2: curated gene sets	4,729	Integrating many databases, such as pubmed, reactome, BioCarta and KEGG
C3: motif gene sets	836	Cis-regulatory motif information, including promoter, 3'-UTR, transcription factor target and microRNA target
C4: computational gene sets	858	Mining cancer-oriented microarray data and dividing into two groups: Cancer Gene Neighborhoods & Cancer module
C5: GO gene sets	6,166	Dividing into BF, CC and MF
C6: oncogenic gene sets	198	Oncogenic signature
C7: immunologic gene sets	4,872	Immunologic signature

All gene sets in MSigDB consist of **human gene symbols*

Step 3: Sorting dataset

- Based on the difference of gene's expression between two phenotype, a sorted dataset is generated.
- Five means to evaluate the difference of gene's expression:
  1. Signal2Noise [default]
  2. tTest (at least 3 samples)
  3. Ratio_of_class (higher FC, higher derivation from 1)
  4. Diff_of_classes (direct subtraction btwn two values)
  5. log2_Ratio_of_Classes (log2FC)

  …



** Correlation calculation is only allowed when phenotype **IS NOT** based on categories*

Step 4: Calculation of Enrichment Score (ES)

Step 5: Calculation of significant difference

Step 6: Calculation of multiple comparisons

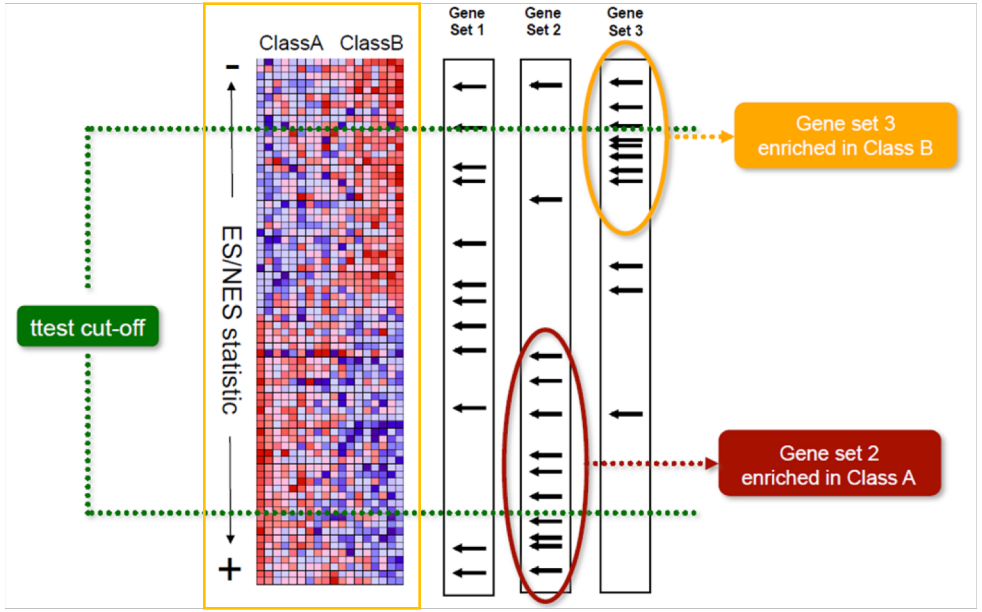
Case Study

Gene sets

Ingenuity Canonical Pathways	-log(p-value)	P value*100		Ratio	z-score	Molecules
Aryl Hydrocarbon Receptor Signaling	7.53E00	2.9512E-06	2.9512E-08	2E-01	0.447	GSTA5,SMARCA4,NR2F1,MYC,GSTM2,ALDH1A1,SP1,HSP90AB1,AL
ILK Signaling	4.54E00	0.00288403	2.884E-05	1.46E-01	-2.041	FLNB,MYH10,MYH9,PPP2R2A,BMP2,VEGFB,MYC,AKT1,RHOD,CREI
Xenobiotic Metabolism Signaling	4.09E00	0.00812831	8.1283E-05	1.25E-01	NaN	UGT2B28,UGT2B7,PPP2R2A,GSTA5,GCLC,CES2,MAPK13,FMO5,UG
AMPK Signaling	3.98E00	0.01047129	0.00010471	1.4E-01	0.500	ARID1A,PPP2R2A,MAPK13,NOS3,SMARCA4,AKT1,FASN,FOXO3,CR
NRF2-mediated Oxidative Stress Response	3.9E00	0.01258925	0.00012589	1.39E-01	-0.632	GSTM1,GSTA2,MGST1,ACTB,GSTA5,NQO1,SLC35A2,DNAJA4,HERF
PXR/RXR Activation	3.76E00	0.01737801	0.00017378	1.97E-01	NaN	SCD,GSTA2,GSTM1,AKT1,ALDH1A1,GSTM2,ALDH3A2,FOXO3,GSTA
EIF2 Signaling	3.75E00	0.01778279	0.00017783	1.36E-01	-0.775	RPL24,RPL22,RPL36A,EIF4G1,RPS4X,AKT1,EIF3B,EIF4G2,RPL21,R
Remodeling of Epithelial Adherens Junctions	3.62E00	0.02398833	0.00023988	1.91E-01	-2.449	CDH1,TUBB3,NME1,ARPC1A,ACTA2,HGF,ACTB,CTNNA1,ZYX,RAB5E
Purine Nucleotides De Novo Biosynthesis II	3.52E00	0.03019952	0.000302	4.55E-01	NaN	ADSL,ADSSL1,GMPS,IMPDH2,IMPDH1
Circadian Rhythm Signaling	3.14E00	0.0724436	0.00072444	2.42E-01	NaN	GRIN1,BHLHE41,CREB1,ATF4,CREB3L4,PER2,GRINA,GRIN3A
phagosome maturation	2.94E00	0.11481536	0.00114815	1.42E-01	NaN	B2M,ATP6V0C,TUBB3,ATP6V0B,VPS41,HLA-A,HLA-B,RAB5B,DYNLT
LPS/IL-1 Mediated Inhibition of RXR Function	2.9E00	0.12589254	0.00125893	1.18E-01	0.707	APOE,GSTA5,CES2,HMGCS2,FMO5,GSTM2,ALDH1A1,SCARB1,ALD
Prostate Cancer Signaling	2.82E00	0.15135612	0.00151356	1.59E-01	NaN	CCNE2,CREB3L4,SIN3A,AR,AKT1,HSP90AB1,CREB1,NKX3-1,ATF4,C
Glutathione-mediated Detoxification	2.8E00	0.15848932	0.00158489	2.41E-01	NaN	GSTM1,GSTA2,MGST1,GSTM2,GSTA5,GGH,GSTA1
Mitochondrial Dysfunction	2.68E00	0.20892961	0.0020893	1.23E-01	NaN	SDHA,COX7B,COX7B2,CASP3,GLRX2,ATP5A1,ATP5L,NDUFB3,NDUI
Palmitate Biosynthesis I (Animals)	2.42E00	0.3801894	0.00380189	1E00	NaN	OXSM,FASN
Fatty Acid Biosynthesis Initiation II	2.42E00	0.3801894	0.00380189	1E00	NaN	OXSM,FASN
Epithelial Adherens Junction Signaling	2.4E00	0.39810717	0.00398107	1.23E-01	NaN	MYH10,TUBB3,MYH9,NOTCH3,ACTB,PVRL3,CTNNA1,IQGAP1,TUBB
Caveolar-mediated Endocytosis Signaling	2.39E00	0.40738028	0.0040738	1.55E-01	NaN	B2M,ITGAE,FLNB,COPZ1,ALB,ACTA2,HLA-A,ACTB,HLA-B,RAB5B,IT
Oxidative Ethanol Degradation III	2.3E00	0.50118723	0.00501187	2.63E-01	NaN	ALDH1A1,ALDH3A2,ACSS2,ALDH9A1,ACSL1
mTOR Signaling	2.22E00	0.60255959	0.0060256	1.12E-01	-1.155	EIF3H,DDIT4,RHOC,PPP2R2A,VEGFB,PLD6,EIF4G1,RPS4X,EIF3F,A
Oleate Biosynthesis II (Animals)	2.18E00	0.66069345	0.00660693	3.08E-01	NaN	SCD,UFSP2,CYB5A,ALDH6A1
Oxidative Phosphorylation	2.13E00	0.74131024	0.0074131	1.28E-01	NaN	SDHA,ATP5C1,NDUFA5,COX7B,COX7B2,NDUFA6,ATP5A1,NDUFB7,N

Why comparing phenotypes A vs B gives different results from B vs A?

- This is because these two comparisons produce different ranked lists of genes. You might expect similar results only if the ranked lists would be perfectly symmetrical, and this usually does not happen.



Ranked lists of genes

(Source:http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ#What_is_the_difference_between_GSEA_and_an_overlap_statistic_.28hypergeometric.29_analysis_tool.3F)

Settings

- Number of permutations: 2000
- Permutation type: phenotype
- Collapse dataset to gene symbols: false
- Max size: 500
- Min size: 0

	Metric for ranking genes parameter
For at least three samples for each phenotype	Signal2noise, tTest
For fewer than three samples	Ratio_of_classes, log2_Ratio_of_classes, Diff_of_classes

Summary of GSEA results with NES, normal p-value and FDR

Gene set	# of genes[a]	NES[b]	Normal p-value[c]	FDR q-value[d]
Data set: M4&22Rv1				
Enriched in M4				
Y1:Aryl Hydrocarbon Receptor Signaling	64/102	1.22734	**<0.001**	**0.052[e]**
Y2:Xenobiotic Metabolism Signaling	46/91	1.214559	**<0.001**	**0.052**
Y4:Glutathione-mediated Detoxification	49/131	1.156147	**<0.001**	**0.082**
P1:Mitochondrial Dysfunction	108/153	1.059494	0.30997	0.251646
P2:Oxidative Phosphorylation	79/101	0.983912	0.516966	0.587326
Enriched in 22Rv1				
Y3:NRF2-mediated Oxidative Stress Response	81/129	-1.15999	**<0.001**	**0.1168**
Y5:Oxidative Ethanol III	38/63	-1.11686	**<0.001**	**0.144925**
R1:Purine Nucleotides De Novo Biosynthesis II	11/29	-0.93169	0.914117	0.836299

[a] The number of gene in gene set that match to the list of gene in dataset/total number of gene in gene set
[b] NES (normalized enrichment score) which is calculated by normalizing the ES (enrichment score) based on the size of each gene set
[c] The nominal p value estimates the statistical significance of the enrichment score for a single gene set
[d] FDR (false discovery rate) which is calculated by comparing tails of the observed and null distributions for the NES
[e] FDR ≤0.25, which indicates that the result is likely to be valid 3 out of 4 times

Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences* 102.43 (2005): 15545-15550.

Enrichment plots

Aryl Hydrocarbon Receptor Signaling

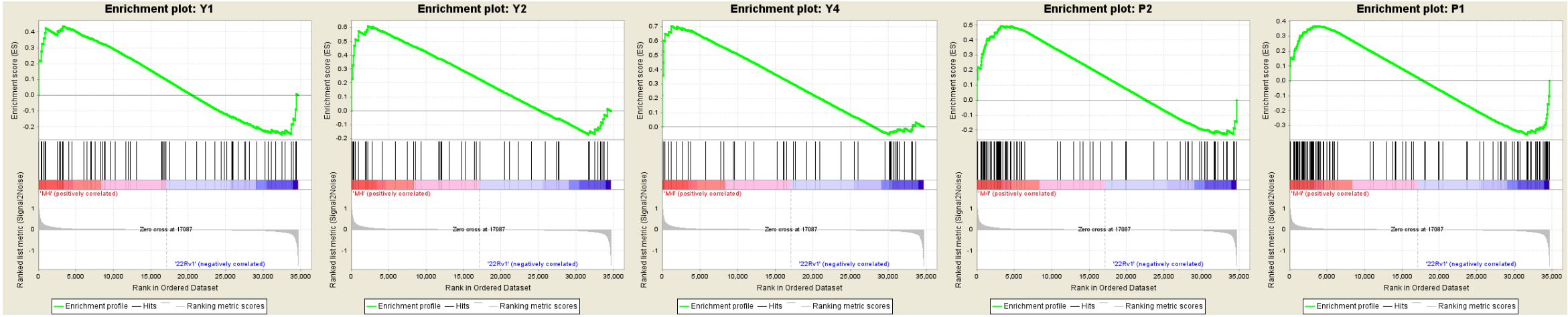
Xenobiotic Metabolism Signaling

Glutathione-mediated Detoxification

Mitochondrial Dysfunction

Oxidative Phosphorylation

Enriched in M4

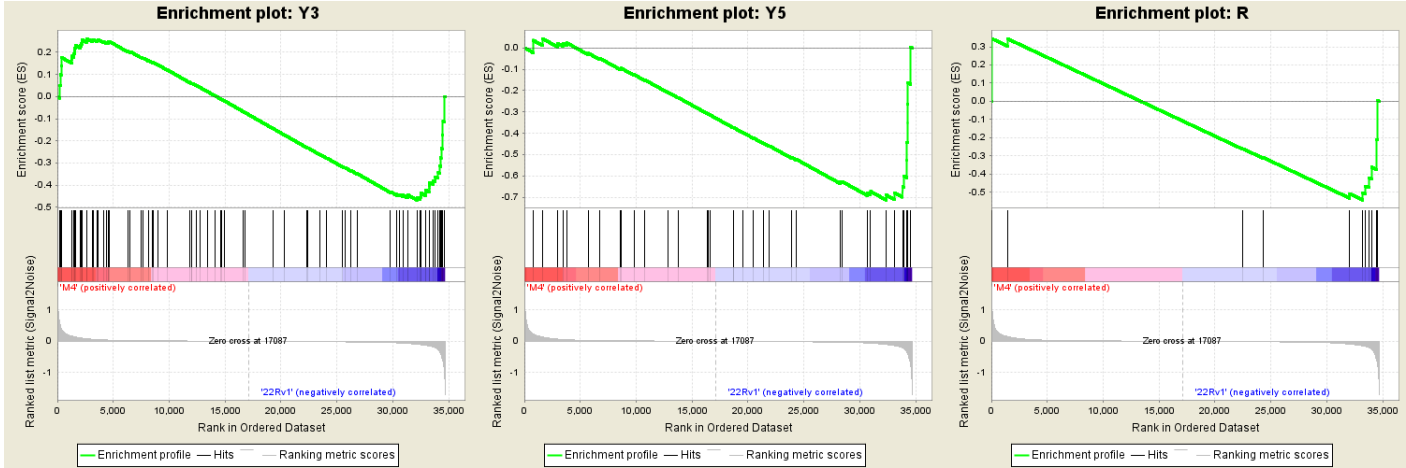


NRF2-mediated Oxidative Stress Response

Oxidative Ethanol III

Purine Nucleotides De Novo Biosynthesis II

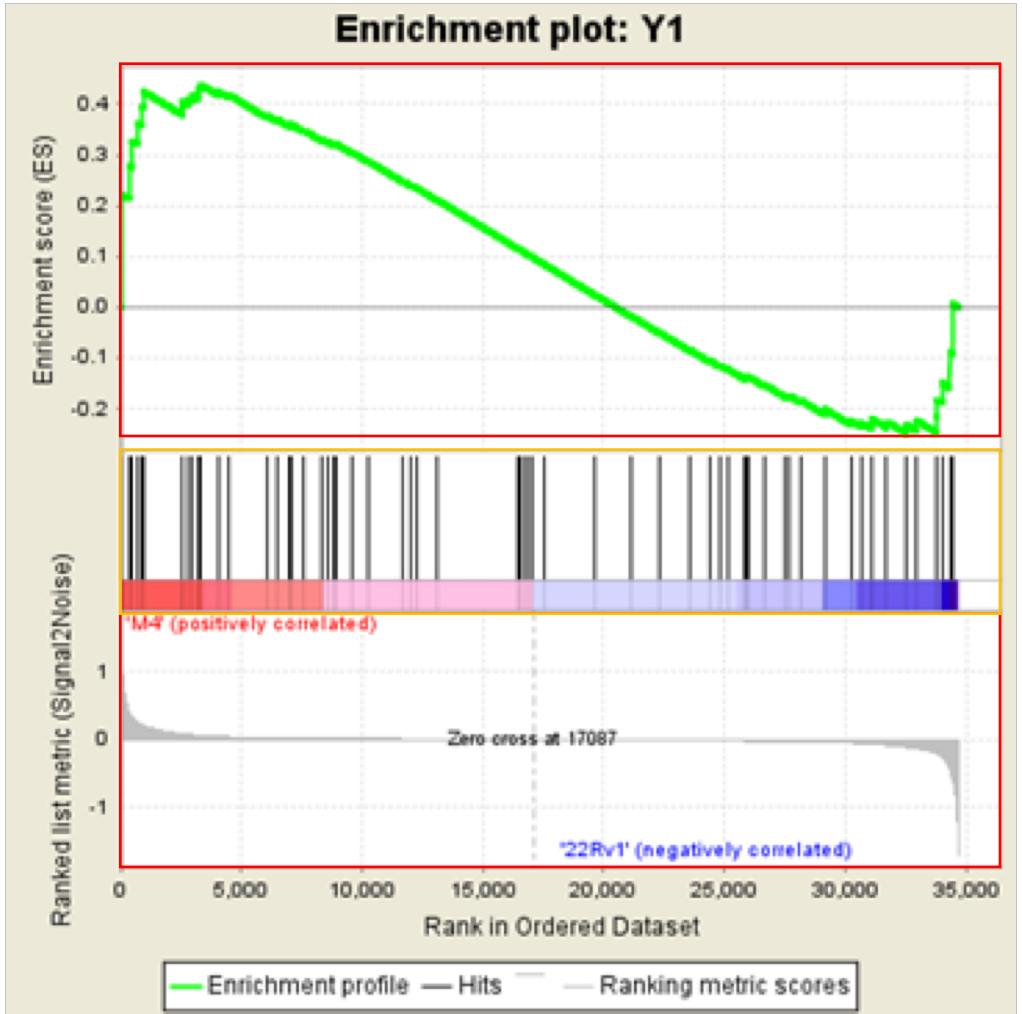
Enriched in 22Rv1



Interpretation of output figure

Positive ES
(M4)

negative ES
(22Rv1)



random walk

The distribution of gene set(S) in Ranked gene list

Correlation with phenotype