מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Gene set enrichment analysis (GSEA)

Ester Feldmesser
Bioinformatics Unit

March 2020

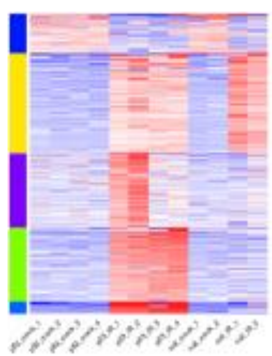
http://dors.weizmann.ac.il/course/GSEA/

After performing a complex high-throughput experiment:

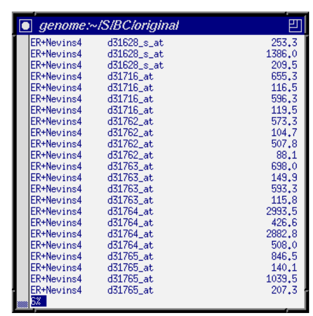
Microarrays
Deep Sequencing
Proteomics
…

What did we get?

Lists of genes

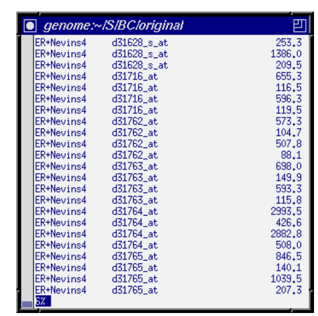


Clusters



ER+Nevins4	d31628_s_at	253.3
ER+Nevins4	d31628_s_at	1386.0
ER+Nevins4	d31628_s_at	209.5
ER+Nevins4	d31716_at	655.3
ER+Nevins4	d31716_at	116.5
ER+Nevins4	d31716_at	596.3
ER+Nevins4	d31716_at	119.5
ER+Nevins4	d31762_at	573.3
ER+Nevins4	d31762_at	104.7
ER+Nevins4	d31762_at	507.8
ER+Nevins4	d31762_at	88.1
ER+Nevins4	d31763_at	698.0
ER+Nevins4	d31763_at	149.9
ER+Nevins4	d31763_at	593.3
ER+Nevins4	d31763_at	115.8
ER+Nevins4	d31764_at	2993.5
ER+Nevins4	d31764_at	426.6
ER+Nevins4	d31764_at	2882.8
ER+Nevins4	d31764_at	508.0
ER+Nevins4	d31765_at	846.5
ER+Nevins4	d31765_at	140.1
ER+Nevins4	d31765_at	1039.5
ER+Nevins4	d31765_at	207.3

Up regulated



ER+Nevins4	d31628_s_at	253.3
ER+Nevins4	d31628_s_at	1386.0
ER+Nevins4	d31628_s_at	209.5
ER+Nevins4	d31716_at	655.3
ER+Nevins4	d31716_at	116.5
ER+Nevins4	d31716_at	596.3
ER+Nevins4	d31716_at	119.5
ER+Nevins4	d31762_at	573.3
ER+Nevins4	d31762_at	104.7
ER+Nevins4	d31762_at	507.8
ER+Nevins4	d31762_at	88.1
ER+Nevins4	d31763_at	698.0
ER+Nevins4	d31763_at	149.9
ER+Nevins4	d31763_at	593.3
ER+Nevins4	d31763_at	115.8
ER+Nevins4	d31764_at	2993.5
ER+Nevins4	d31764_at	426.6
ER+Nevins4	d31764_at	2882.8
ER+Nevins4	d31764_at	508.0
ER+Nevins4	d31765_at	846.5
ER+Nevins4	d31765_at	140.1
ER+Nevins4	d31765_at	1039.5
ER+Nevins4	d31765_at	207.3

Down regulated

Functional Genomics:
Find the Biological Meaning

- Take a list of "interesting" genes and find their biological meaning
  - Gene lists may come from significance/classfication analysis of microarrays, proteomics, or other high-throughput methods
- Requires a reference set of "biological knowledge"

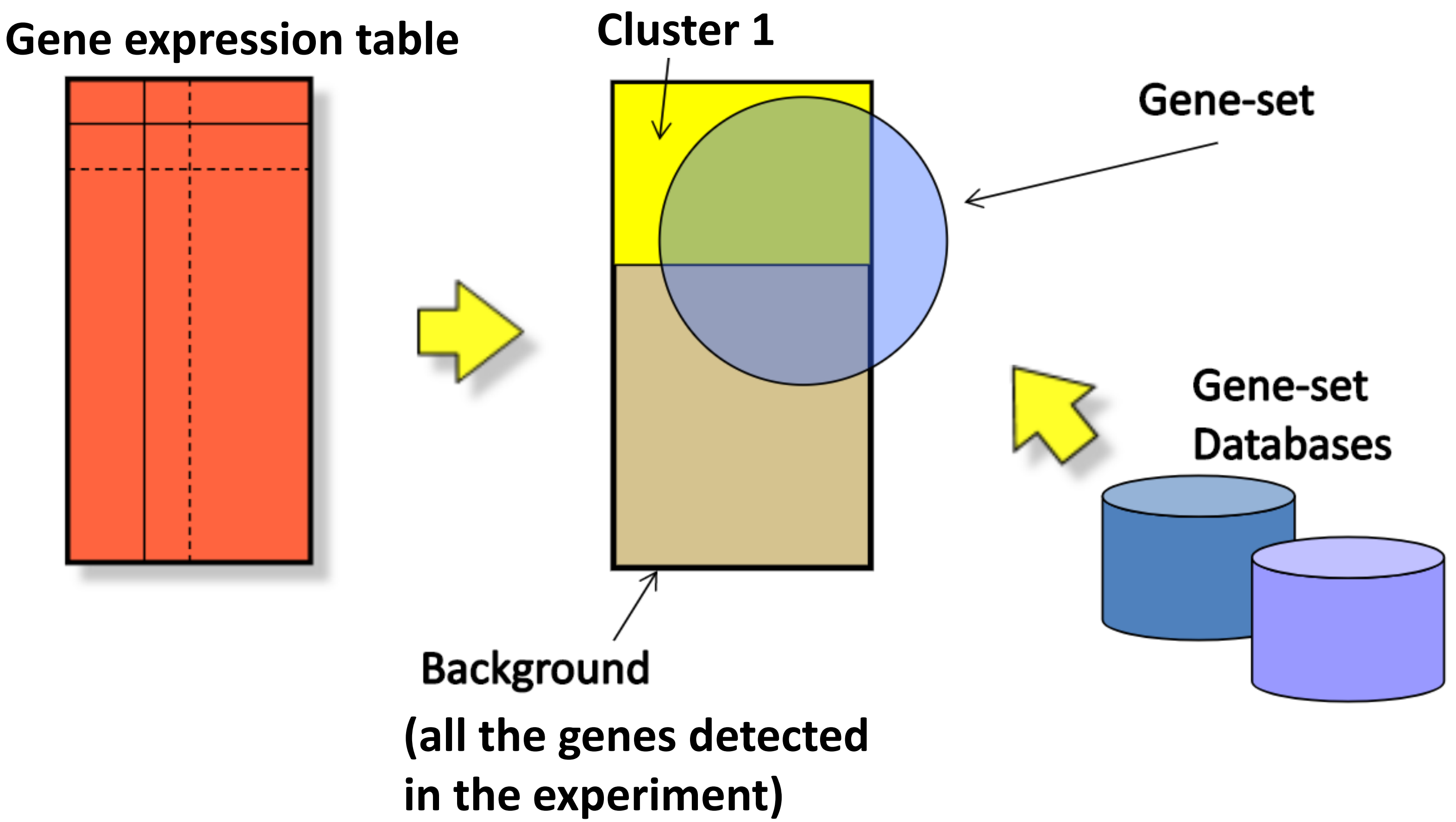
Sets of “Biological Knowledge"

- Linking between genes and biological function:
  - Gene ontology: GO
  - Pathways databases
- Discovery of common sequences in co-regulated genes
- Meta-studies using data from multiple experiments
  - Pubic and private gene or protein expression databases

Enrichment analysis (the most frequently used)

- Find your group of interesting genes (DE, up, down, cluster)
- Identify functional annotations that overlap and are over- represented (hypergeometric test).

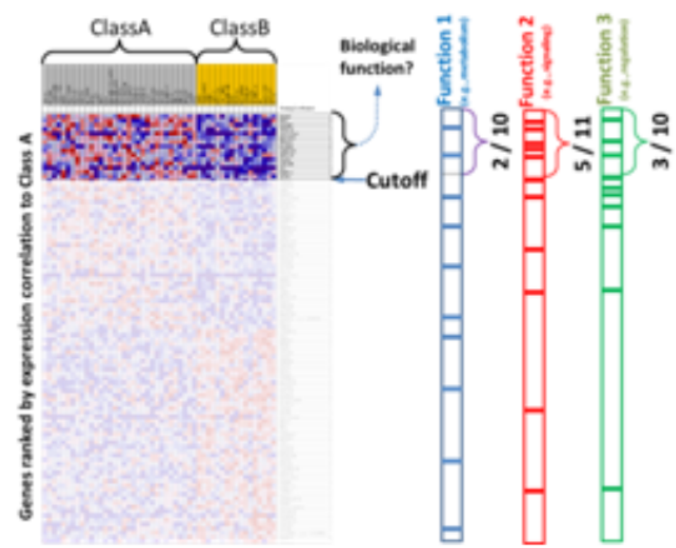
Enrichment test (hypergeometric)



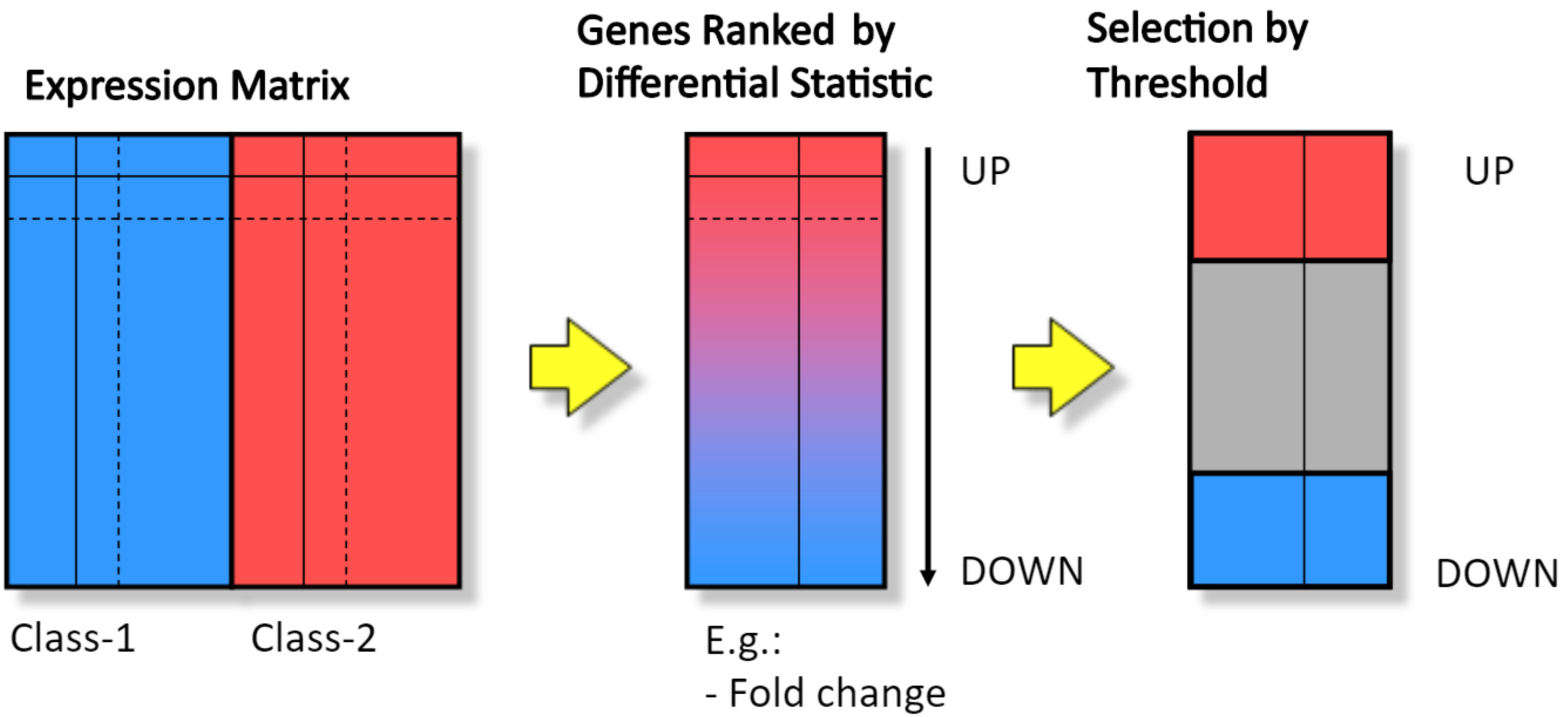
Is the overlap greater than expected by chance (random sampling of the background)?

Problems with cutoff-based analysis

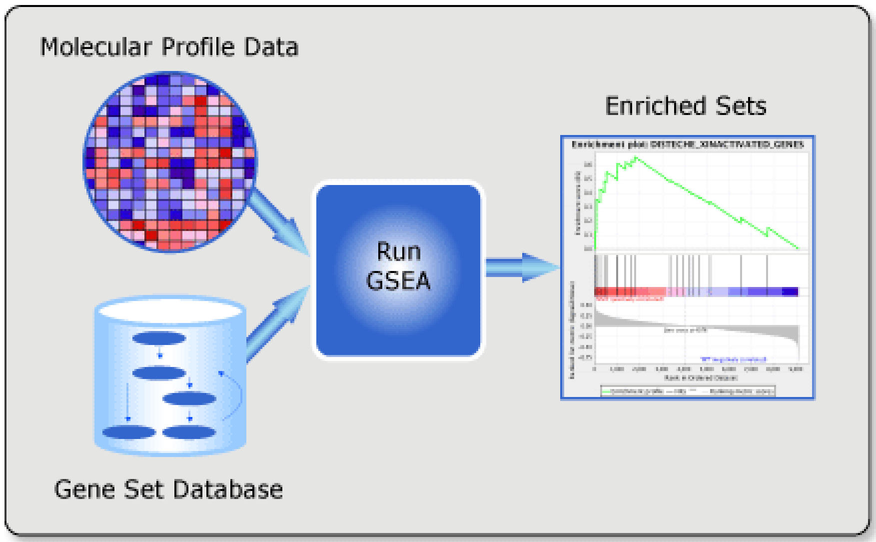
- After correcting for multiple hypotheses testing, no individual gene may meet the threshold due to noise.
- Alternatively, one may be left with a long list of significant genes without any unifying biological theme.
- The cutoff value is often arbitrary!
- **We are really examining only a handful of genes, totally ignoring much of the data**



Design of functional enrichment analysis



Gene Set Enrichment Analysis (GSEA)



http://www.broadinstitute.org/gsea/index.jsp
http://www.pnas.org/content/102/43/15545.full

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

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov

PNAS October 25, 2005 102 (43) 15545-15550; published ahead of print September 30, 2005
https://doi.org/10.1073/pnas.0506580102

Contributed by Eric S. Lander, August 2, 2005

Article | Figures & SI | Info & Metrics | PDF

Abstract

Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.

Gene set database

- The gene sets are defined based on **prior biological knowledge**, e.g., published information about biochemical pathways or co-expression in previous experiments

  and more….

Collections

The MSigDB gene sets are divided into 8 major collections:

H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

MSigDB
Molecular Signatures Database

- Recommended as a starting point.
- Hallmark gene sets summarize and represent specific well-defined biological states or processes.
- The hallmarks reduce noise and redundancy and provide a better delineated biological space for GSEA.

- CGP: chemical and genetic perturbations.
- CP: Canonical pathways
  - CP:BIOCARTA
  - CP:KEGG
  - CP:REACTOME

- The user can define new gene sets

http://www.broadinstitute.org/gsea/msigdb/collections.jsp#H

GSEA features

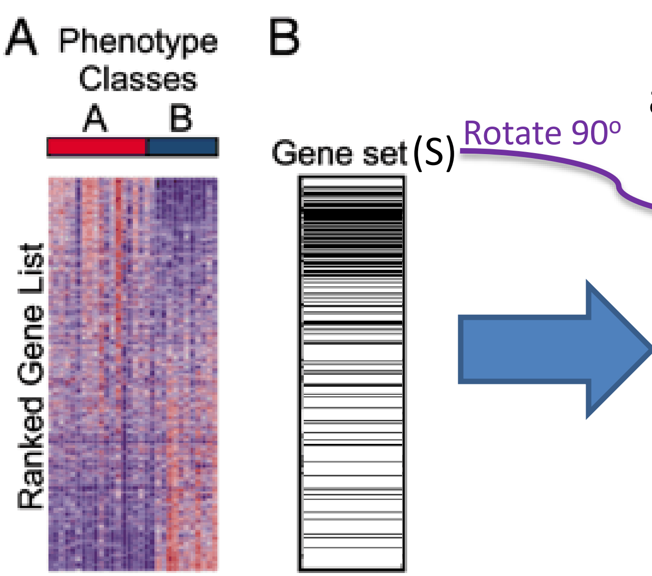
- GSEA performs its analysis on a list of ranked genes derived from comparing between two conditions, there is no need of cutoffs to define up or down regulated genes.

- Given a ranked list of differentially expressed genes, the goal of GSEA is to determine whether members of a gene set tend to occur toward the top (or bottom) of the list , in which case the gene set is correlated with the phenotypic class distinction (conditions).

Subramanian et al, 2005, PNAS ; 102(43): 15545–15550, doi: 10.1073/pnas.0506580102

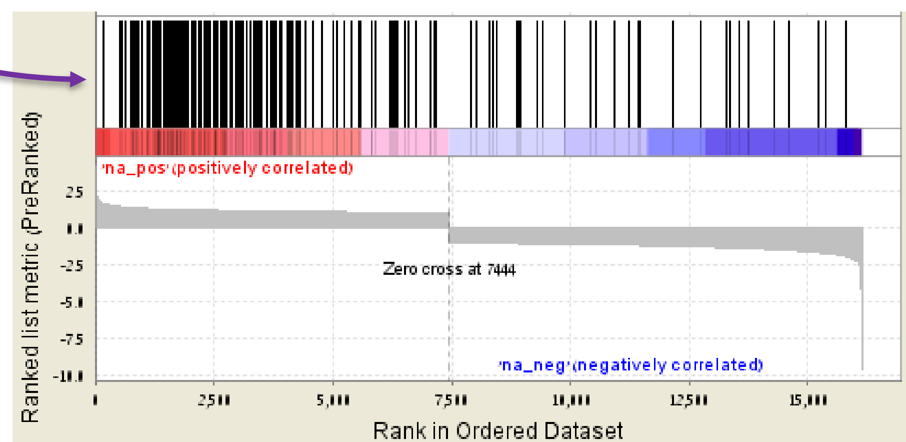
A GSEA overview illustrating the method

- Compute a gene-wise measure for differential expression between A and B and rank the genes according to this measure
- Alternatively a pre-ranked list can be used (L)
- Calculates a score for the enrichment of an **entire set of genes**



Input list: L

Black lines represent genes from the input that appear in the biological knowledge gene set



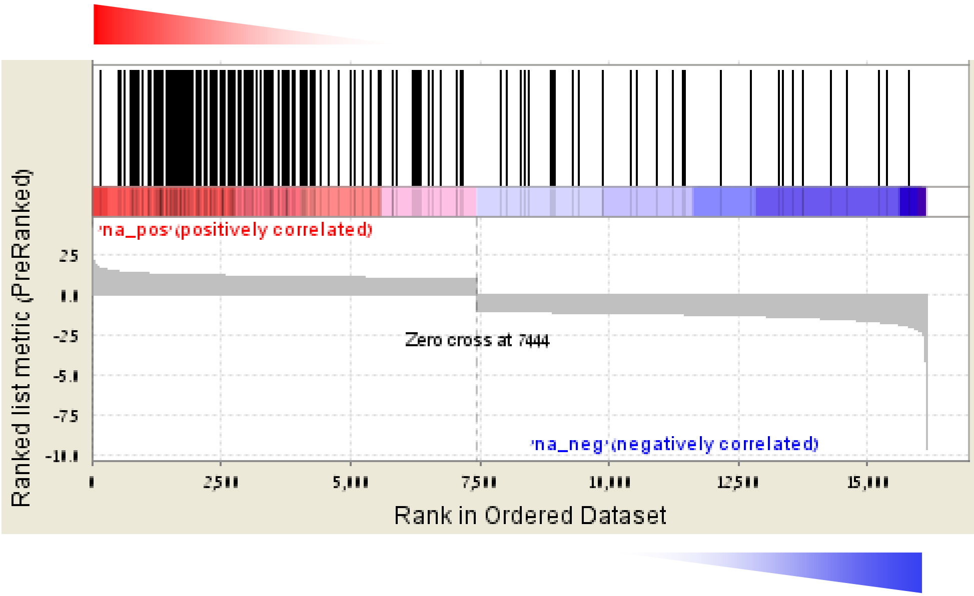
Pre-ranked gene list

Using RNA-seq Datasets with GSEA
https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Using_RNA-seq_Datasets_with_GSEA

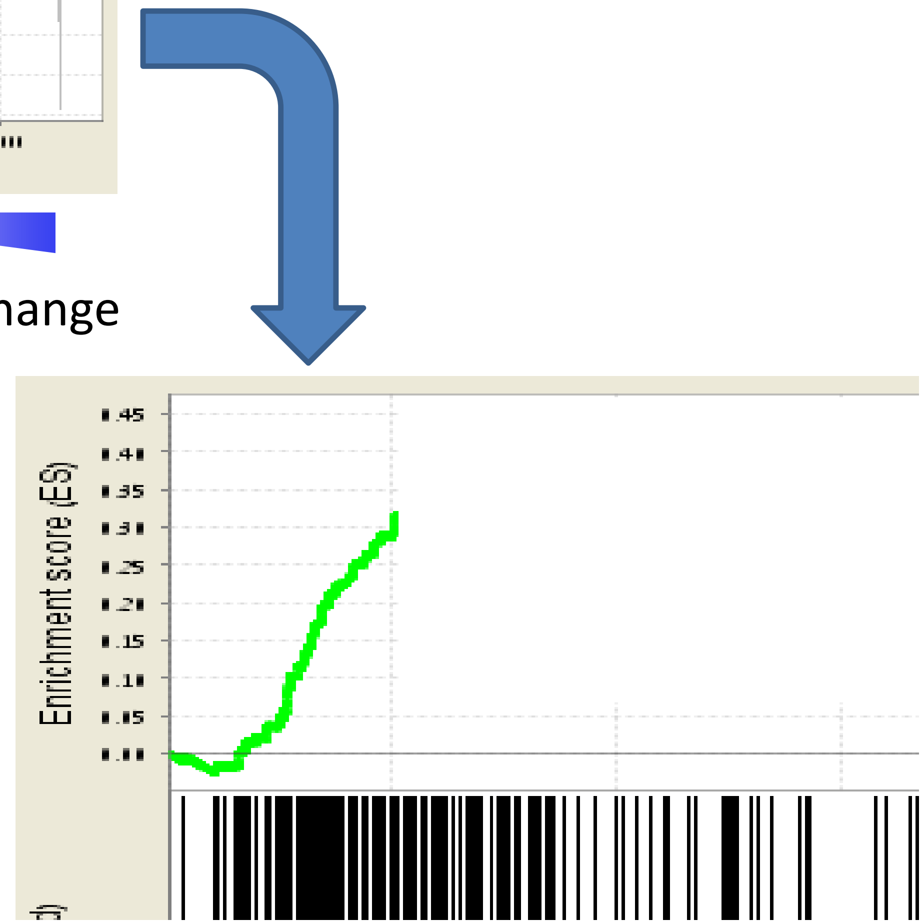
Alternative Method: GSEA-Preranked

1. Prior to conducting gene set enrichment analysis, conduct your differential expression analysis using any of the tools developed by the bioinformatics community (e.g., cuffdiff, edgeR, DESeq, etc).
2. Based on your differential expression analysis, rank your features and capture your ranking in an RNK-formatted file. The ranking metric can be whatever measure of differential expression you choose from the output of your selected DE tool: log2 fold change, p-value (-log10) or p-adjusted.

High fold change



Low fold change

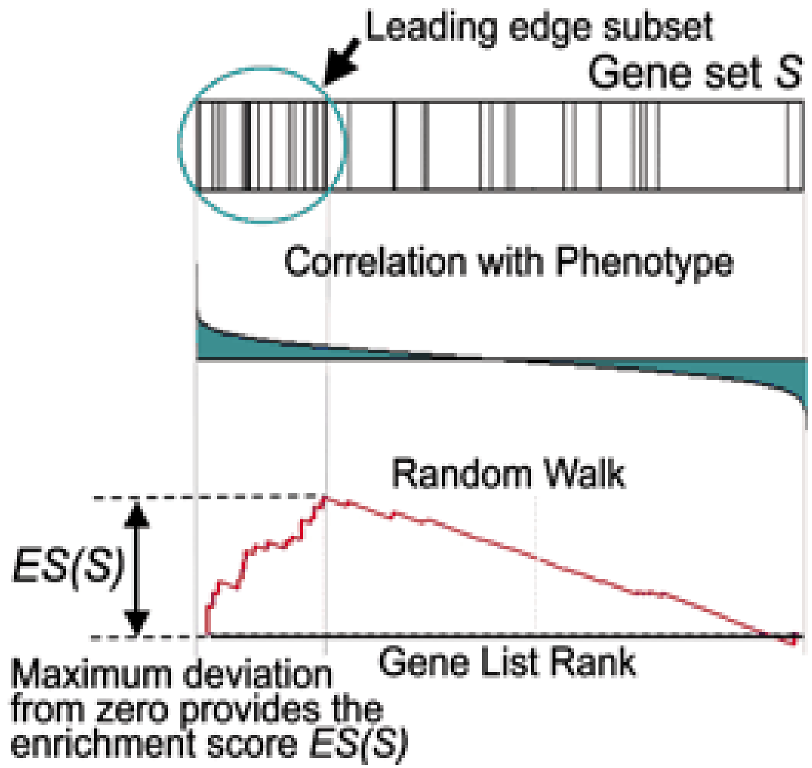


The score is calculated by walking down the input list *L*, increasing a running-sum statistic when we encounter a gene in the gene set *S* and decreasing it when we encounter genes not in *S.*

If up regulated genes in group A are enriched with genes from the Gene set S, many of its genes will have high ranks and we will observe a separation in the ordered list

The more genes found in S, the higher the Enrichment Score (**ES**)

But, when no genes from L are found in S for a long walk down, the ES will decrease



Gene Set Enrichment Analysis

Steps:

1. Calculation of an Enrichment Score (ES): maximum deviation from zero encountered in the walk
2. Normalization of the ES according to the sizes of the input list (L) and gene set (S), obtaining the normalized ES (NES).
3. Estimation of Significance Level of NES by permutations test
4. Adjustment for Multiple Hypothesis Testing

Multiple test correction

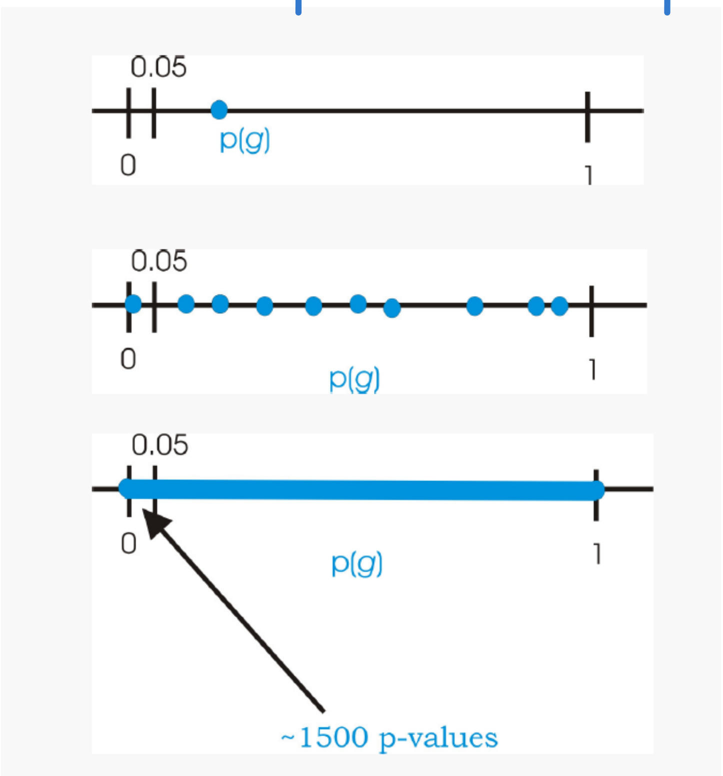
- FDR (False Detection Rate)
- Why?

Multiple testing gene sets
without overlap with our input list

1 comparison

10 comparisons

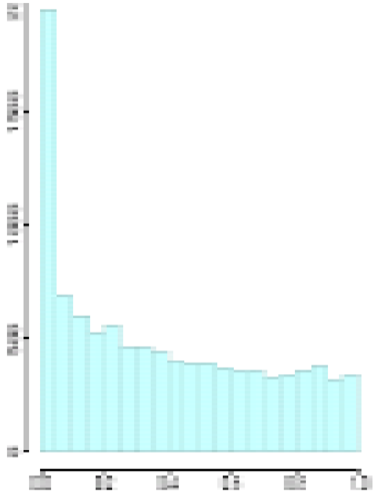
30,0000 comparisons



Looking at
p-values

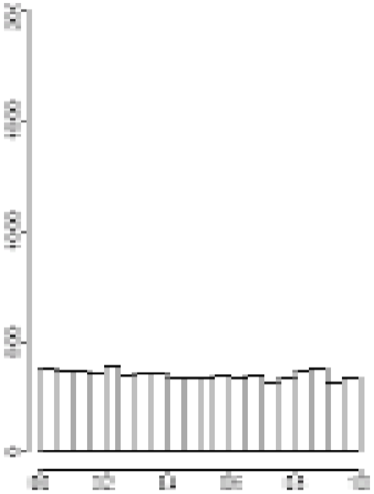
The mixture interpretation of the p-value

Mixture



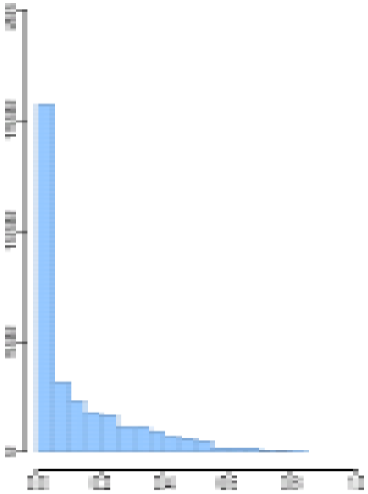
=

Uniform



+

Something



Multiple comparison correction

- False Discovery Rate (FDR) - Adjust the p-value in a way that ensures an expected proportion of false positives
- FDR-controlling procedures are designed to control the expected proportion of "discoveries" that are false

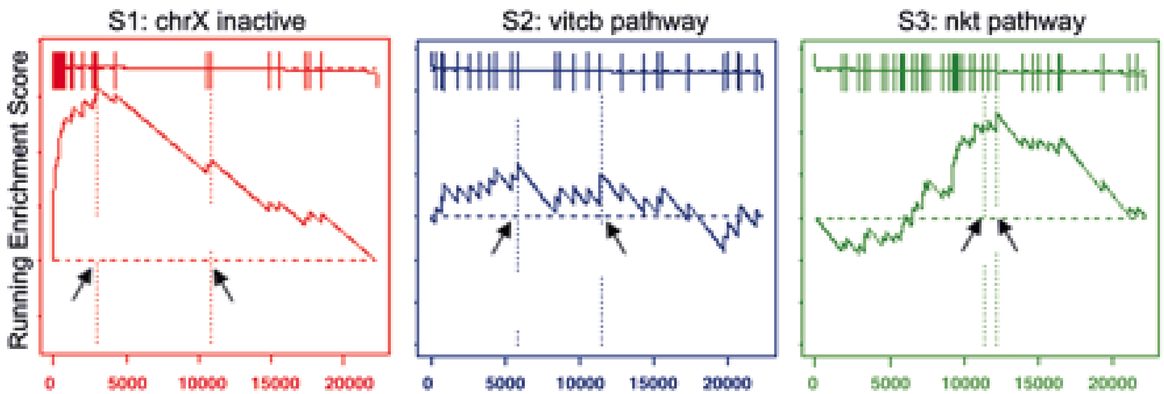
How many comparisons?

The FDR can change when:

- Using different Gene Sets
- Using a redundant Gene Sets

1	REACTOME_THE_CITRIC_ACID_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT
2	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT
3	REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS
4	REACTOME_COMPLEX_I_BIOGENESIS
5	REACTOME_MITOCHONDRIAL_TRANSLATION
6	KEGG_PARKINSONS_DISEASE
7	KEGG_OXIDATIVE_PHOSPHORYLATION
8	KEGG_ALZHEIMERS_DISEASE
9	REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE
10	REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA
11	KEGG_HUNTINGTONS_DISEASE

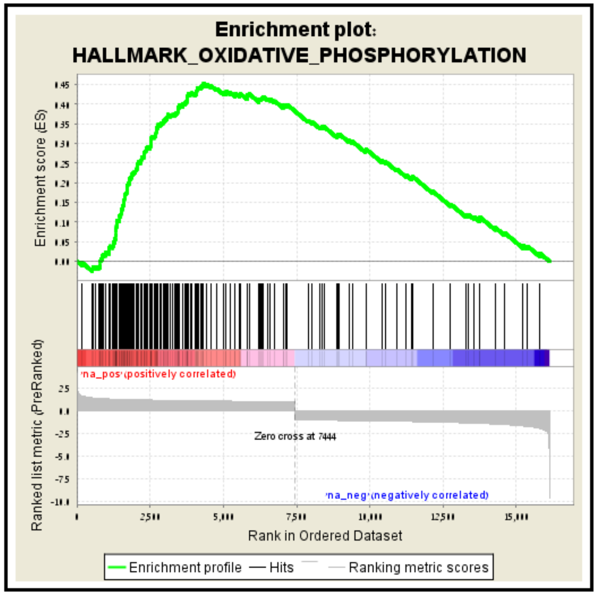
Examples from the paper

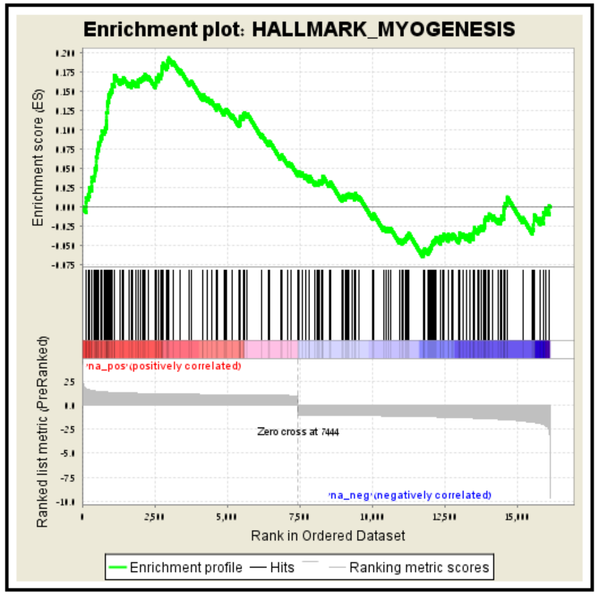


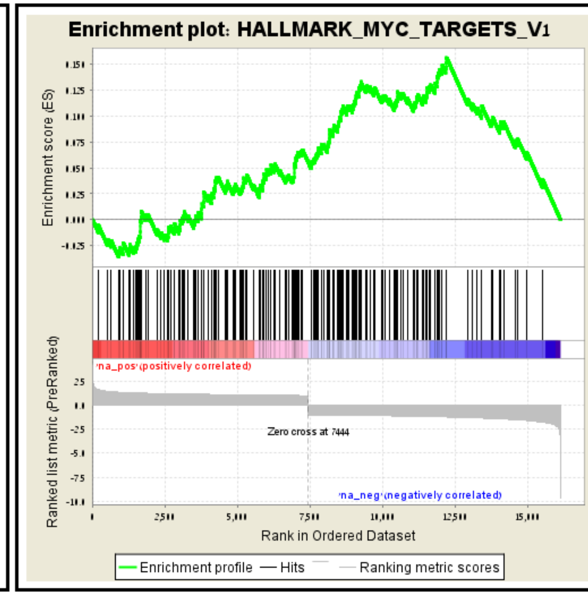
S1 is significantly enriched in females, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom.

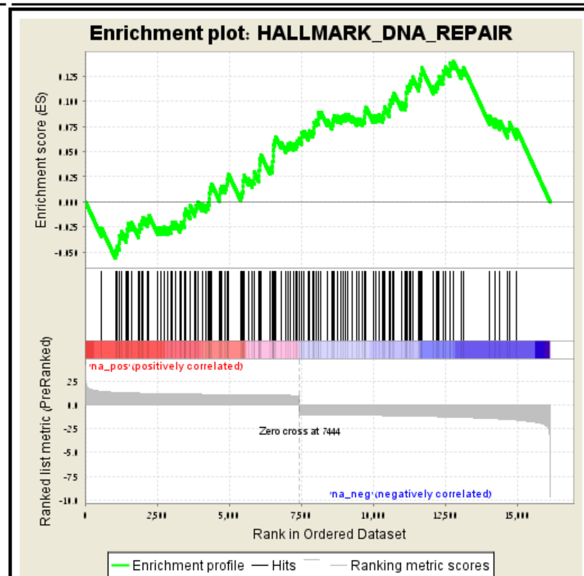
Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero

More examples







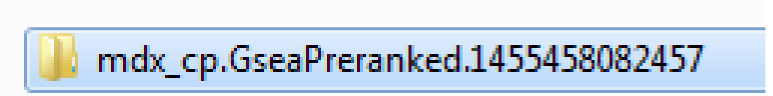


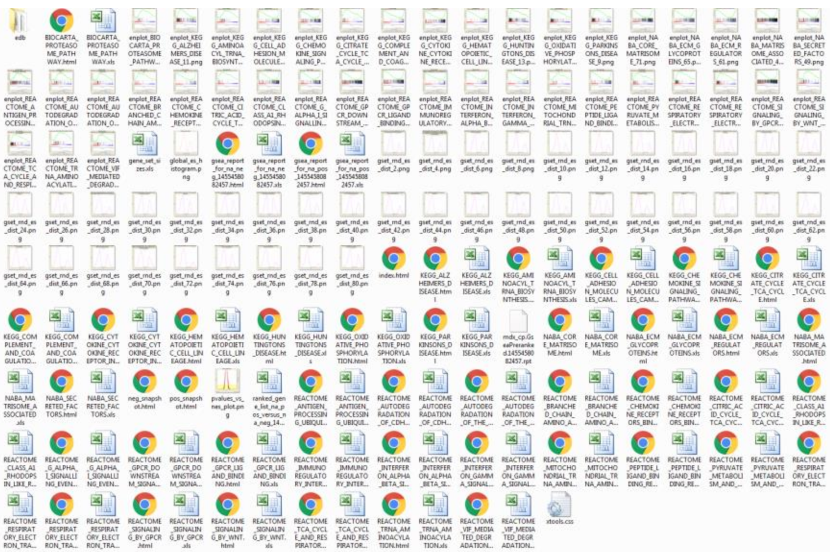
Gene Set Enrichment Analysis

Advantages

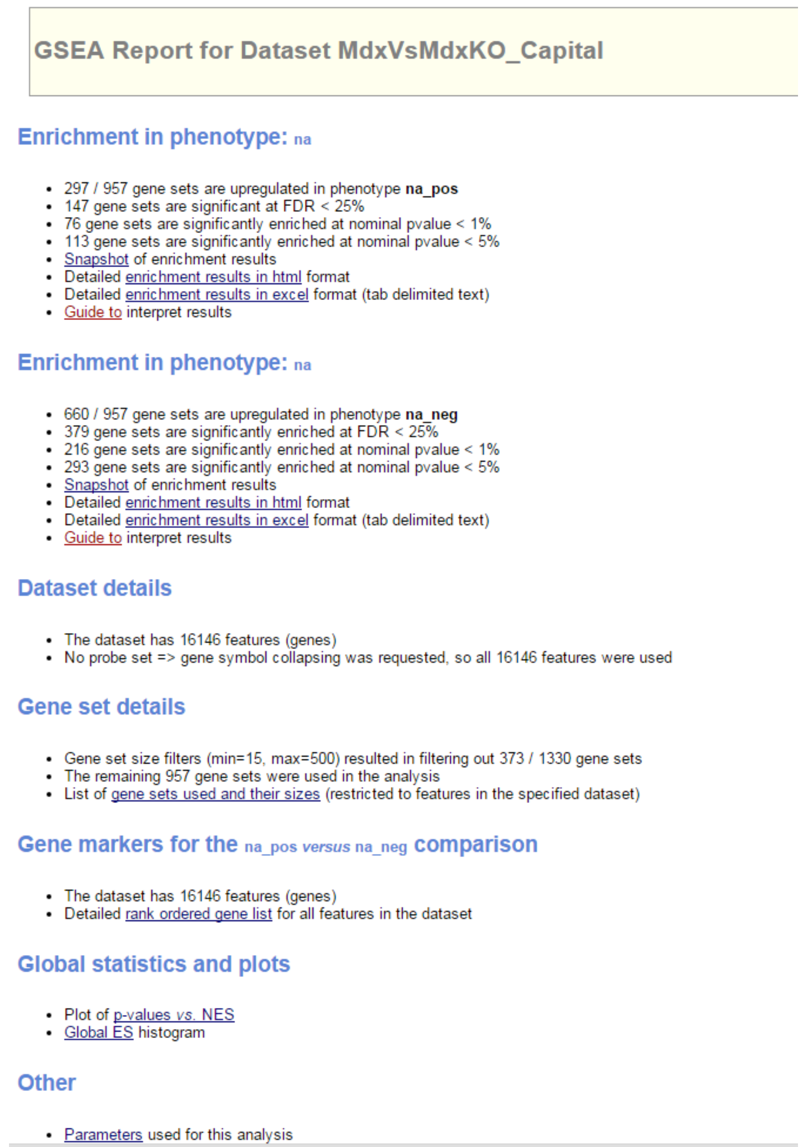
- Ranking of all genes is considered
- No cutoff has to be chosen

GSEA output





index.html



GSEA Report for Dataset MdxVsMdxKO_Capital

Enrichment in phenotype: na

- 297 / 957 gene sets are upregulated in phenotype **na_pos**
- 147 gene sets are significant at FDR < 25%
- 76 gene sets are significantly enriched at nominal pvalue < 1%
- 113 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: na

- 660 / 957 gene sets are upregulated in phenotype **na_neg**
- 379 gene sets are significantly enriched at FDR < 25%
- 216 gene sets are significantly enriched at nominal pvalue < 1%
- 293 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 16146 features (genes)
- No probe set => gene symbol collapsing was requested, so all 16146 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 373 / 1330 gene sets
- The remaining 957 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the na_pos *versus* na_neg comparison

- The dataset has 16146 features (genes)
- Detailed rank ordered gene list for all features in the dataset

Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

Other

- Parameters used for this analysis

Table: Gene sets enriched in phenotype na [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	HALLMARK_OXIDATIVE_PHOSPHORYLATION	Details ...	190	0.45	6.34	0.000	0.000	0.000	4319	tags=70%, list=27%, signal=94%
2	HALLMARK_MYOGENESIS	Details ...	194	0.19	2.70	0.000	0.000	0.000	2965	tags=36%, list=18%, signal=43%
3	HALLMARK_MYC_TARGETS_V1	Details ...	196	0.16	2.16	0.000	0.002	0.008	12217	tags=93%, list=76%, signal=377%
4	HALLMARK_DNA_REPAIR	Details ...	146	0.14	1.73	0.017	0.031	0.157	12784	tags=95%, list=79%, signal=450%
5	HALLMARK_FATTY_ACID_METABOLISM	Details ...	135	0.13	1.61	0.034	0.050	0.302	5634	tags=49%, list=35%, signal=74%
6	HALLMARK_HEME_METABOLISM	Details ...	162	0.12	1.55	0.032	0.058	0.391	3190	tags=30%, list=20%, signal=37%
7	HALLMARK_PROTEIN_SECRETION	Details ...	93	0.11	1.15	0.266	0.340	0.972	11931	tags=87%, list=74%, signal=332%
8	HALLMARK_SPERMATOGENESIS	Details ...	87	0.09	0.92	0.545	0.643	1.000	5292	tags=40%, list=33%, signal=60%
9	HALLMARK_PEROXISOME	Details ...	84	0.08	0.75	0.804	0.810	1.000	2821	tags=24%, list=17%, signal=29%

Gene Set: HALLMARK_OXIDATIVE_PHOSPHORYLATION

Standard name	HALLMARK_OXIDATIVE_PHOSPHORYLATION
Systematic name	M5936
Brief description	Genes encoding proteins involved in oxidative phosphorylation.
Full description or abstract	
Collection	H: hallmark gene sets
Source publication	
Exact source	
Related gene sets	(show 93 founder gene sets for this hallmark gene set)
External links	
Organism	Homo sapiens
Contributed by	Arthur Liberzon (Broad Institute)
Source platform	HUMAN_GENE_SYMBOL
Dataset references	(show 4 hallmark refinement datasets) (show 1 hallmark validation datasets)
Download gene set	format: grp | text | gmt | gmx | xml
Compute overlaps ?	(show collections to investigate for overlap with this gene set)
Compendia expression profiles ?	Human tissue compendium (Novartis) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 200 genes
Gene families ?	Categorize these 200 genes by gene family
Show members	(show 200 members mapped to 200 genes)
Version history	5.0: First introduced

Table: GSEA Results Summary

Dataset	MdxVsMdxKO_Capital
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	HALLMARK_OXIDATIVE_PHOSPHORYLATION
Enrichment Score (ES)	0.4527396
Normalized Enrichment Score (NES)	6.337497
Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

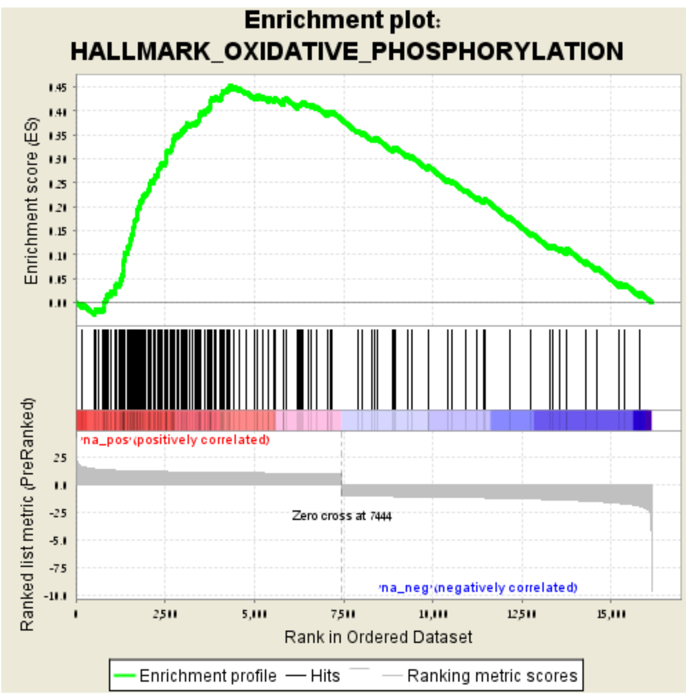


Table: GSEA details [plain text format]

	PROBE	GENE SYMBOL	GENE_TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	MAOB			167	1.659	-0.0033	Yes
2	VDAC1			519	1.447	-0.0190	Yes
3	IDH3A			533	1.441	-0.0136	Yes
4	VDAC3			641	1.413	-0.0141	Yes
5	ATP5O			722	1.393	-0.0131	Yes
6	COX15			773	1.382	-0.0102	Yes
7	COX11			780	1.380	-0.0046	Yes
8	NQO2			781	1.380	0.0014	Yes
9	SDHA			807	1.374	0.0058	Yes
10	ALDH6A1			858	1.363	0.0085	Yes
11	NDUFB2			859	1.363	0.0145	Yes
12	PRDX3			896	1.357	0.0181	Yes
13	CS			973	1.346	0.0192	Yes
14	SLC25A12			977	1.345	0.0248	Yes
15	ATP5E			1075	1.332	0.0245	Yes
16	PDP1			1078	1.331	0.0302	Yes
17	IDH3G			1088	1.329	0.0354	Yes
18	NDUFS1			1117	1.326	0.0394	Yes
19	UQCRC2			1213	1.313	0.0391	Yes
20	FXN			1231	1.311	0.0437	Yes
21	SUCLA2			1237	1.311	0.0491	Yes
22	NDUFV2			1272	1.306	0.0526	Yes
23	NDUFB4			1284	1.305	0.0576	Yes
24	NDUFA5			1305	1.303	0.0620	Yes
25	PMPCA			1306	1.303	0.0676	Yes
26	ACO2			1319	1.301	0.0725	Yes
27	ISCU			1320	1.301	0.0782	Yes
28	BCKDHA			1334	1.300	0.0830	Yes

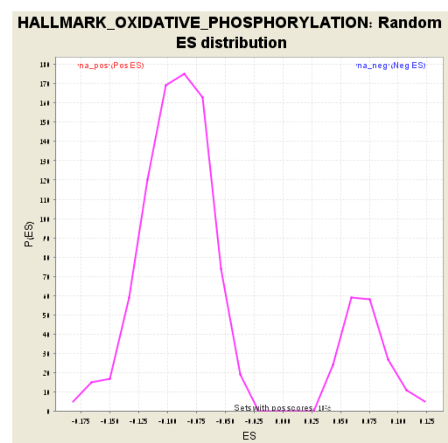


Fig 2: HALLMARK_OXIDATIVE_PHOSPHORYLATION: Random ES distribution
Gene set null distribution of ES for HALLMARK_OXIDATIVE_PHOSPHORYLATION