Gene Set Enrichment Analysis (GSEA)

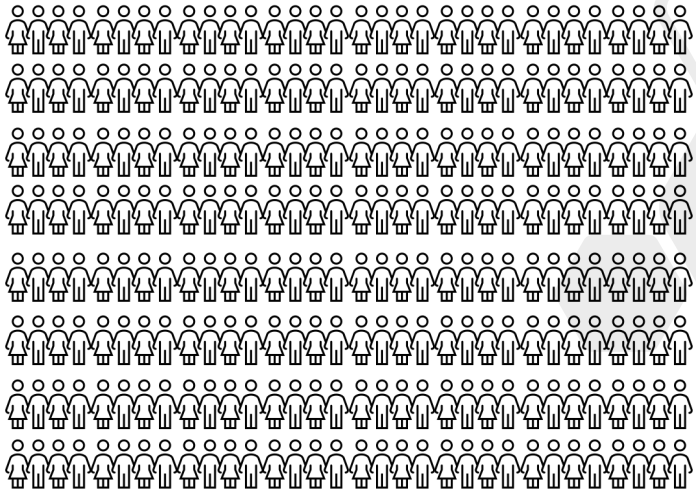
Assistant Professor, Division of Biomedical Informatics, College of Medicine
Manager, Cancer Research Informatics Shared Resource Facility,
Markey Cancer Center
University of Kentucky

UK HealthCare MARKEY CANCER CENTER
An NCI-Designated Cancer Center

October 5, 2020

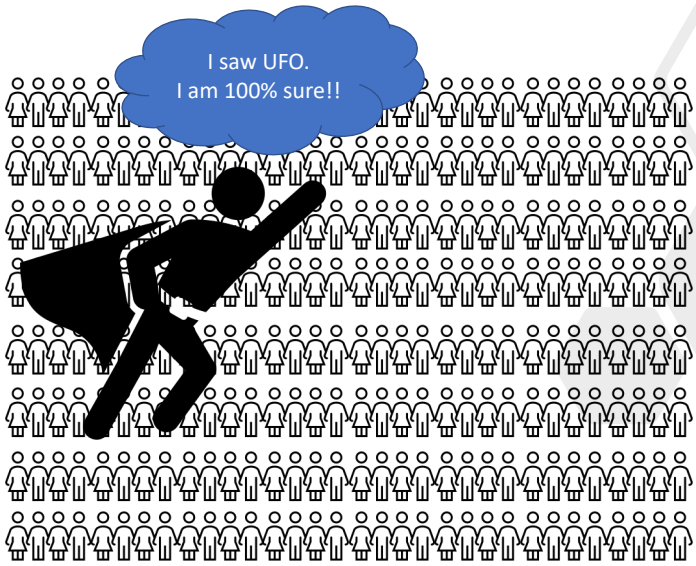
1

Gene Set Enrichment Analysis (GSEA)



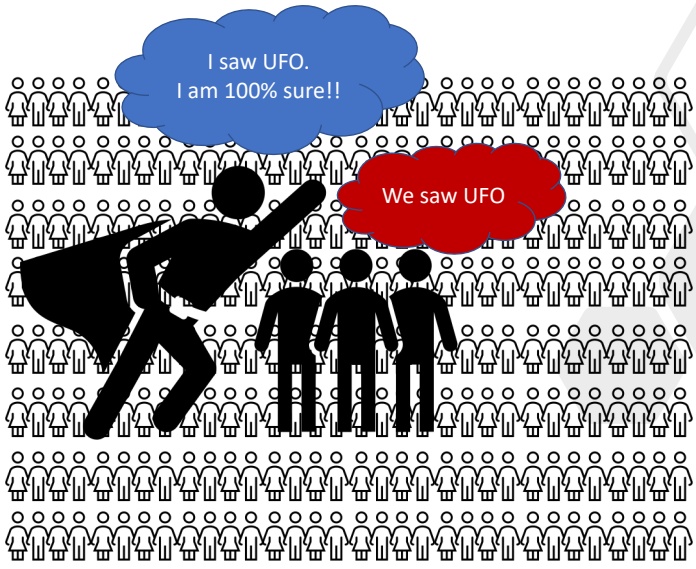
2

Gene Set Enrichment Analysis (GSEA)



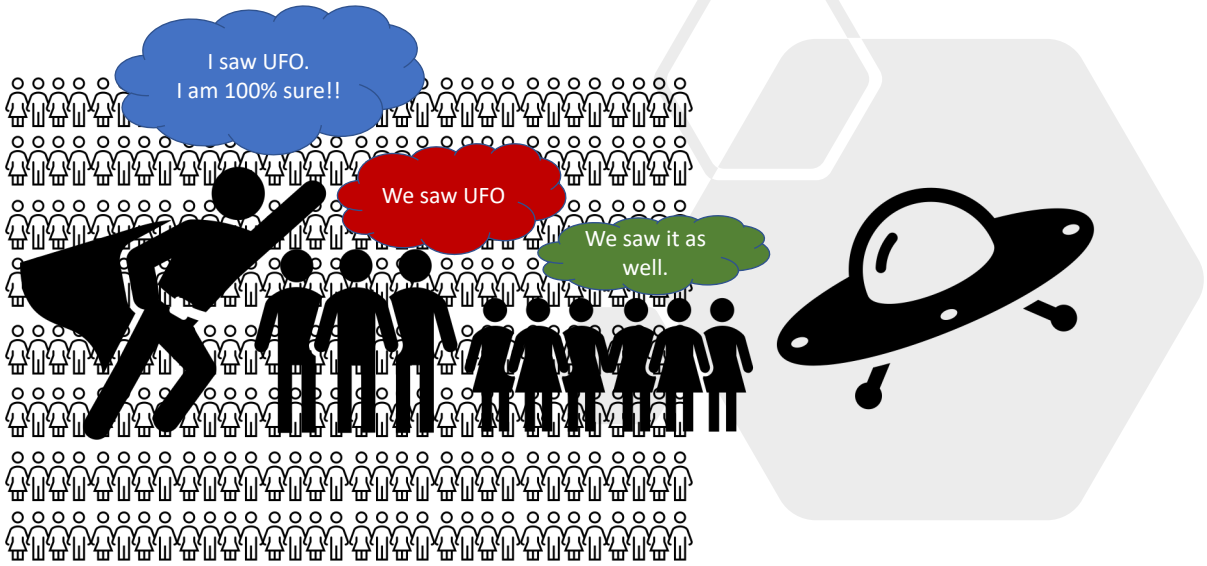
3

Gene Set Enrichment Analysis (GSEA)



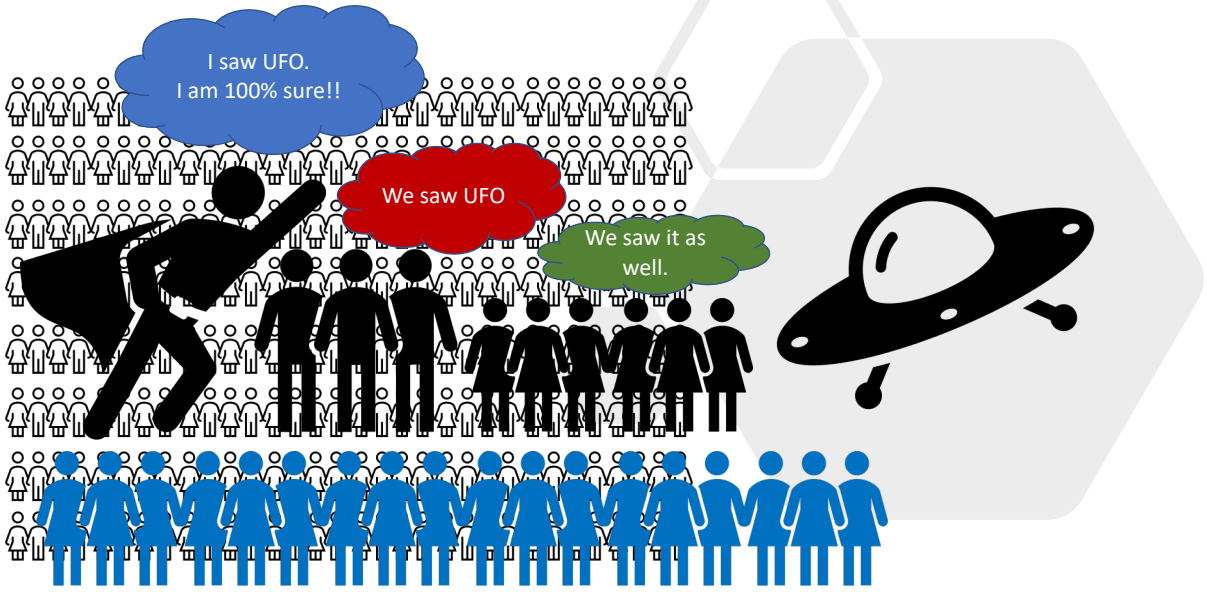
4

Gene Set Enrichment Analysis (GSEA)



5

Gene Set Enrichment Analysis (GSEA)



6

Gene Set Enrichment Analysis (GSEA)

- **Gene Set**: groups of genes that share common biological function, chromosomal location, or regulation. Defined based on prior biological knowledge
- **Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant and concordant differences between two biological states.
- GSEA determines whether members of a gene set *S* tend to occur toward the top (or bottom) of the list *L*, in which case the gene set is correlated with the phenotypic class distinction.

https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

7

GSEA Vs. Differential Expression

Differential Expression

- Single gene may fail to pass statistical tests, due to the noise inherited from the technology rather than biological differences.
- Results can be resulted in a long list of statistically significant genes without any unifying biological theme, so interpretation can be biased toward researcher's area of expertise.
- Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. For example, an increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.

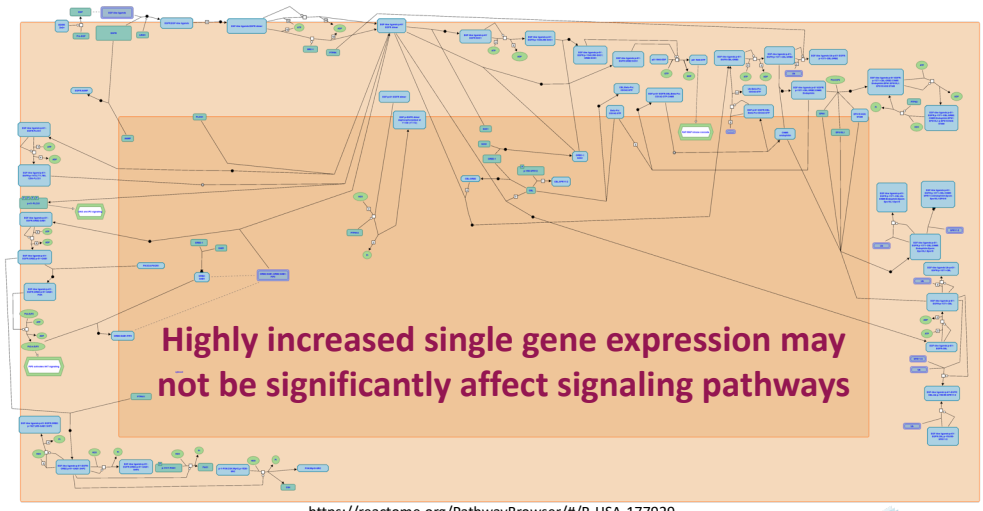
GSEA

- Groups of genes that share common biological function, chromosomal location, or regulation.
- When different groups study the same biological system, GSEA will be likely to show higher overlapping results than the one from differential expression.

Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

8

Signaling by the Epidermal Growth Factor Receptor (EGFR)

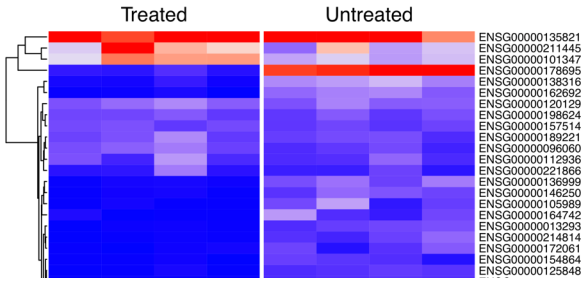


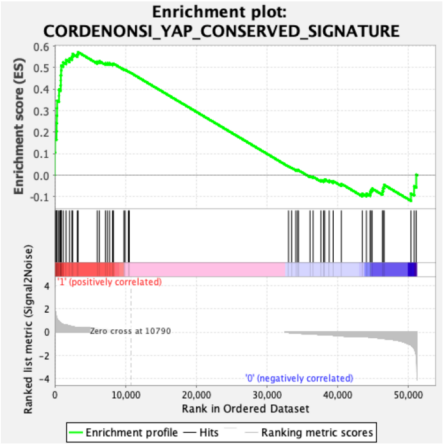
https://reactome.org/PathwayBrowser/#/R-HSA-177929

9

GSEA Vs. Differential Expression

Differential Expression	GSEA





10

GSEA Method

Preprocess 1: Expression data from at least two groups

	Treated / Sample				Normal / Control			
GENE	SRR1039509	SRR1039513	SRR1039517	SRR1039521	SRR1039508	SRR1039512	SRR1039516	SRR1039520
MCHR1	448	408	1047	572	679	873	1138	770
LINC00906	0	0	0	0	0	0	0	0
LRRTM2	515	365	799	508	467	621	587	417
CYP24A1	211	164	331	229	260	263	245	233
ADCY8	55	35	63	60	60	40	78	76
VASH2	0	0	0	0	0	2	1	0
CCL8	3679	4252	11027	7995	3251	6177	6721	5176
GRIN2A	1062	881	1439	1109	1433	1733	1424	1359
VCAM1	380	493	714	704	519	595	820	696
RAMP3	236	175	584	269	394	464	658	360
SMTNL2	168	118	210	177	172	264	241	155
SLITRK6	1867	2657	2751	2905	2112	5137	2735	2467
WNT2	488	357	806	475	524	638	676	493
FER1L6	51	156	38	172	71	211	23	134
LRRC25	394	415	697	599	555	905	727	618

11

GSEA Method

Preprocess 2: Statistical Analysis of Two Groups (optional)

	Treated / Sample				Normal / Control			
GENE	SRR1039509	SRR1039513	SRR1039517	SRR1039521	SRR1039508	SRR1039512	SRR1039516	SRR1039520
MCHR1	448	408	1047	572	679	873	1138	770
LINC00906	0	0	0	0	0	0	0	0
LRRTM2	515	365	799	508	467	621	587	417
CYP24A1	211	164	331	229	260	263	245	233
ADCY8	55	35	63	60	60	40	78	76
VASH2	0	0	0	0	0	2	1	0
CCL8	3679	4252	11027	7995	3251	6177	6721	5176
GRIN2A	1062	881	1439	1109	1433	1733	1424	1359
VCAM1	380	493	714	704	519	595	820	696
RAMP3	236	175	584	269	394	464	658	360
SMTNL2	168	118	210	177	172	264	241	155
SLITRK6	1867	2657	2751	2905	2112	5137	2735	2467
WNT2	488	357	806	475	524	638	676	493
FER1L6	51	156	38	172	71	211	23	134
LRRC25	394	415	697	599	555	905	727	618

Statistical Test

- t-test
- Differential Expression

Ranked Gene List

GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

12

GSEA Method

Gene Set: groups of genes that share common biological function, chromosomal location, or regulation. Defined based on prior biological knowledge

Preprocess 3: Choose Gene Sets

MSigDB Molecular Signatures Database

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 gene sets** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

C5: GO gene sets (browse 9996 gene sets)	Gene sets that contain genes annotated by the same GO term. The C5 collection is divided into three sub-collections based on GO ontologies: BP, CC, and MF. details	Download GMT Files gene symbols NCBI (entrez) gene ids
BP: GO biological process (browse 7350 gene sets)	Gene sets derived from the GO Biological Process Ontology.	Download GMT Files gene symbols NCBI (entrez) gene ids
CC: GO cellular component (browse 1001 gene sets)	Gene sets derived from the GO Cellular Component Ontology.	Download GMT Files gene symbols NCBI (entrez) gene ids
MF: GO molecular function (browse 1645 gene sets)	Gene sets derived from the GO Molecular Function Ontology.	Download GMT Files gene symbols NCBI (entrez) gene ids
C6: oncogenic signatures (browse 189 gene sets)	Gene sets that represent signatures of cellular pathways which are often dis-regulated in cancer. The majority of signatures were generated directly from microarray data from NCBI GEO or from internal unpublished profiling experiments involving perturbation of known cancer genes. details	Download GMT Files gene symbols NCBI (entrez) gene ids
C7: immunologic signatures (browse 4872 gene sets)	Gene sets that represent cell states and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology. details	Download GMT Files gene symbols NCBI (entrez) gene ids

https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

13

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.

S

Query Gene Set

Gene Set
ADCY8
CADM1
CCL8
CYP24A1
DUSP8
LINC00906
LRRTM2
MCHR1
TENT5A
VASH2

L

Reference Signature

Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

14

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.

S

Query Gene Set

Gene Set
ADCY8
CADM1
CCL8
CYP24A1
DUSP8
LINC00906
LRRTM2
MCHR1
TENT5A
VASH2

Ranked Gene List

GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

L

Reference Signature

$$ES = P_{hit} - P_{miss}$$

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} \; where \; N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$
$$where \; N = |L_{gene}|, N_H = |S|$$

Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

15

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.

$$ES = P_{hit} - P_{miss}$$

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} where \; N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$
$$where \; N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

S

Query Gene Set

Gene Set
ADCY8
CADM1
CCL8
CYP24A1
DUSP8
LINC00906
LRRTM2
MCHR1
TENT5A
VASH2

Ranked Gene List

GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

L

Reference Signature

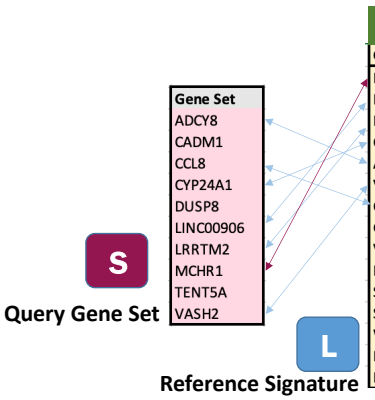
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

16

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} \text{ where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$
$$\text{where } N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

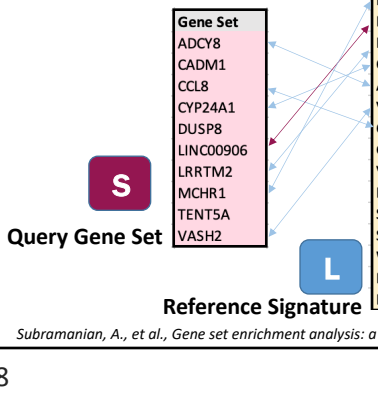
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

17

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} \text{ where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$
$$\text{where } N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

$P_{hit}(S,LINC00906) = \frac{4.8}{28.2} + \frac{4.4}{28.2} = 0.33$ $P_{miss}(S,LINC00906) = 0$ **ES(LINC00906) = 0.33**

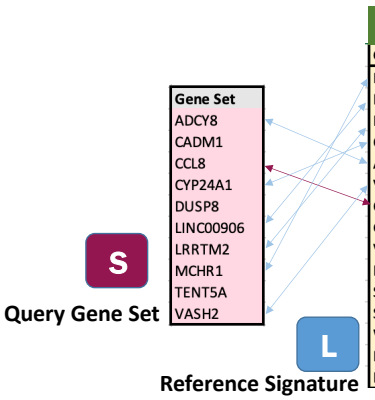
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

18

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} \quad where\ N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$
$$where\ N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

$P_{hit}(S,LINC00906) = \frac{4.8}{28.2} + \frac{4.4}{28.2} = 0.33$ $P_{miss}(S,LINC00906) = 0$ **ES(LINC00906) = 0.33**

$P_{hit}(S,CCL8) = ?$ $P_{miss}(S,CCL8) = ?$ **ES(CCL8) = ?**

Calculate it your self!

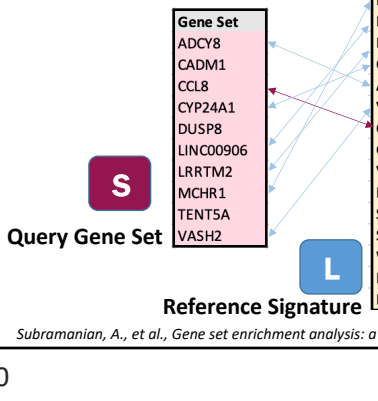
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

19

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R} \quad where\ N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$
$$where\ N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

$P_{hit}(S,LINC00906) = \frac{4.8}{28.2} + \frac{4.4}{28.2} = 0.33$ $P_{miss}(S,LINC00906) = 0$ **ES(LINC00906) = 0.33**

$P_{hit}(S,CCL8) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,CCL8) = 0$ **ES(CCL8) = 1.0**

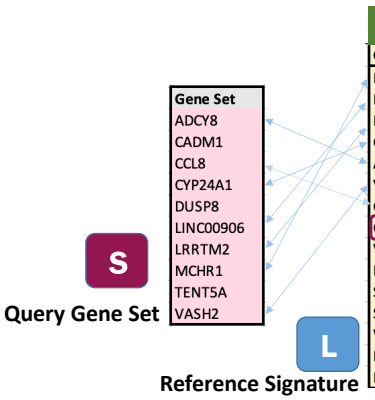
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

20

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g\,j \in S, j \le i} \frac{|r_j|^p}{N_R} \text{ where } N_R = \sum_{g\,j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g\,j \in S, j \le i} \frac{1}{N - N_H}$$
$$\text{where } N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

$P_{hit}(S,LINC00906) = \frac{4.8}{28.2} + \frac{4.4}{28.2} = 0.33$ $P_{miss}(S,LINC00906) = 0$ **ES(LINC00906) = 0.33**

$P_{hit}(S,CCL8) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,CCL8) = 0$ **ES(CCL8) = 1.0**

$P_{hit}(S,GRIN2A) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,GRIN2A) = 0 + \ldots + \frac{1}{15-10} = 0.2$

ES(CCL8) = 1.0 - 0.2 = 0.8

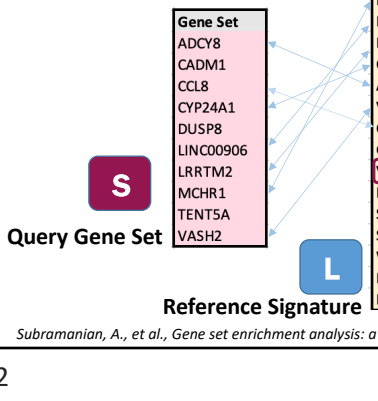
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

21

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g\,j \in S, j \le i} \frac{|r_j|^p}{N_R} \text{ where } N_R = \sum_{g\,j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g\,j \in S, j \le i} \frac{1}{N - N_H}$$
$$\text{where } N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

$P_{hit}(S,LINC00906) = \frac{4.8}{28.2} + \frac{4.4}{28.2} = 0.33$ $P_{miss}(S,LINC00906) = 0$ **ES(LINC00906) = 0.33**

$P_{hit}(S,CCL8) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,CCL8) = 0$ **ES(CCL8) = 1.0**

$P_{hit}(S,GRIN2A) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,GRIN2A) = 0 + \ldots + \frac{1}{15-10} = 0.2$

ES(CCL8) = 1.0 - 0.2 = 0.8

$P_{hit}(S,VCAM1) = ?$ $P_{miss}(S,GRIN2A) = ?$ **ES(CCL8) = ?**

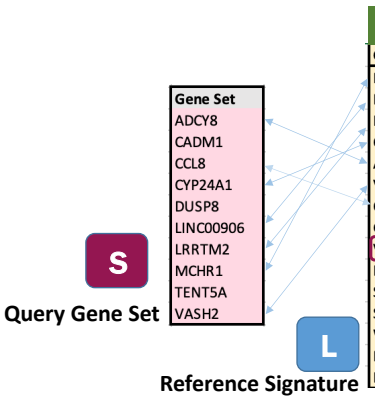
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

22

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974

$ES = P_{hit} - P_{miss}$

$$P_{hit}(S,i) = \sum_{g\,j \in S, j \le i} \frac{|r_j|^p}{N_R} \quad where\ N_R = \sum_{g\,j \in S} |r_j|^p$$

$$P_{miss}(S,i) = \sum_{g\,j \in S, j \le i} \frac{1}{N - N_H}$$
$$where\ N = |L_{gene}|, N_H = |S|$$

Let p=1, N_R=4.8 + 4.4 + 4.0 + 3.9 + 3.7 + 3.7 + 3.7 = 28.2, N=15, N_H=10

$P_{hit}(S,MCHR1) = \frac{4.8}{28.2} = 0.17$ $P_{miss}(S,MCHR1) = 0$ **ES(MCHR1) = 0.17**

$P_{hit}(S,LINC00906) = \frac{4.8}{28.2} + \frac{4.4}{28.2} = 0.33$ $P_{miss}(S,LINC00906) = 0$ **ES(LINC00906) = 0.33**

$P_{hit}(S,CCL8) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,CCL8) = 0$ **ES(CCL8) = 1.0**

$P_{hit}(S,GRIN2A) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,GRIN2A) = 0 + \ldots + \frac{1}{15-10} = 0.2$

ES(VCAM1) = 1.0 - 0.2 = 0.8

$P_{hit}(S,VCAM1) = \frac{4.8}{28.2} + \frac{4.4}{28.2} + \frac{4.0}{28.2} \ldots + \frac{3.7}{28.2} = 1.0$ $P_{miss}(S,VCAM1) = 0 + \ldots + 0.2 + 0.2 = 0.4$

ES(VCAM1) = 1.0 - 0.4 = 0.6

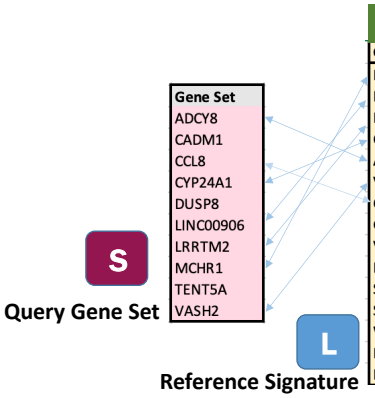
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

23

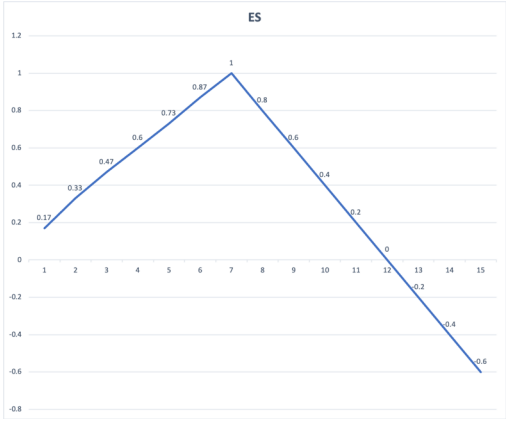
GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.



Ranked Gene List	
GENE	log2FoldChange
MCHR1	4.809435162
LINC00906	4.419081332
LRRTM2	4.069405614
CYP24A1	3.998899926
ADCY8	3.741715095
VASH2	3.732409393
CCL8	3.714507999
GRIN2A	3.709987257
VCAM1	-4.788212961
RAMP3	-4.817964833
SMTNL2	-4.820929664
SLITRK6	-5.107034602
WNT2	-6.501449488
FER1L6	-7.271737091
LRRC25	-11.06074974



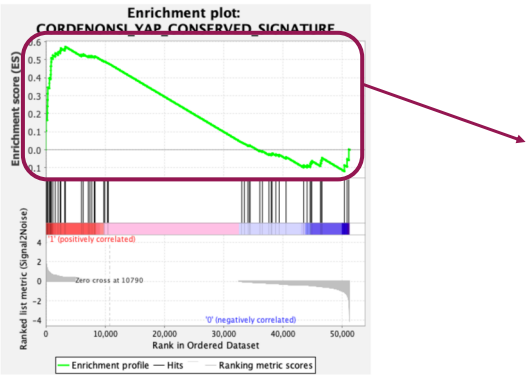
Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

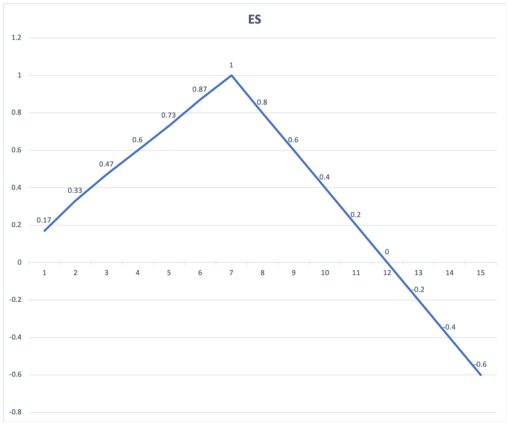
24

GSEA Method

Calculation of Enrichment Score (ES)

Enrichment score (*ES*) reflects the degree to which a set *S* is overrepresented at the extremes (top or bottom) of the entire ranked list *L*.





Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

25

GSEA Method

Statistical Scores

p-value

FDR

26

p-value

p-value is the probability that a random chance generated the data or something else that is equal or rarer (under the null hypothesis).

	TTHHH	TTTHH	
	THTHH	TTHTH	
HHHHH	THHTH	THTTH	TTTTH
	THHHT	HTTTH	TTTHT
THHHH	HTHHT	TTHHT	TTHTT
HTHHH	HTTHH	THTTT	HTTTT
HHTHH	HTHTH	THTHT	TTTTT
HHHTH	HTHHT	HTTHT	
HHHHT	HHTTH	THHTT	
	HHTHT	HTHTT	
	HHHTT	HHTTT	

At least a Head	
Data	Probability
5 Heads	1/32
4 Heads	5/32
3 Heads	10/32
2 Heads	10/32
1 Heads	5/32
0 Heads	0

At least a Tail	
Data	Probability
0 Tails	0
1 Tails	5/32
2 Tails	10/32
3 Tails	10/32
4 Tails	5/32
5 Tails	1/32

p-value of 5 Heads = P(5 Heads) + P(5 Tails) = 1/32 + 1/32 = 0.0625 Equal Prob.

p-value of 4 Heads and 1 Tail = P(4 Heads and 1Tail) + P(4 Tails and 1 Head) + P(5 Heads) + P(5 Tails) = 0.375 Equal + rarer Prob.

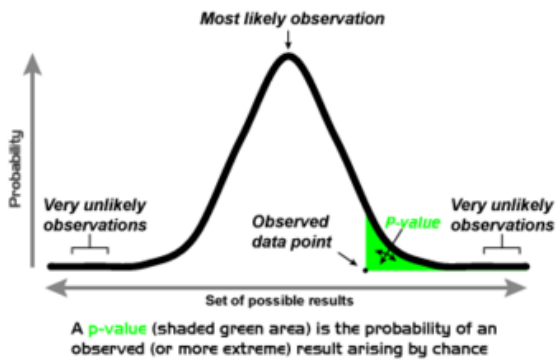
https://towardsdatascience.com/how-to-understand-p-value-in-layman-terms-80a5cc206ec2, https://www.youtube.com/watch?v=5Z9OIYA8He8

27

p-value

p-value is the probability that a random chance generated the data or something else that is equal or rarer (under the null hypothesis).

	TTHHH	TTTHH	
	THTHH	TTHTH	
HHHHH	THHTH	THTTH	TTTTH
	THHHT	HTTTH	TTTHT
THHHH	HTHHT	TTHHT	TTHTT
HTHHH	HTTHH	THTTT	HTTTT
HHTHH	HTHTH	THTHT	TTTTT
HHHTH	HTHHT	HTTHT	
HHHHT	HHTTH	THHTT	
	HHTHT	HTHTT	
	HHHTT	HHTTT	



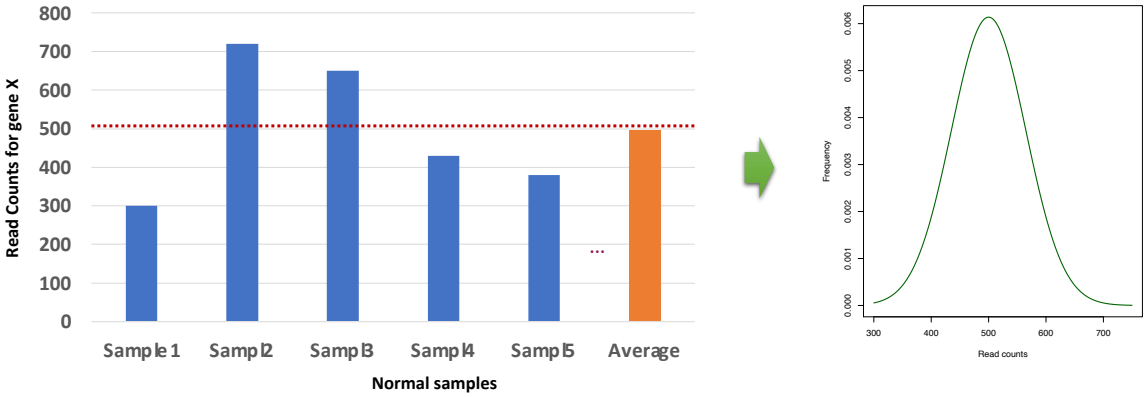
A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

https://towardsdatascience.com/how-to-understand-p-value-in-layman-terms-80a5cc206ec2

28

False Discovery Rate (FDR)

False Discovery Rates are a tool to weed out bad data that looks good!

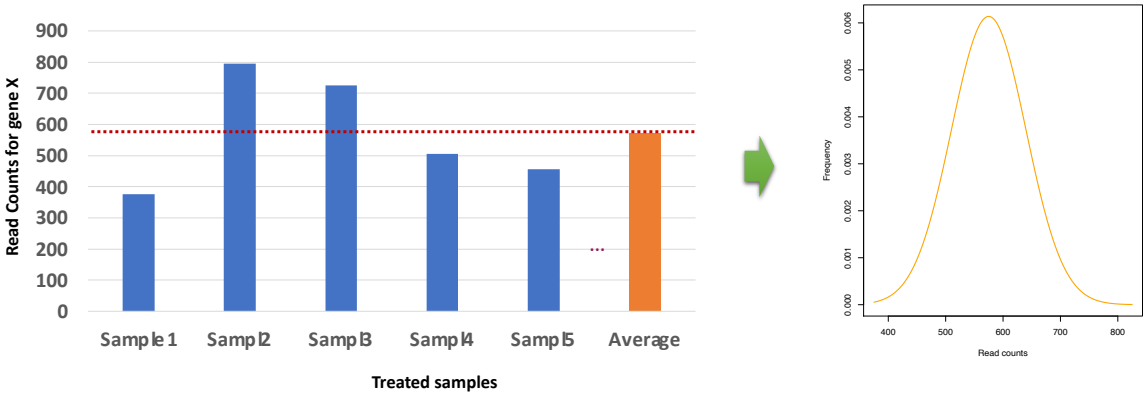


https://www.youtube.com/watch?v=K8LQSvtjcEo

29

False Discovery Rate (FDR)

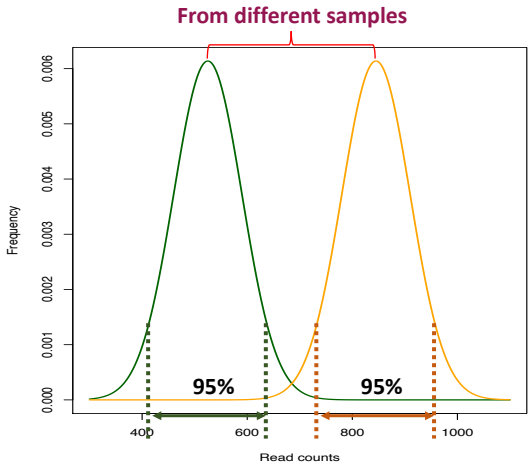
False Discovery Rates are a tool to weed out bad data that looks good!



https://www.youtube.com/watch?v=K8LQSvtjcEo

31

False Discovery Rate (FDR)



In general, if samples follows **normal distribution**, with random sampling, there will be **95%** chance of overlapping (**true positive**), and **5%** chances are not (**false positives**).

Q: 25,000 known genes and sequencing a same sample how may genes can be false positive?

25000 * 0.05 = 1,250

https://www.youtube.com/watch?v=K8LQSvtjcEo

33

False Discovery Rate (FDR)

False Discovery Rates are a tool to weed out bad data that looks good!

However, FDR is not a method to limit false positives, but the term is used interchangeably with the methods

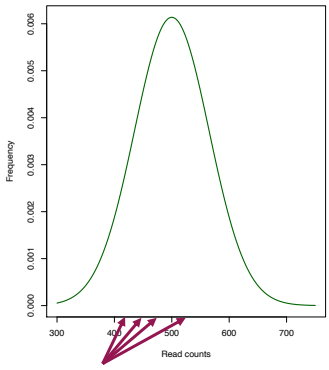
Benjamini-Hochberg method

Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

34

False Discovery Rate (FDR)

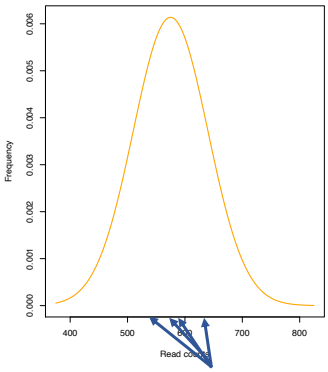
Normal



Run random sampling by taking 1,000 pairs of samples from same distribution and calculate p-values

Normal (p-value)	Test index	Treated (p-value)
0.91	**Test 1**	0.99
0.98	**Test 2**	0.92
:	**:**	:
0.96	**Test 1000**	0.97

Treated

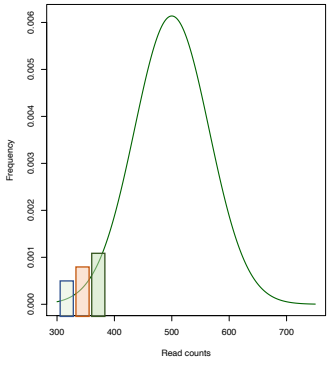


https://www.youtube.com/watch?v=K8LQSvtjcEo

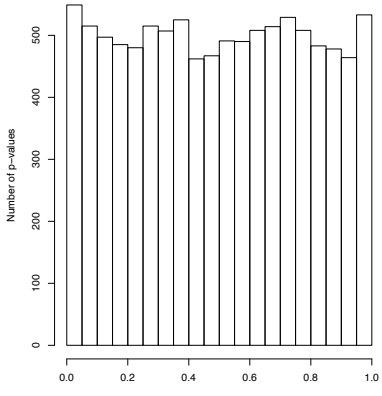
35

False Discovery Rate (FDR)

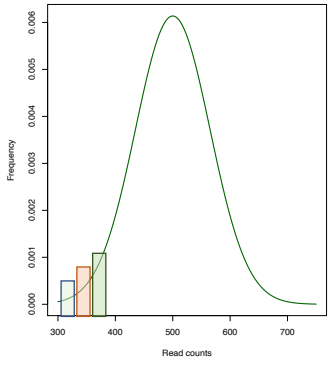
Normal



Regardless of sampling methods p-value distribution will be uniformly distributed



Normal

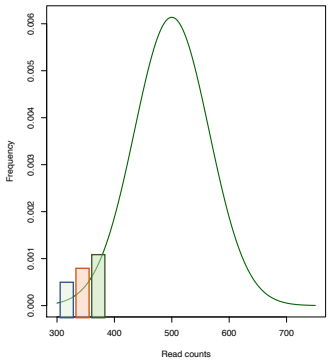


https://www.youtube.com/watch?v=K8LQSvtjcEo

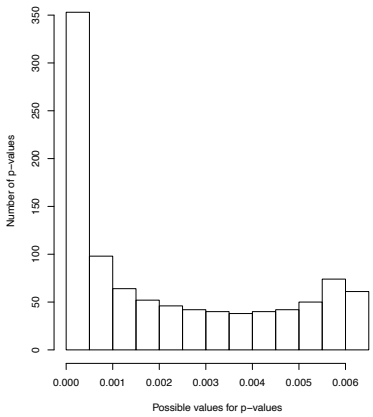
40

False Discovery Rate (FDR)

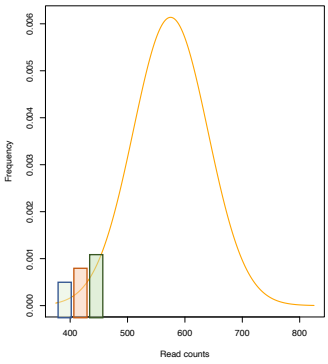
Normal



Regardless of sampling methods p-value distribution will be skewed



Treated

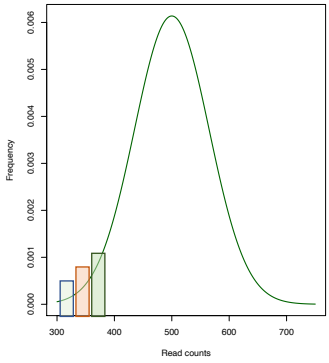


https://www.youtube.com/watch?v=K8LQSvtjcEo

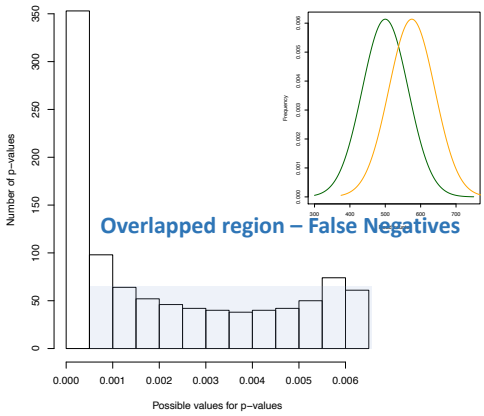
41

False Discovery Rate (FDR)

Normal

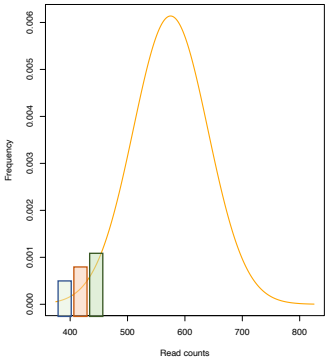


Regardless of sampling methods p-value distribution will be skewed



Overlapped region – False Negatives

Treated

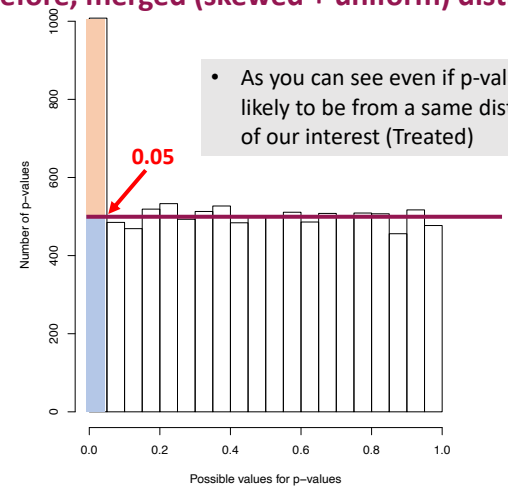


https://www.youtube.com/watch?v=K8LQSvtjcEo

42

False Discovery Rate (FDR)

Therefore, merged (skewed + uniform) distributions are expected



- As you can see even if p-value is low, there are samples likely to be from a same distribution (Normal) that is out of our interest (Treated)

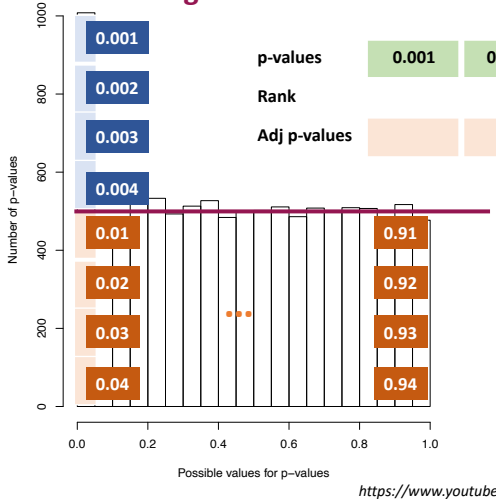
With eyeball measurement we can choose cut off to identify the "True Positives"

https://www.youtube.com/watch?v=K8LQSvtjcEo

43

False Discovery Rate (FDR)

Benjamini-Hochberg method



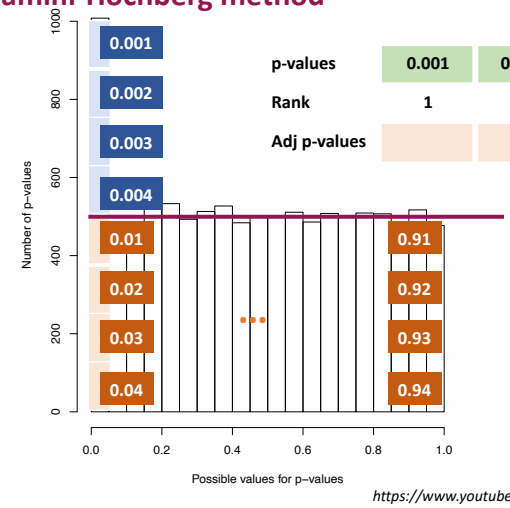
p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank								
Adj p-values								

https://www.youtube.com/watch?v=K8LQSvtjcEo

44

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	...	0.91	0.92	0.93	0.94
Rank	1	2	3	...	997	998	999	1000
Adj p-values								

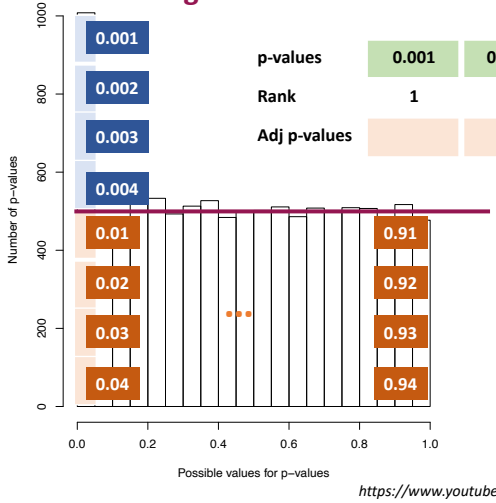
Step1: Sorting p-values in cut-off range

https://www.youtube.com/watch?v=K8LQSvtjcEo

45

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	...	0.91	0.92	0.93	0.94
Rank	1	2	3	...	997	998	999	1000
Adj p-values								0.94

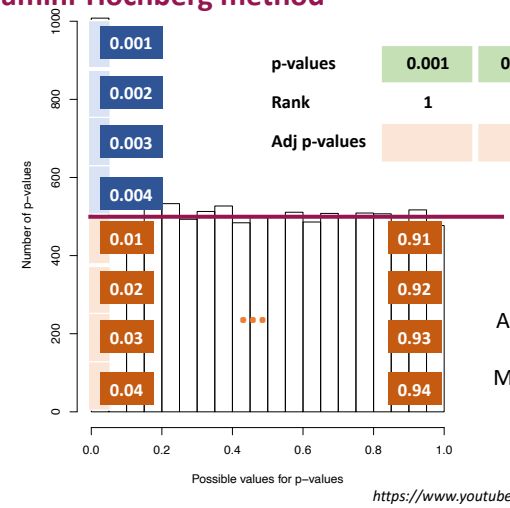
Step2: fill out the largest value

https://www.youtube.com/watch?v=K8LQSvtjcEo

46

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank	1	2	3	…	997	998	999	1000
Adj p-values							0.93	0.94

Step3: Calculate adjust p-value with the equation:

Adjusted p-value= $current\ p-value \times \frac{total\ \#\ of\ p-values}{p-value\ rank}$

Min(adjust p-value, previous adjusted p-value)

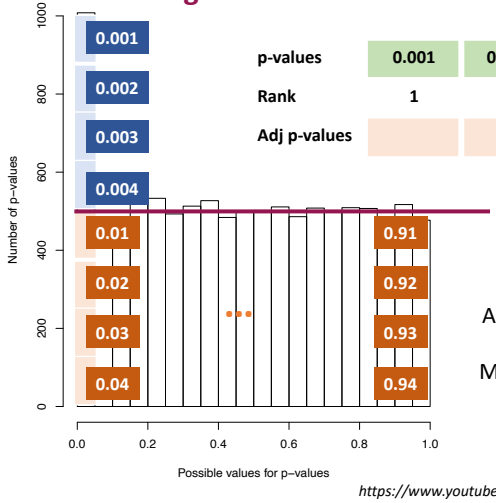
$$0.93 = 0.93 \times \frac{1000}{999}$$ ∴Min(0.93, 0.94) = 0.93

https://www.youtube.com/watch?v=K8LQSvtjcEo

47

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank	1	2	3	…	997	998	999	1000
Adj p-values						0.92	0.93	0.94

Step3: Adjust p-value with the equation:

Adjusted p-value= $current\ p-value \times \frac{total\ \#\ of\ p-values}{p-value\ rank}$

Min(adjust p-value, previous adjusted p-value)

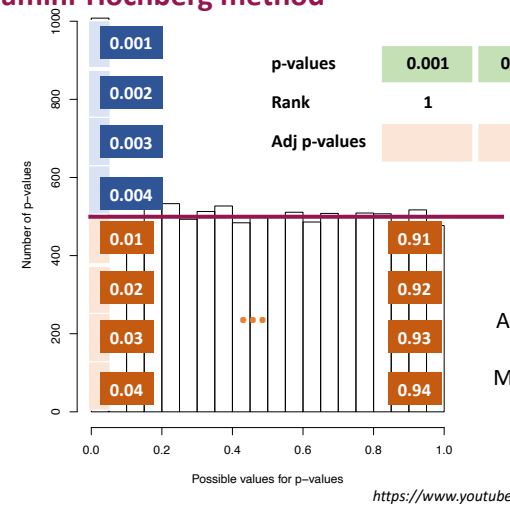
$$0.92 = 0.92 \times \frac{1000}{998}$$ ∴Min(0.92, 0.93) = 0.92

https://www.youtube.com/watch?v=K8LQSvtjcEo

48

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank	1	2	3	…	997	998	999	1000
Adj p-values			0.047		0.91	0.92	0.93	0.94

Step3: Adjust p-value with the equation:

Adjusted p-value= $current\ p-value \times \frac{total\ \#\ of\ p-values}{p-value\ rank}$

Min(adjust p-value, previous adjusted p-value)

$$1.0 = 0.003 \times \frac{1000}{3}$$

∴Min(1.0, 0.047) = 0.047

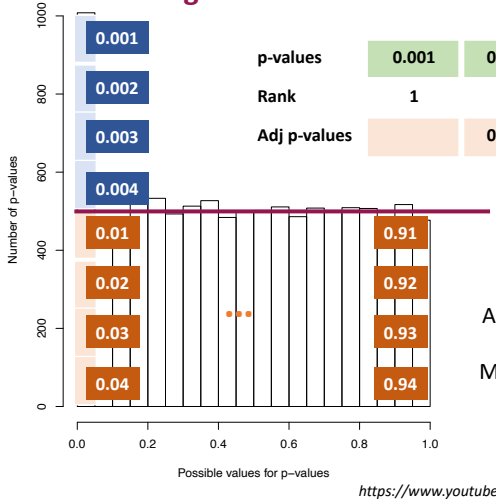
Let previous **adj p-value** is **0.047**

https://www.youtube.com/watch?v=K8LQSvtjcEo

49

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank	1	2	3	…	997	998	999	1000
Adj p-values		0.047	0.047		0.91	0.92	0.93	0.94

Step3: Adjust p-value with the equation:

Adjusted p-value= $current\ p-value \times \frac{total\ \#\ of\ p-values}{p-value\ rank}$

Min(adjust p-value, previous adjusted p-value)

$$1.0 = 0.002 \times \frac{1000}{2}$$

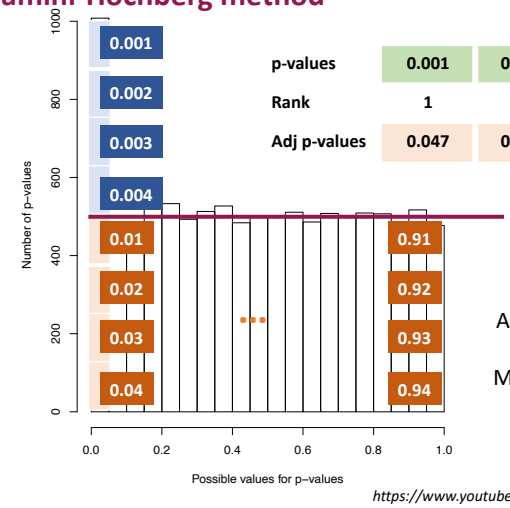
∴Min(1.0, 0.047) = 0.047

https://www.youtube.com/watch?v=K8LQSvtjcEo

50

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank	1	2	3	…	997	998	999	1000
Adj p-values	0.047	0.047	0.047		0.91	0.92	0.93	0.94

Step3: Adjust p-value with the equation:

Adjusted p-value= $current\ p-value \times \frac{total\ \#\ of\ p-values}{p-value\ rank}$

Min(adjust p-value, previous adjusted p-value)

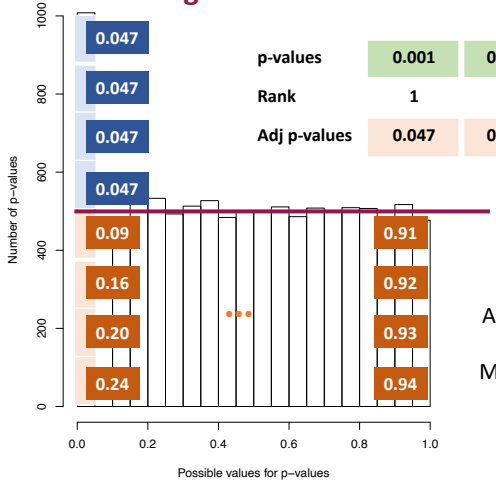
$$1.0 = 0.001 \times \frac{1000}{1}$$ ∴Min(1.0, 0.047) = 0.047

https://www.youtube.com/watch?v=K8LQSvtjcEo

51

False Discovery Rate (FDR)

Benjamini-Hochberg method



p-values	0.001	0.002	0.003	…	0.91	0.92	0.93	0.94
Rank	1	2	3	…	997	998	999	1000
Adj p-values	0.047	0.047	0.047		0.91	0.92	0.93	0.94

Step3: Adjust p-value with the equation:

Adjusted p-value= $current\ p-value \times \frac{total\ \#\ of\ p-values}{p-value\ rank}$

Min(adjust p-value, previous adjusted p-value)

$$1.0 = 0.001 \times \frac{1000}{1}$$ ∴Min(1.0, 0.047) = 0.047

https://www.youtube.com/watch?v=K8LQSvtjcEo

52

GSEA Prerequisites

Ordered gene lists (e.g., t-test, fold-change)

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

DESeq2 https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Gene sets (e.g., Molecular Signature Database)

MSigDB
Molecular Signatures Database

https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

53

Hands-on Practice

Install R and R Studio

https://cloud.r-project.org/

RStudio https://rstudio.com/products/rstudio/download/

54

Hands-on Practice

Install R packages

```
if (!requireNamespace("BiocManager", quietly = TRUE)) {
  install.packages("BiocManager")
}

bio.packages <- c("airway", "DESeq2")
new.packages <- bio.packages[!(bio.packages %in% installed.packages()[,"Package"])]
BiocManager::install(new.packages)
```

55

Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

56

Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```

Data is stored in SummarizedExperiment Class

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

57

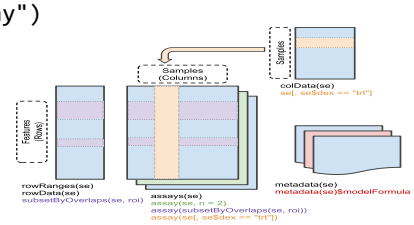
Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```



https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

58

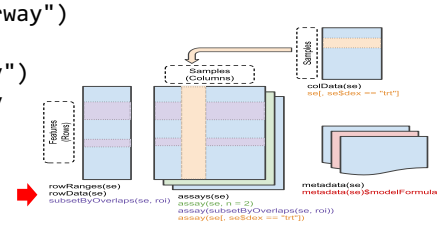
Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```



Rows are features (e.g., gene, transcripts, exons, etc)

```
> rowData(se)
DataFrame with 64102 rows and 0 columns
```

```
> names(se)
  [1] "ENSG00000000003" "ENSG00000000005"
 [10] "ENSG00000001167" "ENSG00000001460"
 [19] "ENSG00000001631" "ENSG00000002016"
 [28] "ENSG00000002746" "ENSG00000002822"
 [37] "ENSG00000003249" "ENSG00000003393"
 [46] "ENSG00000004059" "ENSG00000004139"
 [55] "ENSG00000004660" "ENSG00000004700"
 [64] "ENSG00000004846" "ENSG00000004848"
```

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

59

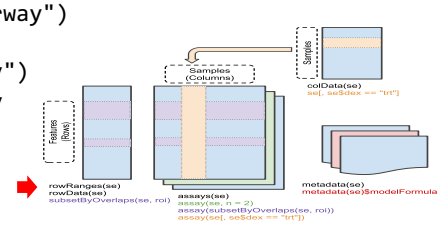
Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```




```
> rowRanges(se)
GRangesList object of length 64102:
$ENSG00000000003 Gene Name
GRanges object with 17 ranges and 2 metadata columns:
       seqnames              ranges strand |   exon_id       exon_name
          <Rle>           <IRanges>  <Rle> | <integer>     <character>
   [1]        X 99883667-99884983      - |    667145 ENSE00001459322
   [2]        X 99885756-99885863      - |    667146 ENSE00000868868
   [3]        X 99887482-99887565      - |    667147 ENSE00000401072
   [4]        X 99887538-99887565      - |    667148 ENSE00001849132
   [5]        X 99888402-99888536      - |    667149 ENSE00003554016
   ...      ...                 ...    ... .       ...             ...
  [13]        X 99890555-99890743      - |    667156 ENSE00003512331
  [14]        X 99891188-99891686      - |    667158 ENSE00001886883
  [15]        X 99891605-99891803      - |    667159 ENSE00001855382
  [16]        X 99891790-99892101      - |    667160 ENSE00001863395
  [17]        X 99894942-99894988      - |    667161 ENSE00001828996

...
<64101 more elements>
-------
seqinfo: 722 sequences (1 circular) from an unspecified genome
```

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

60

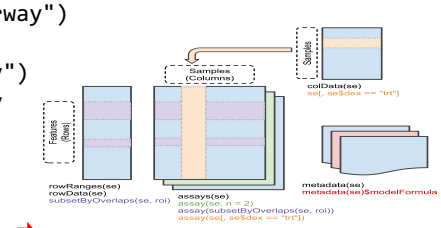
Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```




```
> rowRanges(se)$ENSG00000001460
GRanges object with 42 ranges and 2 metadata columns:
       seqnames            ranges strand |   exon_id       exon_name
          <Rle>         <IRanges>  <Rle> | <integer>     <character>
   [1]        1 24683489-24685109      - |     37300 ENSE00001463438
   [2]        1 24683490-24685109      - |     37301 ENSE00001727580
   [3]        1 24683495-24685109      - |     37303 ENSE00003692602
   [4]        1 24683495-24685109      - |     37302 ENSE00003468551
   [5]        1 24683527-24685109      - |     37304 ENSE00001777368
   ...      ...                 ...    ... .       ...             ...
  [38]        1 24740164-24743424      - |     37337 ENSE00001929731
  [39]        1 24741401-24741587      - |     37338 ENSE00001407060
  [40]        1 24741401-24741588      - |     37339 ENSE00001923991
  [41]        1 24742493-24742643      - |     37340 ENSE00001852206
  [42]        1 24742967-24743085      - |     37341 ENSE00001936284
  -------
  seqinfo: 722 sequences (1 circular) from an unspecified genome
```

Extract a gene (ENSG00000001460) information

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

61

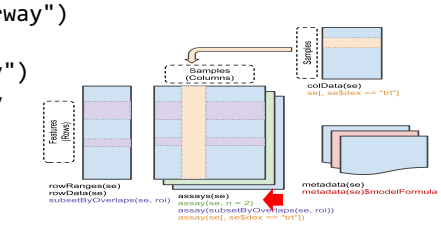
Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```



Assay is the result of experiments

```
> assays(se)
List of length 1
names(1): counts
```

> assays(se)

**** Be careful its plural***

This results contains 'counts' data only. Therefore following commands returns same results.

```
> assays(se)$counts
> assay(se)
```

```
> assays(se)$counts
                SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520 SRR1039521
ENSG00000000003        679        448        873        408       1138       1047        770        572
ENSG00000000005          0          0          0          0          0          0          0          0
ENSG00000000419        467        515        621        365        587        799        417        508
ENSG00000000457        260        211        263        164        245        331        233        229
ENSG00000000460         60         55         40         35         78         63         76         60
ENSG00000000938          0          0          2          0          1          0          0          0
ENSG00000000971       3251       3679       6177       4252       6721      11027       5176       7995
ENSG00000001036       1433       1062       1733        881       1424       1439       1359       1109
ENSG00000001084        519        380        595        493        820        714        696        704
```

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

62

Hands-on Practice

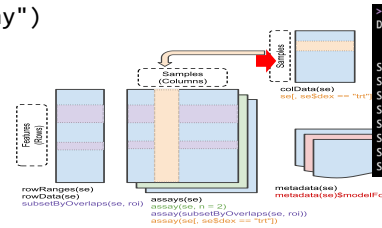
Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

Columns contain the information of samples

```
library("airway")

data("airway")
se <- airway
```

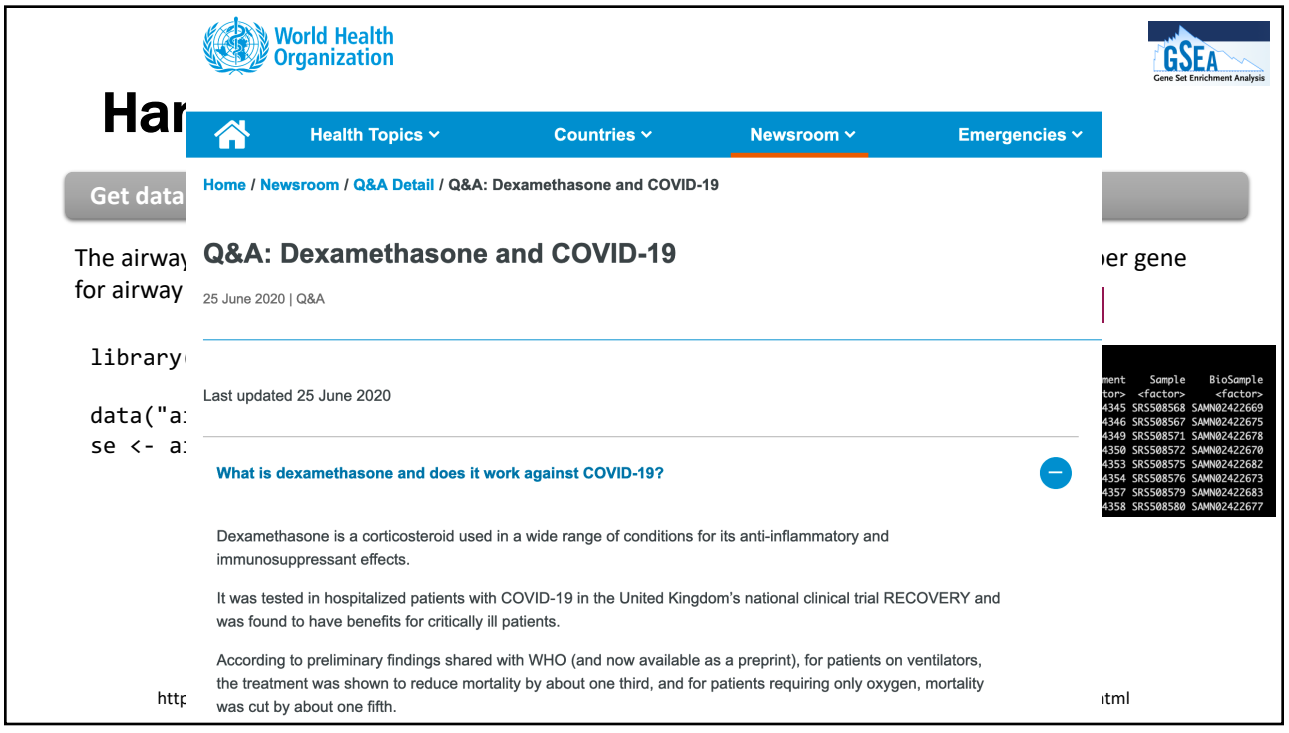



```
> colData(se)
DataFrame with 8 rows and 9 columns
             SampleName     cell      dex    albut        Run avgLength Experiment    Sample    BioSample
               <factor>  <factor> <factor> <factor>   <factor> <integer>   <factor>  <factor>     <factor>
SRR1039508 GSM1275862   N61311    untrt    untrt SRR1039508       126  SRX384345 SRS508568 SAMN02422669
SRR1039509 GSM1275863   N61311      trt    untrt SRR1039509       126  SRX384346 SRS508567 SAMN02422675
SRR1039512 GSM1275866  N052611    untrt    untrt SRR1039512       126  SRX384349 SRS508571 SAMN02422678
SRR1039513 GSM1275867  N052611      trt    untrt SRR1039513        87  SRX384350 SRS508572 SAMN02422670
SRR1039516 GSM1275870  N080611    untrt    untrt SRR1039516       120  SRX384353 SRS508575 SAMN02422682
SRR1039517 GSM1275871  N080611      trt    untrt SRR1039517       126  SRX384354 SRS508576 SAMN02422673
SRR1039520 GSM1275874  N061011    untrt    untrt SRR1039520       101  SRX384357 SRS508579 SAMN02422683
SRR1039521 GSM1275875  N061011      trt    untrt SRR1039521        98  SRX384358 SRS508580 SAMN02422677
```

cell line | Dexamethasone

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

63



64

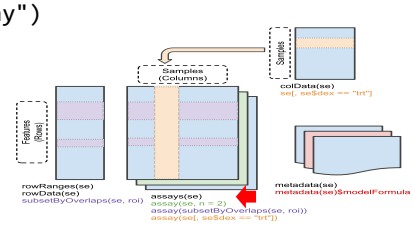
Hands-on Practice

Get data with AirWay Library

The airway package contains an example dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles.

```
library("airway")

data("airway")
se <- airway
```



Using the sample information select 'treat' and 'untreated' samples

```
> assays(se[, se$dex == 'trt'])$counts
                 SRR1039509 SRR1039513 SRR1039517 SRR1039521
ENSG00000000003        448        408       1047        572
ENSG00000000005          0          0          0          0
ENSG00000000419        515        365        799        508
ENSG00000000457        211        164        331        229
ENSG00000000460         55         35         63         60
ENSG00000000938          0          0          0          0
ENSG00000000971       3679       4252      11027       7995
ENSG00000001036       1062        881       1439       1109
ENSG00000001084        380        493        714        704
ENSG00000001167        236        175        584        269
ENSG00000001460        168        118        210        177
```

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

65

Hands-on Practice

Get data with AirWay Library and store them into CSV format

Get expression level data from AirWay and stored them into CSV files

```
library("airway")
library("DESeq2")

data("airway")
se <- airway

treated <- assays(se[, se$dex == 'trt'])$counts
untreat <- assays(se[, se$dex == 'untrt'])$counts
dfTreat <- as.data.frame(treated)
dfUntreat <- as.data.frame(untreat)

write.csv(dfTreat, file='treated.csv')
write.csv(dfUntreat, file='untreated.csv')
```

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

66

Hands-on Practice

Differential Expression

```
library("airway")
library("DESeq2")

data("airway")
se <- airway

ddsSE <- DESeqDataSet(se, design = ~ cell + dex)
```

Convert SE data into DEseq data object

```
> ddsSE
class: DESeqDataSet
dim: 64102 8
metadata(2): '' version
assays(1): counts
rownames(64102): ENSG00000000003 ENSG00000000005 ... LRG_98 LRG_99
rowData names(0):
colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
colData names(9): SampleName cell ... Sample BioSample
> assays(ddsSE)
List of length 1
names(1): counts
```

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

67

Hands-on Practice

Differential Expression

```
library("airway")
library("DESeq2")

data("airway")
se <- airway

ddsSE <- DESeqDataSet(se, design = ~ cell + dex)

keep <- rowSums(assays(ddsSE)$counts) >= 10
dds <- ddsSE[keep,]
```

Filtering small number of reads (i.e., less than 10 reads)

```
> dds
class: DESeqDataSet
dim: 22369 8
metadata(2): '' version
assays(1): counts
rownames(22369): ENSG00000000003 ENSG00000000419 ... ENSG00000273487 ENSG00000273488
rowData names(0):
colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
colData names(9): SampleName cell ... Sample BioSample
```

The number of genes is reduced from 64,102.

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

68

Hands-on Practice

Differential Expression

```
library("airway")
library("DESeq2")

data("airway")
se <- airway

ddsSE <- DESeqDataSet(se, design = ~ cell + dex)

keep <- rowSums(assays(ddsSE)$counts) >= 10
dds <- ddsSE[keep,]
```

Filtering small number of reads (i.e., less than 10 reads)

```
> dds
class: DESeqDataSet
dim: 22369 8
metadata(2): '' version
assays(1): counts
rownames(22369): ENSG00000000003 ENSG00000000419 ... ENSG00000273487 ENSG00000273488
rowData names(0):
colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
colData names(9): SampleName cell ... Sample BioSample
```

The number of genes is reduced from 64,102.

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

69

Hands-on Practice

Differential Expression

```
library("airway")
library("DESeq2")

data("airway")
se <- airway

ddsSE <- DESeqDataSet(se, design = ~ cell + dex)

keep <- rowSums(assays(ddsSE)$counts) >= 10
dds <- ddsSE[keep,]

dds <- DESeq(dds)
res <- results(dds)
resSig <- subset(res, padj <= 0.05)
```

Run DESeq and filter out with p-value <= 0.05

https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

70

Hands-on Practice

Differential Expression https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf

```
library("airway")
library("DESeq2")

data("airway")
se <- airway

ddsSE <- DESeqDataSet(se, design = ~ cell + dex)

keep <- rowSums(assays(ddsSE)$counts) >= 10
dds <- ddsSE[keep,]

dds <- DESeq(dds)
res <- results(dds)
resSig <- subset(res, padj <= 0.05)

dfRes <- as.data.frame(resSig)
resOrder <- dfRes[order(dfRes $log2FoldChange, decreasing=TRUE),]
```

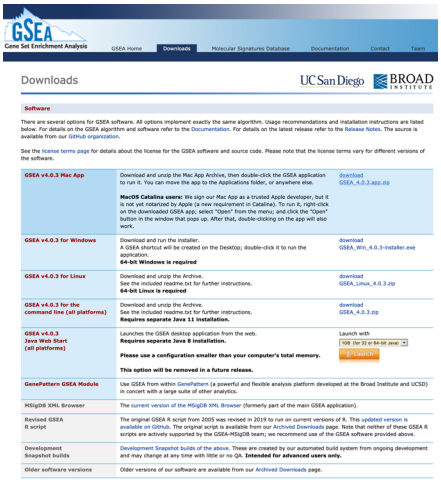
Order results based on log2FoldChange

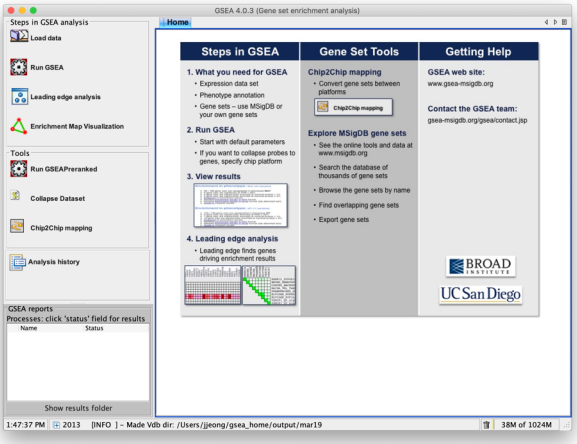
https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

71

Hands-on Practice

GSEA Installation https://www.gsea-msigdb.org/gsea/downloads.jsp

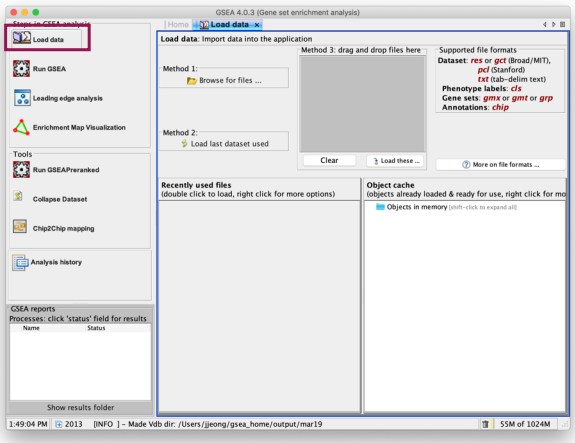




72

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/downloads.jsp



Tab delimited Expression Data

NAME	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	679	448	873	408	1138	1047	770	572
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	467	515	621	365	587	799	417	508
ENSG00000000457	260	211	263	164	245	331	233	229
ENSG00000000460	60	55	40	35	78	63	76	60
ENSG00000000938	0	0	2	0	1	0	0	0
ENSG00000000971	3251	3679	6177	4252	6721	11027	5176	7995
ENSG00000001036	1433	1062	1733	881	1424	1439	1359	1109
ENSG00000001084	519	380	595	493	820	714	696	704
ENSG00000001167	394	236	464	175	658	584	360	269
ENSG00000001460	172	168	264	118	241	210	155	177
ENSG00000001461	2112	1867	5137	2657	2735	2751	2467	2905
ENSG00000001497	524	488	638	357	676	806	493	475
ENSG00000001561	71	51	211	156	23	38	134	172
ENSG00000001617	555	394	905	415	727	697	618	599
ENSG00000001626	10	2	9	2	10	6	5	5
ENSG00000001629	1660	1251	2259	1079	2462	2514	1888	1660
ENSG00000001630	59	54	66	23	84	87	31	59
ENSG00000001631	729	692	943	475	1034	1163	731	744
ENSG00000002016	201	161	256	99	268	257	160	137
ENSG00000002079	3	0	3	1	4	0	0	1
ENSG00000002330	206	174	184	111	194	260	156	177
ENSG00000002549	1459	1294	1317	998	1451	1824	853	1031
ENSG00000002586	7507	7203	9501	6214	10973	12863	6834	7225
ENSG00000002587	2	0	1	0	0	2	0	0
ENSG00000002726	0	0	1	0	0	0	0	0

store all samples into a file "treat_untreat.txt"

Available Data Formats http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

73

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/downloads.jsp

Tab delimited Expression Data

NAME	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	679	448	873	408	1138	1047	770	572
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	467	515	621	365	587	799	417	508
ENSG00000000457	260	211	263	164	245	331	233	229
ENSG00000000460	60	55	40	35	78	63	76	60
ENSG00000000938	0	0	2	0	1	0	0	0
ENSG00000000971	3251	3679	6177	4252	6721	11027	5176	7995
ENSG00000001036	1433	1062	1733	881	1424	1439	1359	1109
ENSG00000001084	519	380	595	493	820	714	696	704
ENSG00000001167	394	236	464	175	658	584	360	269
ENSG00000001460	172	168	264	118	241	210	155	177
ENSG00000001461	2112	1867	5137	2657	2735	2751	2467	2905
ENSG00000001497	524	488	638	357	676	806	493	475
ENSG00000001561	71	51	211	156	23	38	134	172
ENSG00000001617	555	394	905	415	727	697	618	599
ENSG00000001626	10	2	9	2	10	6	5	5
ENSG00000001629	1660	1251	2259	1079	2462	2514	1888	1660
ENSG00000001630	59	54	66	23	84	87	31	59
ENSG00000001631	729	692	943	475	1034	1163	731	744
ENSG00000002016	201	161	256	99	268	257	160	137
ENSG00000002079	3	0	3	1	4	0	0	1
ENSG00000002330	206	174	184	111	194	260	156	177
ENSG00000002549	1459	1294	1317	998	1451	1824	853	1031
ENSG00000002586	7507	7203	9501	6214	10973	12863	6834	7225
ENSG00000002587	2	0	1	0	0	2	0	0
ENSG00000002726	0	0	1	0	0	0	0	0

store all samples into a file "treat_untreat.txt"

CLS: Categorical (e.g tumor vs normal) class file format (*.cls)

Total # of samples

Total # of categories

Always '1'

8		2		1				
# Untreated	Treated							
0	1	0	1	0	1	0	1	

"# Untreated Treated". The key point is that as the third line is processed left-to-right, it will take the first label it finds **no matter what it is** and map it to the first class name from the second line (also left-to-right). Any other instances of that label then map to that same name. After that, the second label found (on the third line) **different from the first** is mapped to the second name (on the second line), and likewise for any other instances.

Available Data Formats http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

74

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/downloads.jsp

Tab delimited Expression Data

NAME	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	679	448	873	408	1138	1047	770	572
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	467	515	621	365	587	799	417	508
ENSG00000000457	260	211	263	164	245	331	233	229
ENSG00000000460	60	55	40	35	78	63	76	60
ENSG00000000938	0	0	2	0	1	0	0	0
ENSG00000000971	3251	3679	6177	4252	6721	11027	5176	7995
ENSG00000001036	1433	1062	1733	881	1424	1439	1359	1109
ENSG00000001084	519	380	595	493	820	714	696	704
ENSG00000001167	394	236	464	175	658	584	360	269
ENSG00000001460	172	168	264	118	241	210	155	177
ENSG00000001461	2112	1867	5137	2657	2735	2761	2467	2905
ENSG00000001497	524	488	638	357	676	806	493	475
ENSG00000001561	71	51	211	156	23	38	134	172
ENSG00000001617	555	394	905	415	727	697	618	599
ENSG00000001626	10	2	9	2	10	6	5	5
ENSG00000001629	1660	1251	2259	1079	2462	2514	1888	1660
ENSG00000001630	59	54	66	23	84	87	31	59
ENSG00000001631	729	692	943	475	1034	1163	731	744
ENSG00000002016	201	161	256	99	268	257	160	137
ENSG00000002079	3	0	3	1	4	0	0	1
ENSG00000002330	206	174	184	111	194	260	156	177
ENSG00000002549	1459	1294	1317	998	1451	1824	853	1031
ENSG00000002586	7507	7203	9501	6214	10973	12863	6834	7225
ENSG00000002587	2	0	1	0	0	2	0	0
ENSG00000002726	0	0	1	0	0	0	0	0

store all samples into a file "treat_untreat.txt"

CLS: Categorical (e.g tumor vs normal) class file format (*.cls)

Total # of samples

Total # of categories

Always '1'

8	2	1					
# Untreated	Treated						
0	1	0	1	0	1	0	1

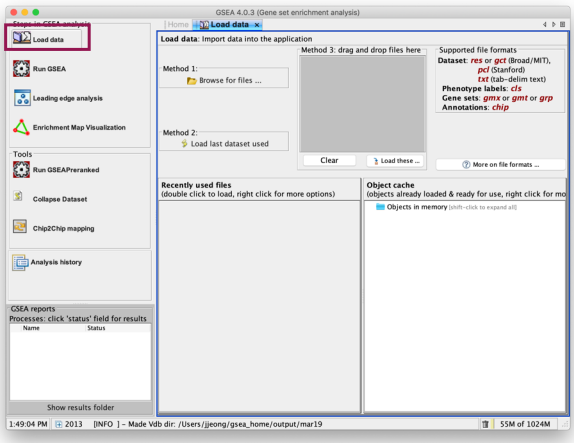
Save as "untreat_treat.cls"

Available Data Formats http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

75

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/downloads.jsp

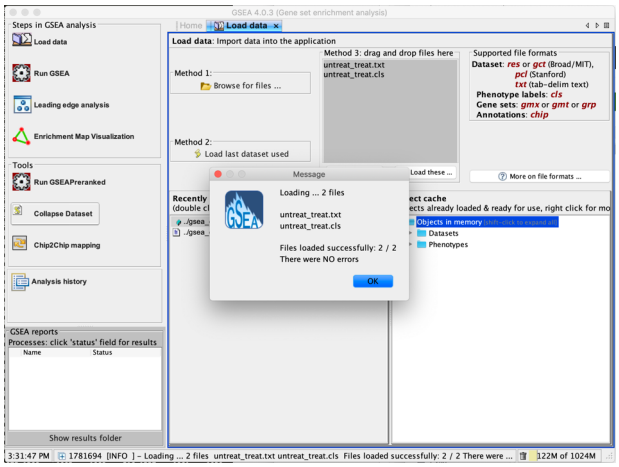


1. Click Load data

76

Hands-on Practice

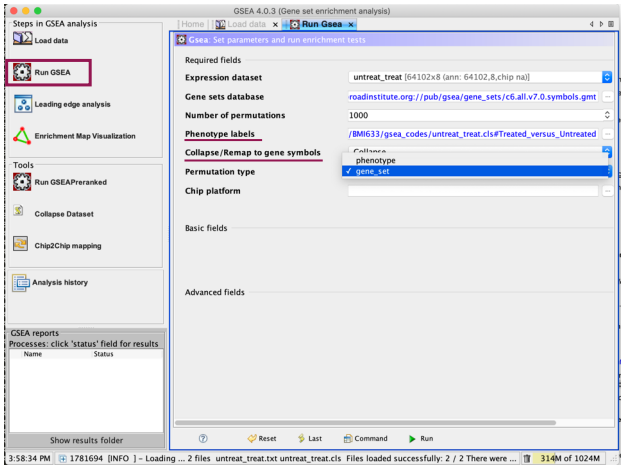
Run GSEA https://www.gsea-msigdb.org/gsea/downloads.jsp



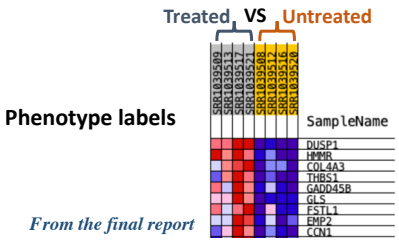
1. Click Load data
2. Drag and drop data into “Method3”

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_Run_GSEA_Page



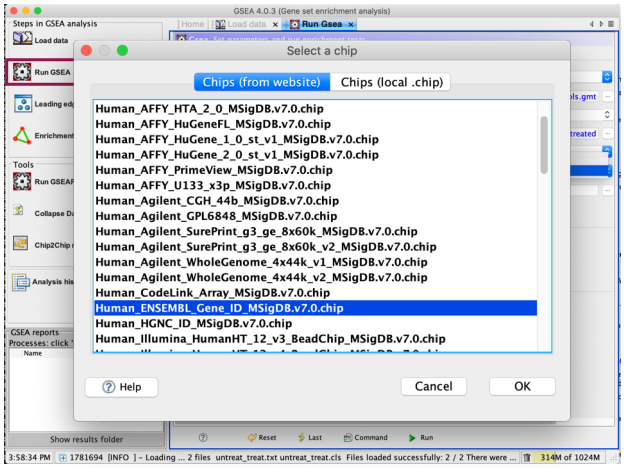
1. Click Load data
2. Drag and drop data into “Method3”
3. Run GSEA
   - Choose “No_Collapse” if using gene name in the data
   - If there are at least seven (7) samples in each phenotype then choose “Phenotype” otherwise choose “gene_set”



From the final report

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_Run_GSEA_Page

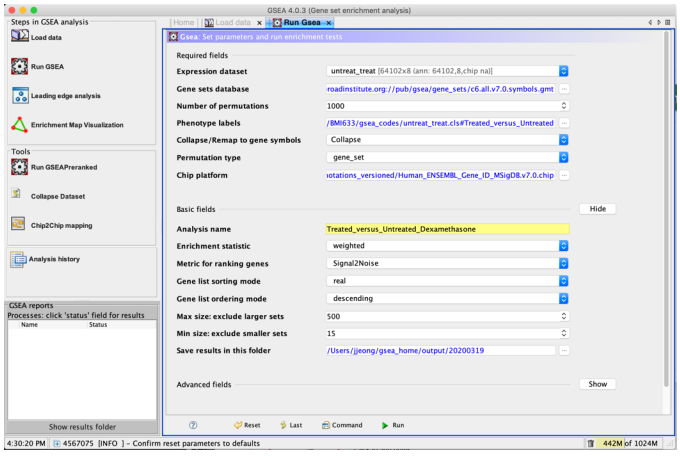


1. Click Load data
2. Drag and drop data into “Method3”
3. Run GSEA
   - Choose “No_Collapse” if using gene name in the data
   - If there are at least seven (7) samples in each phenotype then choose “Phenotype” otherwise choose “gene_set”
   - Chip platform: Human_ENSEMBL_Gene_ID_MSigDB…

79

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_Run_GSEA_Page

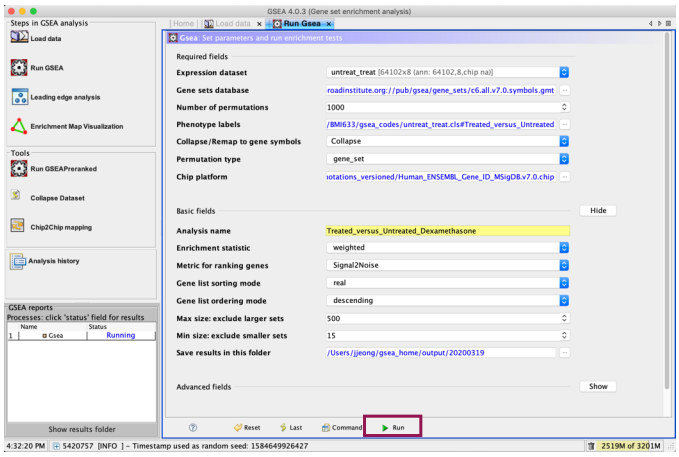


1. Click Load data
2. Drag and drop data into “Method3”
3. Run GSEA
   - Choose “No_Collapse” if using gene name in the data
   - If there are at least seven (7) samples in each phenotype then choose “Phenotype” otherwise choose “gene_set”
   - Chip platform: Human_ENSEMBL_Gene_ID_MSigDB…
   - **Basic fields:**
     - Signal2Noise: at least two categorical phenotypes and your expression dataset must contain at least three (3) samples for each phenotype.

80

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_Run_GSEA_Page

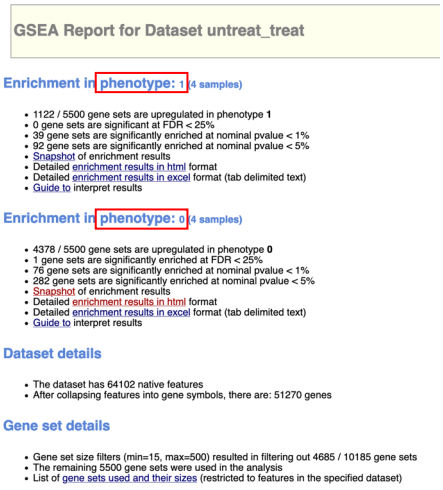


1. Click Load data
2. Drag and drop data into "Method3"
3. Run GSEA
   - Choose "No_Collapse" if using gene name in the data
   - If there are at least seven (7) samples in each phenotype then choose "Phenotype" otherwise choose "gene_set"
   - Chip platform: Human_ENSEMBL_Gene_ID_MSigDB…
   - **Basic fields:**
     - Signal2Noise: at least two categorical phenotypes and your expression dataset must contain at least three (3) samples for each phenotype.
4. Run

81

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_Run_GSEA_Page


GSEA Report for Dataset untreat_treat

Enrichment in phenotype: 1 (4 samples)

- 1122 / 5500 gene sets are upregulated in phenotype **1**
- 0 gene sets are significant at FDR < 25%
- 39 gene sets are significantly enriched at nominal pvalue < 1%
- 92 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: 0 (4 samples)

- 4378 / 5500 gene sets are upregulated in phenotype **0**
- 1 gene sets are significantly enriched at FDR < 25%
- 76 gene sets are significantly enriched at nominal pvalue < 1%
- 282 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 64102 native features
- After collapsing features into gene symbols, there are: 51270 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 4685 / 10185 gene sets
- The remaining 5500 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

1. Click Load data
2. Drag and drop data into "Method3"
3. Run GSEA
   - Choose "No_Collapse" if using gene name in the data
   - If there are at least seven (7) samples in each phenotype then choose "Phenotype" otherwise choose "gene_set"
   - Chip platform: Human_ENSEMBL_Gene_ID_MSigDB…
   - **Basic fields:**
     - Signal2Noise: at least two categorical phenotypes and your expression dataset must contain at least three (3) samples for each phenotype.
4. Run
5. Go to output directory and open "index.html"

Phenotype **1** (treated) vs Phenotype **0** (untreated)

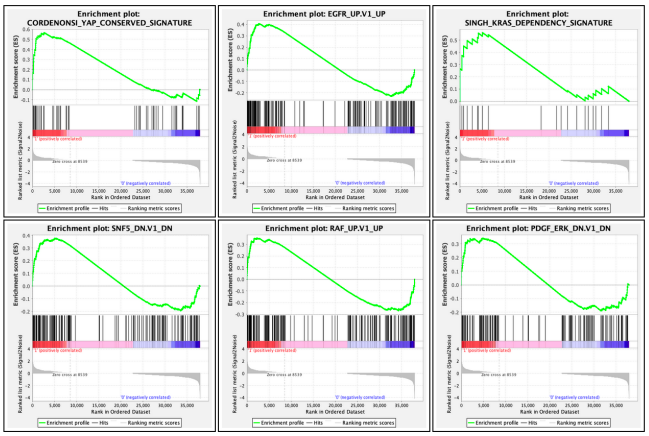
(+ value) (- values)

82

Hands-on Practice

Run GSEA https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_Run_GSEA_Page

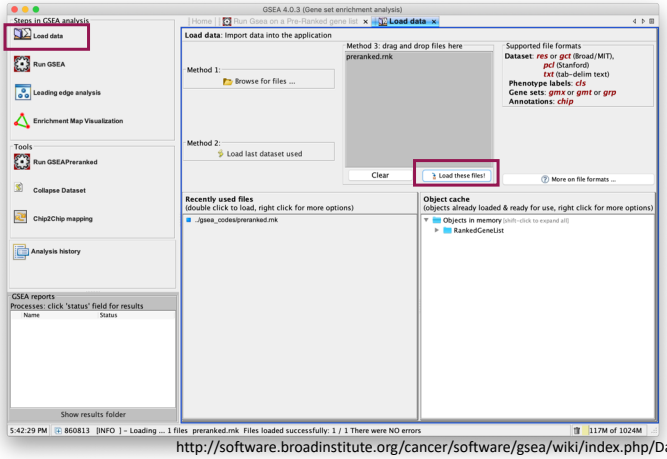
Table: Snapshot of enrichment results



83

Hands-on Practice

Run GSEA Preranked https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_GSEAPreranked_Page



1. Prepare ranked gene list file (*.rnk)

Comments

Tab delimited ranked list with statistic values

```
# hgnc_symbol    log2FoldChange
# treated_vs_untreated
MCHR1    4.809435162
LINC00906    4.419081332
LRRTM2   4.069405614
CYP24A1 3.998899926
ADCY8    3.741715095
VASH2    3.732409393
CCL8     3.714507999
GRIN2A   3.709987257
TLDC2    -3.973738941
SLC16A12     -4.018680095
TSPAN8   -4.193009497
RASL11B -4.395220702
KLF15    -4.433733615
LGI3     -4.471955285
SPARCL1 -4.55996944
PRODH    -4.776927499
FAM107A -4.788212961
SERTM2   -4.817964833
ANGPTL7 -4.820929664
STEAP4   -5.107034602
GUCY2D   -6.501449488
ZBTB16   -7.271737091
AL0X15B -11.06074974
```

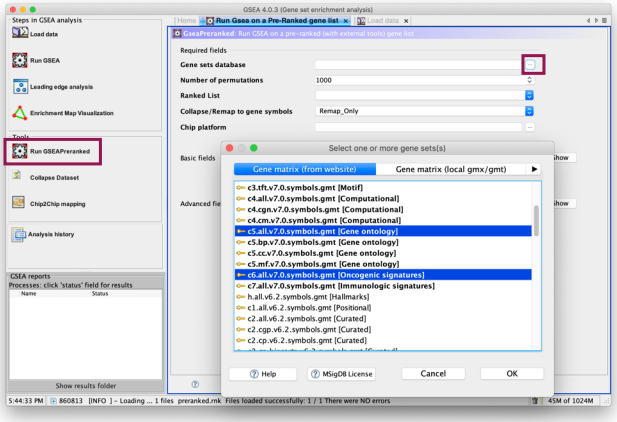
Saved as "preranked.rnk"

http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29

84

Hands-on Practice

Run GSEA Preranked https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_GSEAPreranked_Page



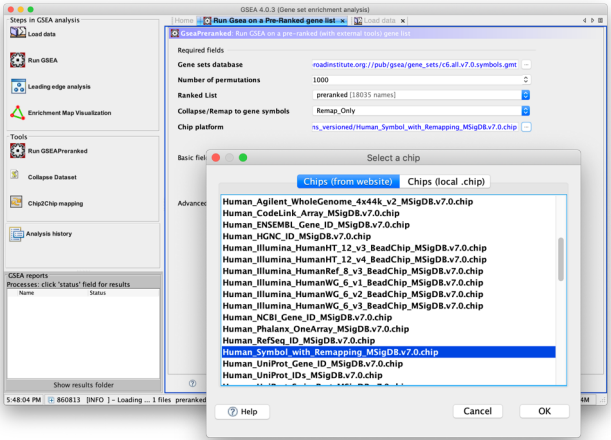
1. Prepare ranked gene list file (*.rnk)
2. Click GSEAPreranked
3. Select Gene Set

http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29

85

Hands-on Practice

Run GSEA Preranked https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_GSEAPreranked_Page



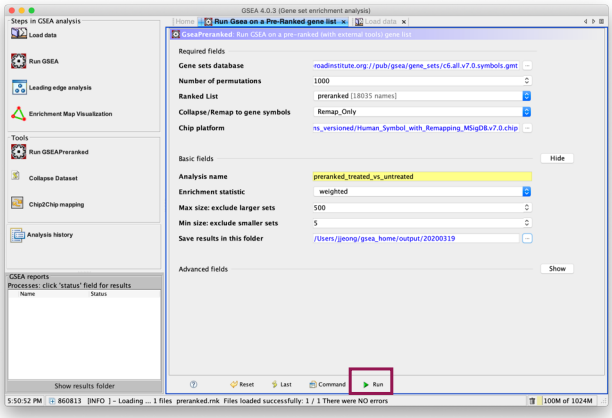
1. Prepare ranked gene list file (*.rnk)
2. Click GSEAPreranked
3. Select Gene Set
4. Select Chip platform
   - "Human_Symbol_with_Remapping_MSigDB…."

http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29

86

Hands-on Practice

Run GSEA Preranked https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_GSEAPreranked_Page



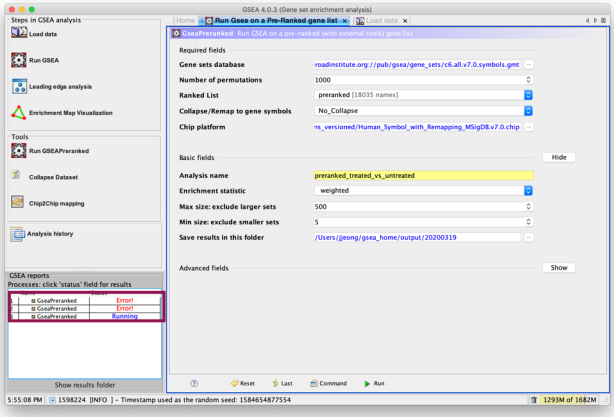
1. Prepare ranked gene list file (*.rnk)
2. Click GSEAPreranked
3. Select Gene Set
4. Select Chip platform
   - "Human_Symbol_with_Remapping_MSigDB…."
5. Fill out the Basic fields
6. Run

http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29

87

Hands-on Practice

Run GSEA Preranked https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_GSEAPreranked_Page



1. Prepare ranked gene list file (*.rnk)
2. Click GSEAPreranked
3. Select Gene Set
4. Select Chip platform
   - "Human_Symbol_with_Remapping_MSigDB…."
5. Fill out the Basic fields
6. Run
7. If errors occurs, correct the errors and run again

http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29

88

Hands-on Practice

Run GSEA Preranked https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideTEXT.htm#_GSEAPreranked_Page

GSEA Report for Dataset preranked

Enrichment in phenotype: na

- 4054 / 8973 gene sets are upregulated in phenotype **na_pos**
- 99 gene sets are significant at FDR < 25%
- 166 gene sets are significantly enriched at nominal pvalue < 1%
- 456 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

- 4919 / 8973 gene sets are upregulated in phenotype **na_neg**
- 21 gene sets are significantly enriched at FDR < 25%
- 181 gene sets are significantly enriched at nominal pvalue < 1%
- 518 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

1. Prepare ranked gene list file (*.rnk)
2. Click GSEAPreranked
3. Select Gene Set
4. Select Chip platform
   - "Human_Symbol_with_Remapping_MSigDB…."
5. Fill out the Basic fields
6. Run
7. If errors occurs, correct the errors and run again
8. If success, go to output directory and open "index.htmL"

GSEA doesn't know what the meaning of positive and negative ranked values, but you already know.

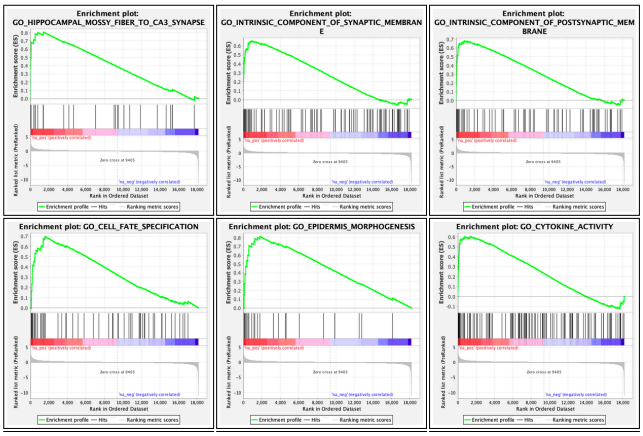
http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29

89

Hands-on Practice

Run GSEA Preranked

Table: Snapshot of enrichment results



90

Thank you

Jong Cheol (JC) Jeong

JongCheol.Jeong@uky.edu

91