

Gene Set Enrichment Analysis

february 2020

MARGenomics



Gene Set Enrichment Analysis (GSEA)

List of genes with statistics (p.Value and logFC)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	Symbol	FC: Tumor1 vs Ctrl	logFC: Tumor1 vs Ctrl	P-Value: Tumor1 vs Ctrl	adj. P-Val.: Tumor1 vs Ctrl	mean Ctrl	mean LINC	CONTROL 1.RMA	CONTROL 3.RMA	CONTROL 4.RMA	CONTROL 5.RMA	CONTROL 6.RMA	Tumoral1.RMA	Tumoral2.RMA	Tumoral3.RMA	Tumoral4.RMA	Tumoral5.RMA
1	FAM231D	-2.02	-1.02	0.00000	0.01856	7.25	6.23	7.11	7.46	7.37	7.38	6.91	6.39	6.31	6.10	6.00	6.35
2	CHD1	-1.97	-0.98	0.00000	0.02406	7.16	6.18	6.97	7.42	7.34	7.01	7.06	6.25	6.36	5.90	6.00	6.37
3	DOCK5	-1.66	-0.73	0.00012	0.41294	5.04	4.31	5.25	5.11	5.23	4.95	4.69	4.46	4.18	4.51	4.45	3.97
4	XG	-1.58	-0.66	0.00056	0.58249	7.10	6.45	7.09	7.25	7.09	6.88	7.19	6.58	6.58	6.81	6.29	5.96
5	DST	-1.57	-0.65	0.00140	0.81165	5.68	5.03	5.71	5.98	5.48	5.80	5.41	4.81	4.62	5.07	5.16	5.48
6	MPC1L	-1.44	-0.53	0.00008	0.34905	4.58	4.05	4.39	4.64	4.71	4.43	4.71	4.03	4.04	3.86	4.10	4.20
7	BLOC1S1F	-1.43	-0.51	0.00766	0.98116	6.24	5.73	6.02	6.41	6.42	5.96	6.37	6.14	5.82	5.26	5.51	5.91
8	MKRN1	-1.41	-0.49	0.03340	0.98116	6.42	5.92	6.38	6.61	6.26	6.61	6.22	6.18	5.63	5.73	5.40	6.67
9	SNUPN	-1.41	-0.49	0.00041	0.53985	5.59	5.10	5.71	5.63	5.69	5.34	5.57	4.87	4.97	5.23	5.31	5.11
10	RPL36A	-1.40	-0.49	0.00002	0.11241	7.25	6.76	7.29	7.22	7.22	7.20	7.30	6.67	6.80	6.63	6.94	6.75
11	SERF2	-1.40	-0.48	0.00538	0.98116	5.22	4.74	5.59	5.09	5.39	5.20	4.86	4.80	4.56	4.69	4.55	5.11
12	DHRS4L2	-1.39	-0.48	0.00133	0.81165	4.93	4.45	4.79	5.03	5.16	4.69	4.99	4.51	4.61	4.18	4.33	4.65
13	LGALS14	-1.39	-0.48	0.00251	0.86826	4.20	3.72	4.05	4.16	4.05	4.26	4.47	3.67	3.86	3.81	3.30	3.96

Pre-ranked analysis

Gene Expression matrix (microarrays)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	Symbol	FC: Tumor1 vs Ctrl	logFC: Tumor1 vs Ctrl	P-Value: Tumor1 vs Ctrl	adj. P-Val.: Tumor1 vs Ctrl	mean Ctrl	mean LINC	CONTROL 1.RMA	CONTROL 3.RMA	CONTROL 4.RMA	CONTROL 5.RMA	CONTROL 6.RMA	Tumoral1.RMA	Tumoral2.RMA	Tumoral3.RMA	Tumoral4.RMA	Tumoral5.RMA
1	FAM231D	-2.02	-1.02	0.00000	0.01856	7.25	6.23	7.11	7.46	7.37	7.38	6.91	6.39	6.31	6.10	6.00	6.35
2	CHD1	-1.97	-0.98	0.00000	0.02406	7.16	6.18	6.97	7.42	7.34	7.01	7.06	6.25	6.36	5.90	6.00	6.37
3	DOCK5	-1.66	-0.73	0.00012	0.41294	5.04	4.31	5.25	5.11	5.23	4.95	4.69	4.46	4.18	4.51	4.45	3.97
4	XG	-1.58	-0.66	0.00056	0.58249	7.10	6.45	7.09	7.25	7.09	6.88	7.19	6.58	6.58	6.81	6.29	5.96
5	DST	-1.57	-0.65	0.00140	0.81165	5.68	5.03	5.71	5.98	5.48	5.80	5.41	4.81	4.62	5.07	5.16	5.48
6	MPC1L	-1.44	-0.53	0.00008	0.34905	4.58	4.05	4.39	4.64	4.71	4.43	4.71	4.03	4.04	3.86	4.10	4.20
7	BLOC1S1F	-1.43	-0.51	0.00766	0.98116	6.24	5.73	6.02	6.41	6.42	5.96	6.37	6.14	5.82	5.26	5.51	5.91
8	MKRN1	-1.41	-0.49	0.03340	0.98116	6.42	5.92	6.38	6.61	6.26	6.61	6.22	6.18	5.63	5.73	5.40	6.67
9	SNUPN	-1.41	-0.49	0.00041	0.53985	5.59	5.10	5.71	5.63	5.69	5.34	5.57	4.87	4.97	5.23	5.31	5.11
10	RPL36A	-1.40	-0.49	0.00002	0.11241	7.25	6.76	7.29	7.22	7.22	7.20	7.30	6.67	6.80	6.63	6.94	6.75
11	SERF2	-1.40	-0.48	0.00538	0.98116	5.22	4.74	5.59	5.09	5.39	5.20	4.86	4.80	4.56	4.69	4.55	5.11
12	DHRS4L2	-1.39	-0.48	0.00133	0.81165	4.93	4.45	4.79	5.03	5.16	4.69	4.99	4.51	4.61	4.18	4.33	4.65
13	LGALS14	-1.39	-0.48	0.00251	0.86826	4.20	3.72	4.05	4.16	4.05	4.26	4.47	3.67	3.86	3.81	3.30	3.96

Classic analysis

List of over represented pathways

	GS	SIZE	ES	NES	NOM	FDR	FWER
	follow link to M SigDB						
NADLER_OBESITY_DN	Details...	46	0.80	2.65	0.000	0.000	0.000
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	Details...	42	0.73	2.48	0.000	0.000	0.000
HSHAO_LIVER_SPECIFIC_GENES	Details...	218	0.59	2.47	0.000	0.000	0.000
KIM_LIVER_CANCER_POOR_SURVIVAL_DN	Details...	38	0.75	2.42	0.000	0.000	0.000
WOO_LIVER_CANCER_RECURRENCE_DN	Details...	70	0.66	2.42	0.000	0.000	0.000
LEE_LIVER_CANCER_CIPROFIBRATE_DN	Details...	61	0.65	2.33	0.000	0.000	0.000
CHIANG_LIVER_CANCER_SURCLASS_PROLIFERATION_DN	Details...	161	0.56	2.31	0.000	0.000	0.000
LEE_LIVER_CANCER_SURVIVAL_UP	Details...	159	0.55	2.27	0.000	0.000	0.000
KEGG_PROPANOATE_METABOLISM	Details...	31	0.73	2.26	0.000	0.000	0.000
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	Details...	31	0.73	2.26	0.000	0.000	0.000
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	Details...	48	0.67	2.25	0.000	0.000	0.000
LEE_LIVER_CANCER_DENA_DN	Details...	73	0.62	2.21	0.000	0.000	0.003



Molecular Signatures Database (MSigDB)

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

The screenshot shows the MSigDB website interface. At the top is the GSEA logo and navigation links: GSEA Home, Downloads, Molecular Signatures Database (active), Documentation, Contact, and Team. A left sidebar contains links: MSigDB Home, About Collections, Browse Gene Sets, Search Gene Sets, Investigate Gene Sets, View Gene Families, and Help. The main content area is titled 'Molecular Signatures Database v7.0' and features an 'Overview' section with a list of actions: Search, Browse, Examine, Download, and Investigate. The 'Investigate' section is expanded, showing options to compute overlaps, categorize members, and view expression profiles. To the right, a 'Collections' section lists 8 major categories: H (hallmark gene sets), C1 (positional gene sets), C2 (curated gene sets), C3 (motif gene sets), C4 (computational gene sets), and C5 (GO gene sets). The bottom of the page includes 'License Terms' and a registration prompt.

GSEA
Gene Set Enrichment Analysis

GSEA Home Downloads **Molecular Signatures Database** Documentation Contact Team

► MSigDB Home
► About Collections
► Browse Gene Sets
► Search Gene Sets
► Investigate Gene Sets
► View Gene Families
► Help

MSigDB
Molecular Signatures Database

Molecular Signatures Database v7.0

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the [GO_NOTCH_SIGNALING_PATHWAY](#) gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
 - **Compute overlaps** between your gene set and gene sets in MSigDB.
 - **Categorize** members of a gene set by gene families.
 - **View the expression profile** of a gene set in a provided public expression compendia.

License Terms

GSEA and MSigDB are available for use under [these license terms](#).

Please [register](#) to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can

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.



Molecular Signatures Database (MSigDB)

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 1350 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



Gene Sets: GMT format

- Genes are grouped in families describing biological pathways, molecular functions, disease mechanisms etc..
- Example: Collection 5 from MSigDB Biological processes (c5.bp.v7.0.symbols.gmt)

GO Biological Process
(Column A)

Extra information
(Column B)

Genes in the GO BP

A	B	C	D	E	F	G	H	I	J	K
GO_CARDIAC_CHAMBER_DEVELOPMENT	http://www.gsea-msi	ZMPSTE24	UBE4B	CITED2	TAB1	SEMA3C	FRS2	PPP1R13L	ADAMTS6	XIRP2
GO_DNA_DEPENDENT_DNA_REPLICATION_MAINTENANCE_OF_FIDELITY	http://www.gsea-msi	ALYREF	CDK9	EME1	DDX11	DNA2	EME2	CENPX	PRIMPOL	ZNF365
GO_CIRCADIEN_RHYTHM	http://www.gsea-msi	ADA	NR1H3	NAMPT	CDK4	HNRNP	PRMT5	NCOA2	MYBBP1A	ATG7
GO_PHOSPHATIDYLSELINE_ACYL_CHAIN_REMODELING	http://www.gsea-msi	PLA2G4B	LPCAT3	PLAAT3	OSBPL5	OSBPL8	OSBPL10	PLA2G4E	MBOAT1	LPCAT4
GO_SPINAL_CORD_DEVELOPMENT	http://www.gsea-msi	OLIG2	GDF11	ADARB1	LBX1	DBX1	FOXN4	GDF7	DAB1	MDGA2
GO_PLATELET_DERIVED_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY	http://www.gsea-msi	TXNIP	RGS14	F3	F7	FER	ABL1	TIPARP	RAPGEF1	HIP1
GO_CELLULAR_RESPONSE_TO_LIPOPROTEIN_PARTICLE_STIMULUS	http://www.gsea-msi	CDH13	TLR6	TICAM1	ABCA1	AKT1	FCER1G	FGF21	HMGCS1	APOE
GO_POSITIVE_REGULATION_OF_EPITHELIAL_CELL_DIFFERENTIATION	http://www.gsea-msi	CD24	CDH5	TRIM16	ADD1	KDF1	CTNNB1	CYP27B1	ETV2	FOXCI
GO_POSITIVE_REGULATION_OF_KINASE_ACTIVITY	http://www.gsea-msi	ABI1	CD24	TOM1L1	ADAM8	RAD50	RASGRP1	TENM1	CDK5	SPRY2
GO_POTASSIUM_ION_TRANSPORT	http://www.gsea-msi	KCNE3	KCNJ18	HCN4	ABCC9	KCNK7	AKAP9	CDK2	KCNMB2	CDKN1B
GO_REGULATION_OF_T_CELL_RECEPTOR_SIGNALING_PATHWAY	http://www.gsea-msi	ADA	BTN2A2	CD226	MALT1	LILRB4	CD160	CD300A	CCR7	RC3H1
GO_NEGATIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION	http://www.gsea-msi	MIR2355	CDK6	RIDA	CDKN1B	CDKN1C	CDKN2B	CNMD	STRAP	DUSP10
GO_MOVEMENT_IN_ENVIRONMENT_OF_OTHER_ORGANISM_INVOLVED_IN_SYMBIOTIC_INTERACTION	http://www.gsea-msi	TRIM28	TRIM13	MID2	TRIM31	CHMP4B	DDB1	TRIM32	TRIM35	CHMP2B
GO_REGULATION_OF_PROTEIN_TARGETING_TO_MITOCHONDRION	http://www.gsea-msi	SAE1	HUWE1	ARIH2	HAX1	UBE2J2	LRRK2	CSNK2A2	UBL4B	ABLIM3
GO_APICAL_PROTEIN_LOCALIZATION	http://www.gsea-msi	HCN1	SHROOM2	ARF4	MAL	GOPC	VANGL2	SHROOM3	RDX	NAPA
GO_REGULATION_OF_ESTABLISHMENT_OF_PLANAR_POLARITY	http://www.gsea-msi	GPC6	PSME3	PSMD14	SEC24B	CTHRC1	AP2M1	AP2S1	CLTC	PSMB11
GO_FOREBRAIN_NEURON_DEVELOPMENT	http://www.gsea-msi	SEMA3A	DCLK2	ARX	DRD1	DRD2	FGF8	FGFR2	ATF5	FOXP1
GO_POSITIVE_REGULATION_OF_PROTEIN_MATURATION	http://www.gsea-msi	ADAM8	SPON1	TIMM17A	CCBE1	ENO1	F12	GSN	HPN	KLKB1
GO_NEUROMUSCULAR_JUNCTION_DEVELOPMENT	http://www.gsea-msi	GPHN	CACNG2	UNC13B	AFG3L2	RER1	CHRNA1	LRRK2	COL4A1	COL4A5
GO_MITOTIC_CYTOKINESIS	http://www.gsea-msi	PDCD6IP	KIF20A	CENPA	STAMBP	CFL1	JTB	RAB35	CIT	SNX18
GO_NEGATIVE_REGULATION_OF_RESPONSE_TO_ENDOPLASMIC_RETICULUM_STRESS	http://www.gsea-msi	RACK1	CREB3	HYOU1	OS9	PARK7	CLU	LRRK2	ATF6B	PPP1R15A
GO_SMAD_PROTEIN_SIGNAL_TRANSDUCTION	http://www.gsea-msi	TOB1	GDF11	RBM14	LEFTY1	SUB1	RBPMS	BTBD11	PARP1	GDF7
GO_CYTOPLASMIC_TRANSLATION	http://www.gsea-msi	DPH3P1	EIF3M	DNAJC24	CPEB2	JAKMIP1	AARS	DHX9	DHX36	DPH1
GO_MEIOTIC_CHROMOSOME_SEGREGATION	http://www.gsea-msi	SYCE1L	RNF212B	SMC4	ACTR3	ACTR2	MEI4	SYCP2	SMC2	CENPC
GO_POSITIVE_REGULATION_OF_CALCIIUM_ION_TRANSPORT	http://www.gsea-msi	KCNE3	RAMP3	TRDN	LILRA2	ADCYAP1R1	CCR1	CRH	TRPV3	DRD1
GO_REGULATION_OF_DOUBLE_STRAND_BREAK_REPAIR	http://www.gsea-msi	PARP3	ACTR2	MAD2L2	RAD51AP1	POLQ	CHEK1	SHLD3	CGAS	RM12
GO_RNA_DEPENDENT_DNA_BIOSYNTHETIC_PROCESS	http://www.gsea-msi	TEN1	RAD50	CCT7	CCT4	CCT2	CCT8	PTGES3	PNKP	NEK7
GO_REGULATION_OF_B_CELL_RECEPTOR_SIGNALING_PATHWAY	http://www.gsea-msi	CD300A	FCRL3	CMTM3	GCSAML	NFAM1	ELF1	ELF2	FCGR2B	PLCL2
GO_DENDRITE_DEVELOPMENT	http://www.gsea-msi	ABI1	HDAC6	FOXO6	ARMCX5-GP1	DNM1L	PQBPI	ACTR2	ABI2	FARP1



GSEA statistics

Enrichment Score (ES): Degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes (the default ranking metric is signal-to-noise ratio)

Examples for positive enriched gene sets in the contrast LINC vs CTRL (enriched in LINC)

Table: GSEA Results Summary	
Dataset	LINC vs Ctrl
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	GO_CELLULAR_RESPONSE_TO_ZINC_ION
Enrichment Score (ES)	0.8632142
Normalized Enrichment Score (NES)	2.4628952
Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

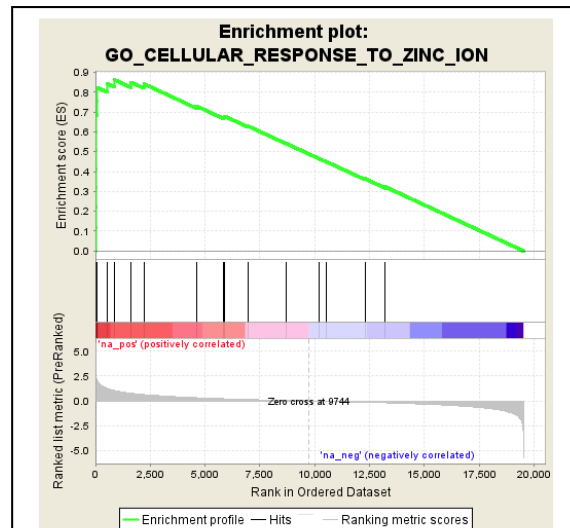


Fig 1: Enrichment plot: GO_CELLULAR_RESPONSE_TO_ZINC_ION
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA Results Summary	
Dataset	LINC vs Ctrl
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	GO_ACTIN_MEDIATED_CELL_CONTRACTION
Enrichment Score (ES)	0.38933018
Normalized Enrichment Score (NES)	1.5310684
Nominal p-value	0.0055452865
FDR q-value	0.5675065
FWER p-Value	1.0

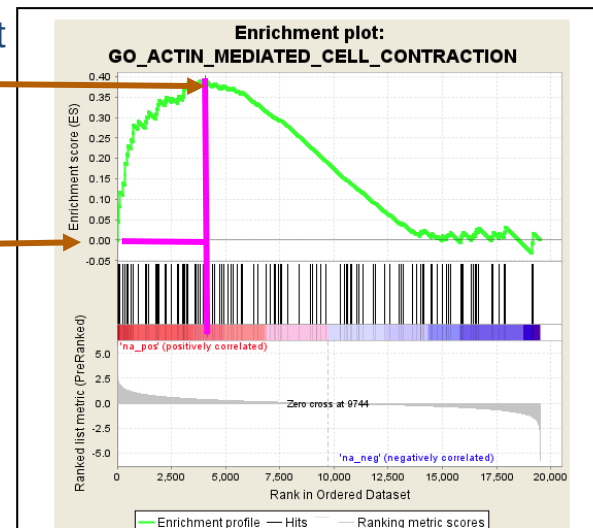


Fig 1: Enrichment plot: GO_ACTIN_MEDIATED_CELL_CONTRACTION
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Enrichment
score
Leading
edge
subset

GSEA statistics

Enrichment Score (ES): Degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes (the default ranking metric is signal-to-noise ratio)

Examples for negative enriched gene sets in the contrast LINC vs CTRL (enriched in CTRL)

Table: GSEA Results Summary	
Dataset	LINC vs Ctrl
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_neg
GeneSet	GO_LONG_CHAIN_FATTY_ACID_IMPORT_INTO_CELL
Enrichment Score (ES)	-0.8595188
Normalized Enrichment Score (NES)	-2.2617462
Nominal p-value	0.0
FDR q-value	0.0023107266
FWER p-Value	0.002

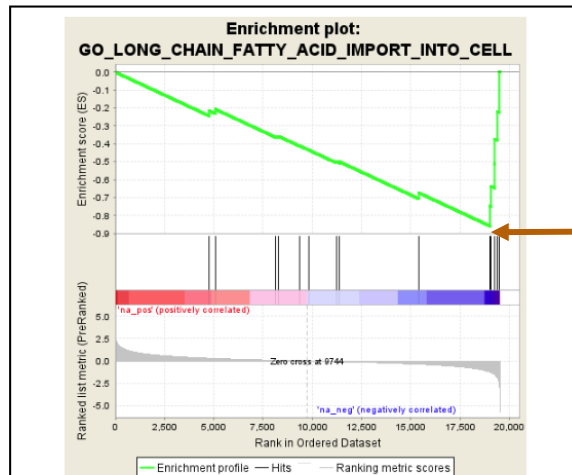


Fig 1: Enrichment plot: GO_LONG_CHAIN_FATTY_ACID_IMPORT_INTO_CELL

Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA Results Summary	
Dataset	LINC vs Ctrl
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_neg
GeneSet	GO_CELLULAR_RESPONSE_TO_IONIZING_RADIATION
Enrichment Score (ES)	-0.45991984
Normalized Enrichment Score (NES)	-1.650384
Nominal p-value	0.006289308
FDR q-value	0.49168018
FWER p-Value	1.0

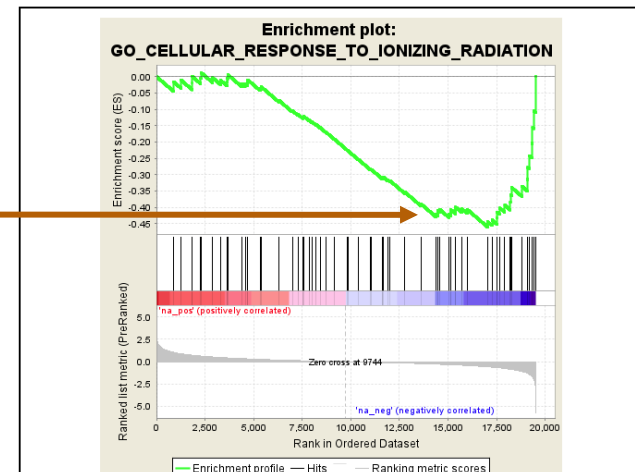


Fig 1: Enrichment plot: GO_CELLULAR_RESPONSE_TO_IONIZING_RADIATION

Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Enrichment
score



GSEA statistics

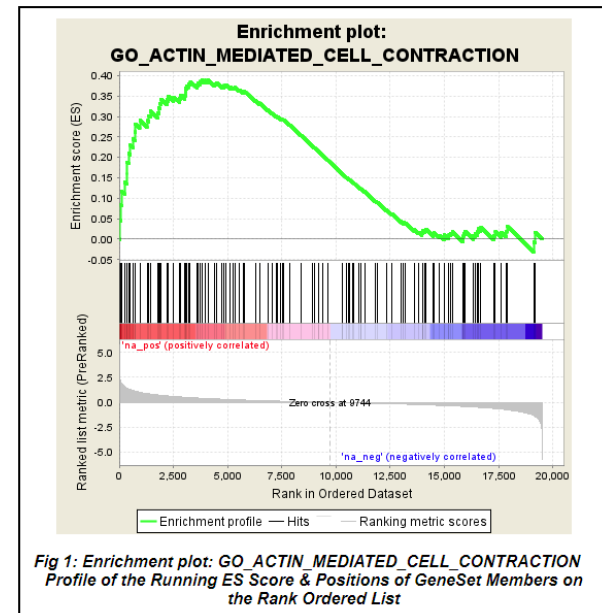
Normalized enrichment score (NES)

- Obtained by normalizing the enrichment score
- Accounts for differences in gene set size and in correlations between gene sets and the expression dataset
- Used to compare analysis results across gene sets

$$NES = \frac{\text{actual ES}}{\text{mean(ESs against all permutations of the dataset)}}$$

Table: GSEA Results Summary

Dataset	LINC vs. Ctrl
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	GO_ACTIN_MEDIATED_CELL_CONTRACTION
Enrichment Score (ES)	0.38933018
Normalized Enrichment Score (NES)	1.5310684
Nominal p-value	0.0055452865
FDR q-value	0.5675065
FWER p-Value	1.0

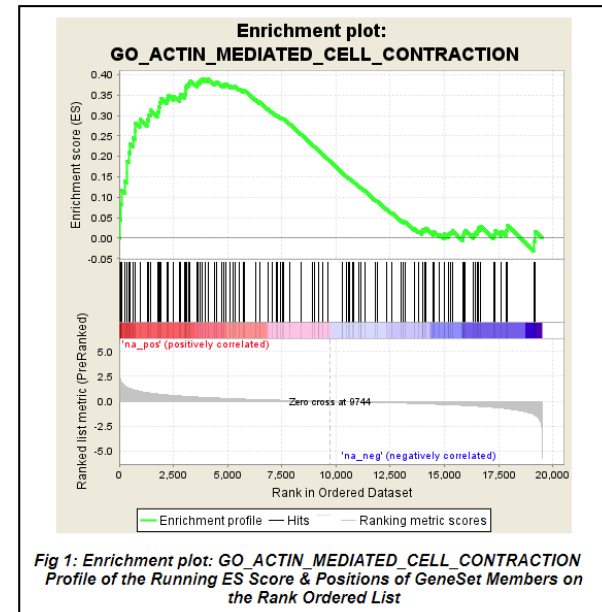


GSEA statistics

Nominal p-value: Estimates the statistical significance of the enrichment score for a single gene set

FDR q-value: The false discovery rate (FDR) is the estimated probability that a gene set with a given NES represents a false positive finding (Ex: an FDR of 25% indicates that the result is likely to be valid 3 out of 4 times)

Table: GSEA Results Summary	
Dataset	LINC vs. Ctrl
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	GO_ACTIN_MEDIATED_CELL_CONTRACTION
Enrichment Score (ES)	0.38933018
Normalized Enrichment Score (NES)	1.5310684
Nominal p-value	0.0055452865
FDR q-value	0.5675065
FWER p-Value	1.0



GSEA tool practical

Input data preparation

Input Data: Lung Cancer dataset example

- Transcriptome of Human NSCLC tissues (GEO accession: GSE74706)
 - NSCLC: n=18
 - TumorFree: n=18

The screenshot shows the NCBI GEO Accession Display page for GSE74706. The page includes a search bar, navigation links, and a detailed description of the dataset. The dataset is titled "Transcriptome of human NSCLC tissues" and is from the organism "Homo sapiens". It is an expression profiling by array experiment. The overall design is described as "RNA from patient samples was isolated to examine the TGFβ pathway expression between matching pairs of tumor-free lung and NSCLC specimen". The contributor is listed as "Marwitz S, Ammerpohl O, Reck M, Klingmueller U, Goldmann T". The citation is "Marwitz S, Depner S, Dvornikov D, Merkle R et al. Downregulation of the TGFβ Pseudoreceptor BAMBI in Non-Small Cell Lung Cancer Enhances TGFβ Signaling and Invasion. Cancer Res 2016 Jul 1;76(13):3785-801. PMID: 27197161". The submission date is Nov 05, 2015, and the last update date is Jan 09, 2018. The contact name is Sebastian Marwitz, and the organization is Research Center Borstel - Leibniz Lung Center. The department is Pathology, and the street address is Parkallee 3a, Borstel, 23845, Germany. The platforms are listed as GPL13497 Agilent-026652 Whole Human Genome Microarray 4x44K v2 (Probe Name version).

Series matrix

The screenshot shows the GEO2R interface. The "Download family" section is highlighted with a red box. The "Series Matrix File(s)" option is selected. The "Supplementary file" section shows the file "GSE74706_RAW.tar" with a size of 326.0 Mb and a download link. The "Format" section shows the "Format" dropdown menu with options "SOFT", "MINIML", and "TXT". The "Series Matrix File(s)" option is selected. The "Download family" section is highlighted with a red box. The "Series Matrix File(s)" option is selected. The "Supplementary file" section shows the file "GSE74706_RAW.tar" with a size of 326.0 Mb and a download link. The "Format" section shows the "Format" dropdown menu with options "SOFT", "MINIML", and "TXT".



Input data: GCT format

GCT format: For microarrays, contains the expression data. Tabulator separated.

First row is always
#1.2

Second row has two tab separated
values:

34183 = Number of geneID
36 = Number of samples

From the **third row** on
starts the expression
matrix

```
1 #1.2
2 34183 36
3 NAME Description 17962_08_Lung 17962_08_Tumor 21577_08_Lung 21577_08_Tumor 6495_08_Lung 6495_08_Tumor 12097_07_Lung
4 (+)E1A_r60_1 NA 0 0.1795826 0.1795826 0.1795826 0 -0.069934845 0.1795826 0.1795826 0 -0.069934845 -0.19537
5 (+)E1A_r60_3 NA -0.19623804 -0.28856254 -0.5115266 0.39665508 -0.71220696 1.0335908 2.3404715 -0.14526844 -0.23276138 0.3
6 (+)E1A_r60_a104 NA -0.8340554 -0.28793335 -0.93885803 1.4892027 -1.1166267 2.0433853 4.0047445 -0.84728265 -0.29745603 -0.0
7 (+)E1A_r60_a107 NA 0.15162134 0.45030403 0.36336517 1.3216305 -0.23709536 0.7641797 1.7247491 0.5285454 0.48011923 0.4
8 (+)E1A_r60_a135 NA 0.48870468 0.6699066 0.99248886 1.3114834 0.0710907 0.48975086 0.56334877 0.58070755 0.8026047 0.3
9 (+)E1A_r60_a20 NA 0.5134697 0.49859428 1.0381546 1.1965628 -0.030184746 0.24079514 0.47081947 0.43220806 0.78427505
10 (+)E1A_r60_a22 NA 0.574193 0.681386 1.1696777 1.3308182 0.16920853 0.33121872 0.49439335 0.5239725 0.8721428 0.0
11 (+)E1A_r60_a97 NA 0.711298 0.69098663 1.1606741 1.3366041 0.32015896 0.37965775 0.6580162 0.51055336 0.8807583 0.0
12 (+)E1A_r60_n11 NA 0.5108471 0.4702568 0.66748047 0.70708656 0.21434402 0.13879585 0.5108471 0.18255806 0.65517616 0.0
13 (+)E1A_r60_n9 NA 0.15073204 0.20399284 0.25653076 0.4207363 0.039978027 0.08213234 0.29883766 0.08213234 0.3340416 -0.0
14 (-)3xSLv1 NA -0.07265639 0.026640415 -0.04011154 -0.09190309 -0.01863134 -0.15329444 -0.2357111 0.01543498 -0.0070072412 0.1
15 A_23_P100001 NA 0.30399132 0.8295469 -0.5054226 0.24987602 0.2578745 -2.989729 -0.48598194 0.41997814 -0.17772484 -2.
16 A_23_P100022 NA 0.5760064 -1.5653286 0.9171052 -0.8337636 0.38460922 0.33457136 2.8440013 -2.824328 0.059636593 3.8
17 A_23_P100056 NA 0.3298936 -2.118638 0.82557726 -1.1536093 0.65477705 -1.0410714 -0.09235954 -1.2997136 -0.016194344
18 A_23_P100074 NA -0.74025345 -0.80999184 -1.1920676 -0.8951044 -0.7765732 1.0460758 -1.1037989 1.1644545 -0.59459496 0.9
19 A_23_P100127 NA -1.8846049 1.1270428 -0.6877289 1.700746 -1.1477237 1.763371 -0.9462967 3.7809863 -1.1842289 3.0
20 A_23_P100141 NA 0.60105515 -0.4087453 0.7880449 0.5643873 0.1839261 -0.03980875 1.0823107 -1.0503159 -0.21138 1.0
21 A_23_P100189 NA 1.5875831 0.56400204 0.9729352 0.87494993 0 -0.106552124 1.253355 0.6427555 0.73570824 0.831431
22 A_23_P100196 NA -0.15721703 -0.33642197 -0.05117798 0.21472263 0 0.65099144 -0.8198786 0.67990875 0.33212185 1.1101093
```

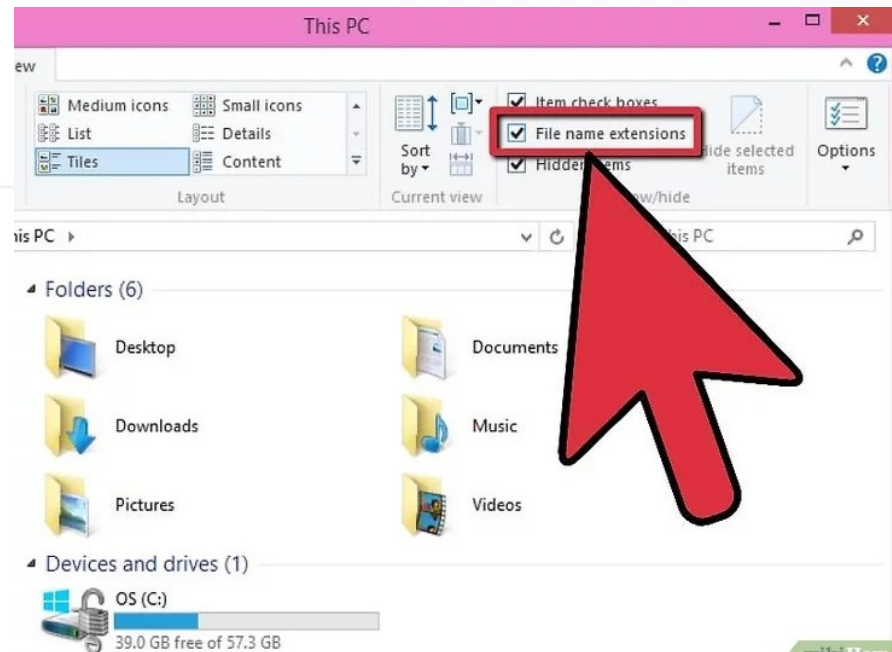
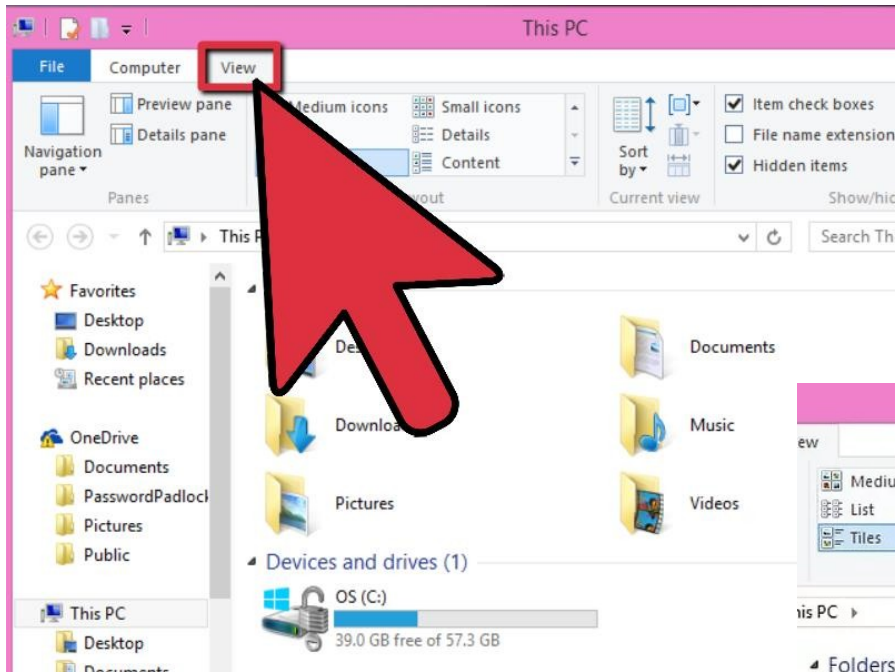
First column in the
expression matrix
contains gene IDs

Second column in the expression matrix
contains description but is ignored by the
program (NA in this case)

The rest of the columns in the
expression matrix contain
expression values for each
sample



Change file extension in Windows



CLS format: For microarrays, contains the phenotypic data. Blank space separated.

Always 1

Second row: Names of the phenotypic traits.
IMPORTANT: First condition is Tumorfrees lung, like in the third row.

GSEA tool

Available at: <https://software.broadinstitute.org/gsea/index.jsp>

GSEA
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

Downloads

UC San Diego BROAD INSTITUTE

Software

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. For details on the GSEA algorithm and software refer to the [Documentation](#). For details on the latest release refer to the [Release Notes](#). The source is available from our [GitHub organization](#).

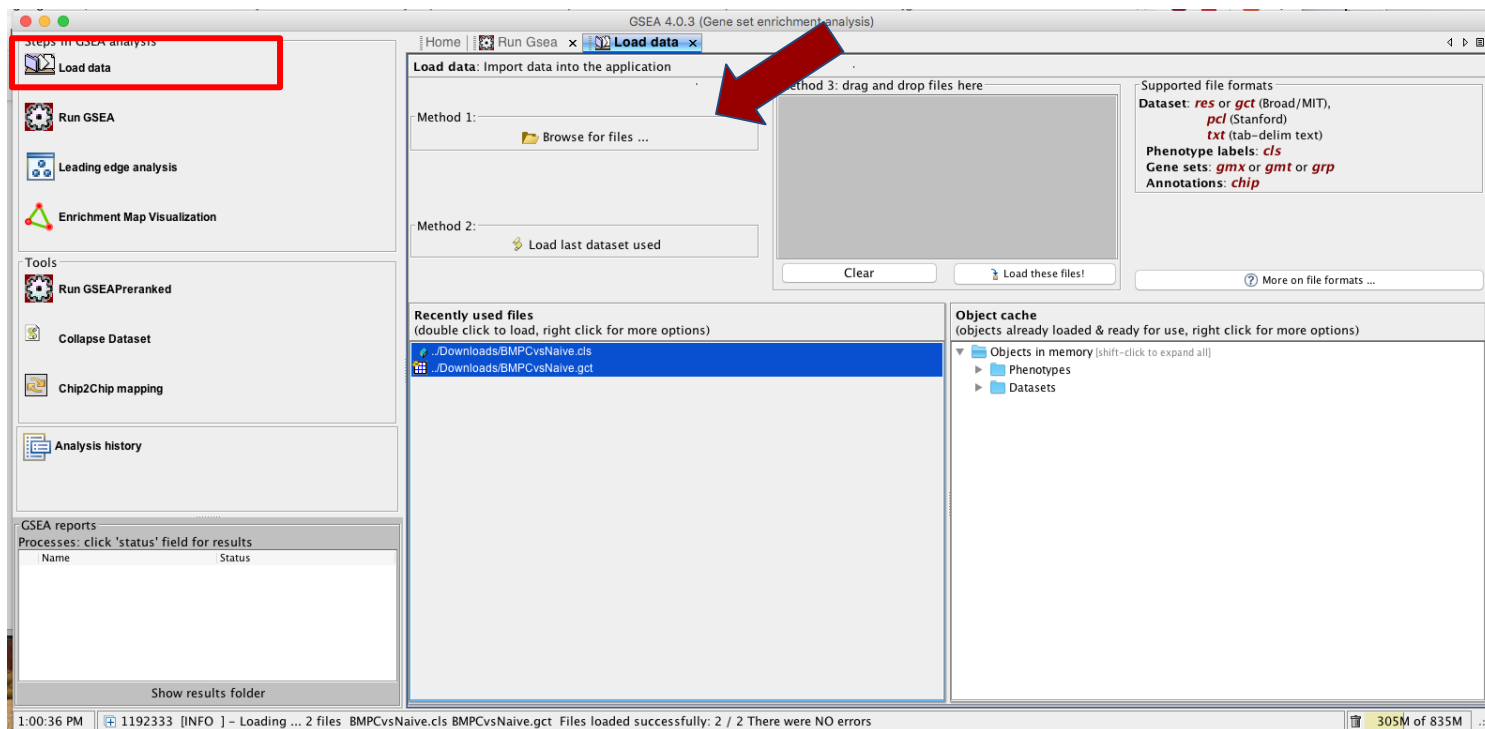
See the [license terms page](#) for details about the license for the GSEA software and source code. Please note that the license terms vary for different versions of the software.

GSEA v4.0.3 Mac App	Download and unzip the Mac App Archive, then double-click the GSEA application to run it. You can move the app to the Applications folder, or anywhere else. MacOS Catalina users: We sign our Mac App as a trusted Apple developer, but it is not yet notarized by Apple (a new requirement in Catalina). To run it, right-click on the downloaded GSEA app; select "Open" from the menu; and click the "Open" button in the window that pops up. After that, double-clicking on the app will also work.	download GSEA_4.0.3.app.zip
GSEA v4.0.3 for Windows	Download and run the installer. A GSEA shortcut will be created on the Desktop; double-click it to run the application. 64-bit Windows is required	download GSEA_Win_4.0.3-installer.exe
GSEA v4.0.3 for Linux	Download and unzip the Archive. See the included readme.txt for further instructions. 64-bit Linux is required	download GSEA_Linux_4.0.3.zip
GSEA v4.0.3 for the command line (all platforms)	Download and unzip the Archive. See the included readme.txt for further instructions. Requires separate Java 11 installation.	download GSEA_4.0.3.zip
GSEA v4.0.3 Java Web Start (all platforms)	Launches the GSEA desktop application from the web. Requires separate Java 8 installation. Please use a configuration smaller than your computer's total memory. This option will be removed in a future release.	Launch with 4GB (for 64-bit Java only) ▼ Launch
GenePattern GSEA Module	Use GSEA from within GenePattern (a powerful and flexible analysis platform developed at the Broad Institute and UCSD) in concert with a large suite of other analytics.	



GSEA Classic Analysis

1. Download GSEA: GSEA v4.0.3 Java Web Start (all platforms)
2. Download gct and cls files in the folder “MaterialsCursGSEA”
3. Open the app
4. “Load data” - “Browse for files” - load the .cls and .gct files



GSEA Classic Analysis

1. Go to “Run GSEA” - “Required fields” and “Basic fields”

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis
- Enrichment Map Visualization

Tools

- Run GSEAPreranked
- Collapse Dataset
- Chip2Chip mapping
- Analysis history

GSEA reports

Processes: click 'status' field for results

	Name	Status
1	Gsea	Success
2	Gsea	Running

Required fields

Expression dataset: NSCLCndNT [34183x36 (ann: 34183,36,chip na)]

Gene sets database: roadinstitute.org://pub/gsea/gene_sets_final/c5.bp.v7.0.symbols.gmt

Number of permutations: 1000

Phenotype labels: EA_BicoH.GSE74706\NSCLCndNT.cls#NSCLC_versus_Tumorfree_lung

Collapse dataset to gene symbols: true

Permutation type: phenotype

Chip platform: adinstitute.org://pub/gsea/annotations/Agilent_HumanGenome.chip

Basic fields

Analysis name: NSCLCvsNT

Enrichment statistic: weighted

Metric for ranking genes: Signal2Noise

Gene list sorting mode: real

Gene list ordering mode: descending

Max size: exclude larger sets: 500

Min size: exclude smaller sets: 15

Save results in this folder: C:\Users\mamaf\Documents\CursGSEA_BicoH

Advanced fields

phenotype

gene_set = When you have fewer than seven (7) samples in any phenotype

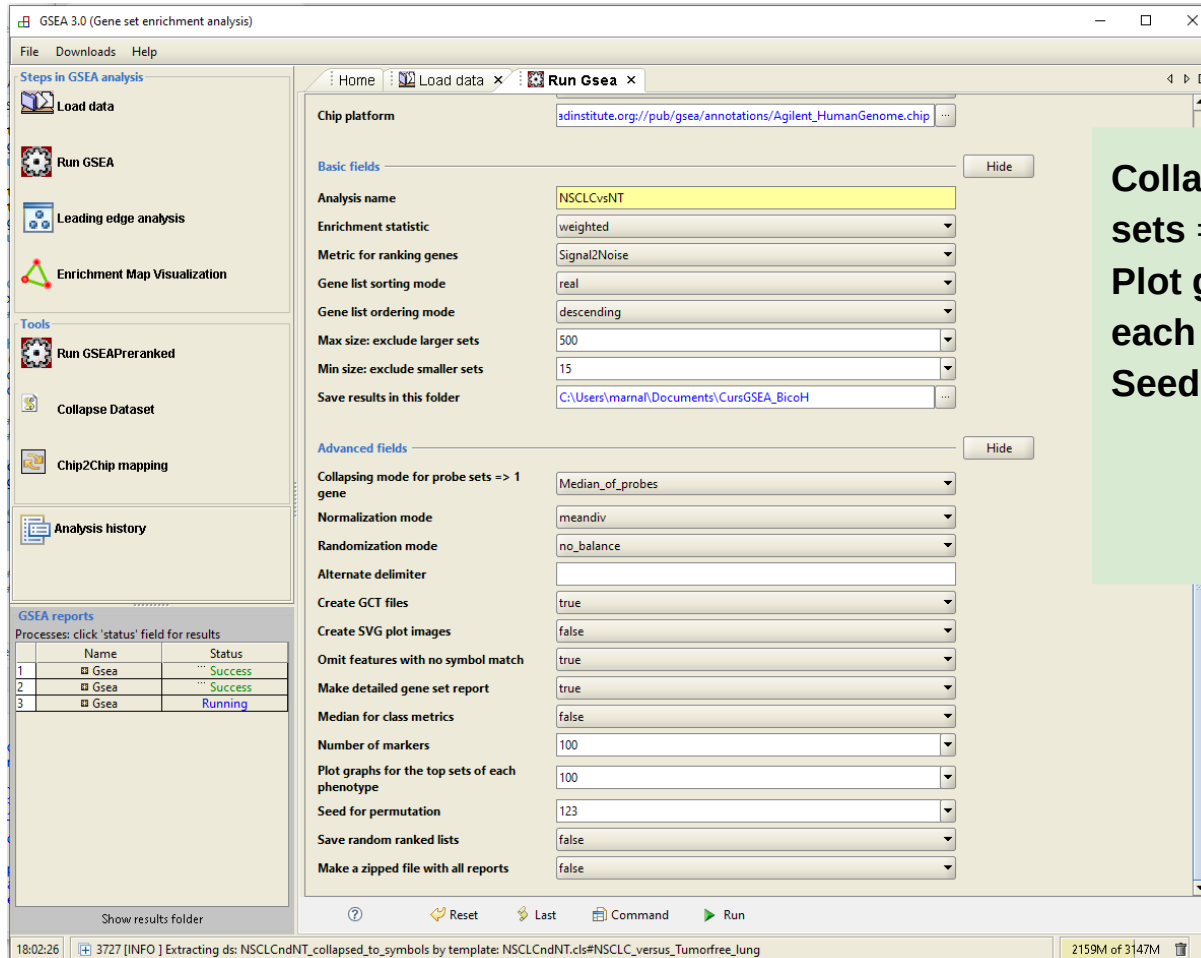
Agilent_HumanGenome in this case

Output folder name

Output folder where the results are stored

GSEA Classic Analysis

1. Go to “Run GSEA” - “Advanced fields”



GSEA Classic results

GSEA Report for Dataset NSCLCndNT

Enrichment in phenotype: NSCLC (18 samples)

- 958 / 3305 gene sets are upregulated in phenotype NSCLC
- 384 gene sets are significant at FDR < 25%
- 189 gene sets are significantly enriched at nominal pvalue < 1%
- 288 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](#)

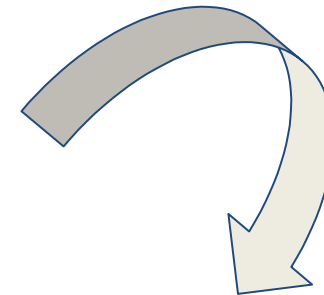
Enriched in NSCLC

Enrichment in phenotype: Tumorfree_lung (18 samples)

- 2347 / 3305 gene sets are upregulated in phenotype Tumorfree_lung
- 1291 gene sets are significant at FDR < 25%
- 490 gene sets are significantly enriched at nominal pvalue < 1%
- 868 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](#)

Enriched in TumorFree

index.html in the folder contains the results summary



GSEA_Plots (Details)

Gene set information

Dataset details

- The dataset has 34183 native features
- After collapsing features into gene symbols, there are: 14110 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 4045 / 7350 gene sets
- The remaining 3305 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

G5 Follow link to MSigDB		GS DETAILS	SIZE	FDR	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	GO_MEIOTIC_CHROMOSOME_SEGREGATION	Details...	41	0.72	2.09	0.000	0.041	0.043	904	
2	GO_REGULATION_OF_CHROMOSOME_SEGREGATION	Details...	66	0.67	2.05	0.000	0.042	0.088	1023	
3	GO_MEIOTIC_CELL_CYCLE_PROCESS	Details...	96	0.59	2.05	0.000	0.030	0.092	1401	
4	GO_MEIOSIS_I_CELL_CYCLE_PROCESS	Details...	65	0.61	2.04	0.000	0.025	0.099	904	
5	GO_SIGNAL_TRANSDUCTION_INVOLVED_IN_CELL_CYCLE_CHECKPOINT	Details...	44	0.61	2.03	0.000	0.024	0.118	1162	
6	GO_DNA_PACKAGING	Details...	125	0.67	2.03	0.000	0.020	0.119	1156	
7	GO_NEGATIVE_REGULATION_OF_DNA_REPLICATION	Details...	22	0.75	2.02	0.000	0.020	0.131	1272	
8	GO_REGULATION_OF_NUCLEAR_DIVISION	Details...	132	0.51	2.01	0.000	0.022	0.154	1002	
9	GO_MITOTIC_CELL_CYCLE_CHECKPOINT	Details...	105	0.58	2.00	0.000	0.021	0.164	1162	
10	GO_NEGATIVE_REGULATION_OF_CHROMOSOME_SEGREGATION	Details...	29	0.72	2.00	0.000	0.019	0.164	876	
11	GO_REGULATION_OF_SISTER_CHROMATID_SEGREGATION	Details...	52	0.64	2.00	0.000	0.017	0.164	972	
12	GO_MEIOTIC_CELL_CYCLE	Details...	137	0.54	2.00	0.000	0.017	0.169	1401	

GSEA classic results

GSEA plots in “Details...”

- Enrichment plot
- Table with genes in the gene set
- Heatmap
- Random ES distribution

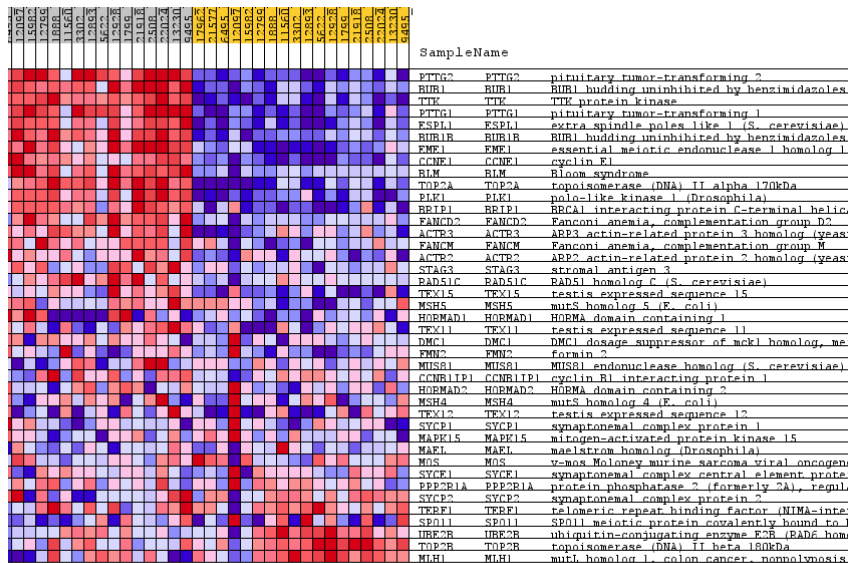
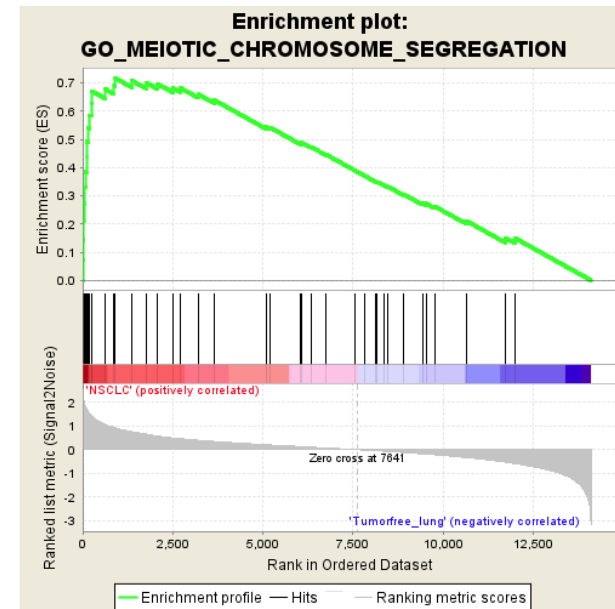


Fig 2: GO_MEIOTIC_CHROMOSOME_SEGREGATION
Blue-Pink O' Gram in the Space of the Analyzed GeneSet



Input data: RNK format

RNK file:

- Valid for microarrays and RNAseq
- A rank score is given for each gene: Rank Score = $-\log_{10}(\text{pvalue}) * \text{sign}(\log\text{FC})$
- A gene differentially expressed at a significant level (low pvalue close to 0) will be assigned with a high score

GSE74706_geo2r_NSCLCvsNT - Microsoft Excel

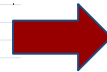
Archivo Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista Programador

Calibri 11 Fuente Alineación Combinar y centrar Número Formato condicional Dar formato Estilos de celdas Insertar

Portapapeles Copiar Copiar formato

J2 =SIGNO(F2)*-LOG10(C2)

	A	B	C	D	E	F	G	H	I	J	K
1	ID	adj.P.Val	P.Value	t	B	logFC	SPOT_ID	GB_ACC	GENE_SYMBOL	RANK_SCORE	
2	A_23_P23783	1.64E-16	4.81E-21	-19.529454	37.625655	-6.63	A_23_P23783	NM_000261	MYOC	-2.03E+01	
3	A_23_P21120	9.99E-16	7.68E-20	-17.978458	34.979518	-3.05	A_23_P21120	NM_001112	ADARB1	-1.91E+01	
4	A_23_P93360	9.99E-16	8.77E-20	-17.906747	34.85214	-5.65	A_23_P93360	NM_001136	AGER	-1.91E+01	
5	A_23_P33173	2.24E-15	2.62E-19	-17.323259	33.798303	-2.36	A_23_P33173	NM_001495	GFRA2	-1.86E+01	
6	A_23_P42630	1.04E-14	1.97E-18	-16.28866	31.850378	-3.98	A_23_P42630	NM_003734	AOC3	-1.77E+01	
7	A_23_P30294	1.04E-14	2.25E-18	-16.222511	31.722245	-4.34	A_23_P30294	NM_001801	CDO1	-1.76E+01	
8	A_23_P53390	1.04E-14	2.40E-18	-16.189176	31.657506	-3.32	A_23_P53390	NM_002837	PTPRB	-1.76E+01	
9	A_23_P91334	1.04E-14	2.44E-18	-16.180222	31.640098	-3.25	A_23_P91334	NM_052970	HSPA12B	-1.76E+01	
10	A_23_P34442	1.05E-14	2.76E-18	-16.120576	31.523925	-3.38	A_23_P34442	NM_019055	ROBO4	-1.76E+01	
11	A_23_P33768	1.15E-14	3.38E-18	-16.019764	31.326748	-3.39	A_23_P33768	NM_153714	C10orf67	-1.75E+01	
12	A_23_P84860	1.25E-14	4.01E-18	-15.935286	31.160713	-5.18	A_23_P84860	NM_007177	FAM107A	-1.74E+01	
13	A_23_P33813	1.92E-14	7.06E-18	-15.657744	30.610016	-4.43	A_23_P33813	NM_019105	TNXB	-1.72E+01	
14	A_23_P32094	1.92E-14	7.29E-18	-15.642224	30.578984	-2.98	A_23_P32094	NM_022648	TNS1	-1.71E+01	
15	A_23_P14434	2.45E-14	1.01E-17	-15.482525	30.258181	-3.15	A_23_P14434	NM_004787	SLIT2	-1.70E+01	
16	A_23_P33785	2.45E-14	1.13E-17	-15.428981	30.150014	-3.11	A_23_P33785	NM_001083	PDE5A	-1.69E+01	
17	A_23_P32935	2.45E-14	1.17E-17	-15.414878	30.121473	-2.58	A_23_P32935	NM_170600	SH2D3C	-1.69E+01	
18	A_23_P67661	2.45E-14	1.38E-17	-15.334439	29.958277	-2.55	A_23_P67661	NM_001864	COX7A1	-1.69E+01	
19	A_23_P41677	2.45E-14	1.40E-17	-15.329226	29.947677	-4.46	A_23_P41677	NM_016929	CLIC5	-1.69E+01	
20	A_23_P25247	2.45E-14	1.40E-17	-15.326698	29.942536	-2.64	A_23_P25247	NM_000442	PECAM1	-1.69E+01	
21	A_23_P32483	2.45E-14	1.43E-17	-15.316614	29.922019	-4.98	A_23_P32483	NM_001199219	INMT	-1.68E+01	
22	A_23_P21655	2.68E-14	1.65E-17	-15.24925	29.784684	-3.51	A_23_P21655	NM_153366	SVEP1	-1.68E+01	
23	A_23_P15180	2.96E-14	1.90E-17	-15.180693	29.644413	-3.38	A_23_P15180	NM_006329	FBN5	-1.67E+01	
24	A_23_P32527	2.96E-14	1.99E-17	-15.15919	29.600311	-3	A_23_P32527	NM_024768	CCDC48	-1.66E+01	
25	A_23_P16677	3.09E-14	2.26E-17	-15.098942	29.476477	-3.67	A_23_P16677	NM_024065	LINC00312	-1.66E+01	
26	A_23_P50040	3.09E-14	2.32E-17	-15.087841	29.453617	-2.75	A_23_P50040	NM_080284	ABCA6	-1.66E+01	
27	A_23_P33641	3.09E-14	2.38E-17	-15.075025	29.427208	-2.78	A_23_P33641	NM_152536	FGD5	-1.66E+01	
28	A_23_P16654	3.09E-14	2.44E-17	-15.063036	29.402487	-4.48	A_23_P16654	NM_024768	CCDC48	-1.66E+01	
29	A_23_P14254	4.01E-14	3.36E-17	-14.912111	29.089932	-1.92	A_23_P14254	NM_014795	ZEB2	-1.65E+01	
30	A_23_P15683	4.01E-14	3.40E-17	-14.906832	29.078954	-2.53	A_23_P15683	NM_005822	RCAN2	-1.65E+01	



GSE74706_NSCLCvsNT.rnk.txt - Microsoft Excel

Archivo Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista Programador

Calibri 11 Fuente Alineación Combinar y centrar Número Formato condicional Dar formato Estilos de celdas Insertar

Portapapeles Copiar Copiar formato

K3

	A	B	C	D	E	F	G	H	I
1	MYOC	-2.03E+01							
2	ADARB1	-1.91E+01							
3	AGER	-1.91E+01							
4	GFRA2	-1.86E+01							
5	AOC3	-1.77E+01							
6	CDO1	-1.76E+01							
7	PTPRB	-1.76E+01							
8	HSPA12B	-1.76E+01							
9	ROBO4	-1.76E+01							
10	C10orf67	-1.75E+01							
11	FAM107A	-1.74E+01							
12	TNXB	-1.72E+01							
13	TNS1	-1.71E+01							
14	SLIT2	-1.70E+01							
15	PDE5A	-1.69E+01							
16	SH2D3C	-1.69E+01							
17	COX7A1	-1.69E+01							
18	CLIC5	-1.69E+01							
19	PECAM1	-1.69E+01							
20	INMT	-1.68E+01							
21	SVEP1	-1.68E+01							
22	FBN5	-1.67E+01							
23	PLAC9	-1.67E+01							
24	LINC00312	-1.66E+01							
25	ABCA6	-1.66E+01							
26	FGD5	-1.66E+01							
27	CCDC48	-1.66E+01							
28	ZEB2	-1.65E+01							
29	RCAN2	-1.65E+01							
30	LMO2	-1.64E+01							
31	LD82	-1.64E+01							
32	SIPR1	-1.64E+01							
33	ABCA8	-1.64E+01							
34	GPX3	-1.63E+01							
35	MND1	-1.63E+01							
36	DES	-1.62E+01							

IMPORTANT: Delete missings and dates in gene symbol column



Steps in GSEA analysis



Load data



Run GSEA



Leading edge analysis



Enrichment Map Visualization

Tools



Run GSEAPreranked



Collapse Dataset



Chip2Chip mapping



Analysis history

GSEA reports

Processes: click 'status' field for results

	Name	Status
1	GseaPreran...	Running

Show results folder

GseaPreranked: Run GSEA on a pre-ranked (with external tools) gene list

Required fields

Gene sets database

roadinstitute.org://pub/gsea/gene_sets_final/c5.bp.v7.0.symbols.gmt

Number of permutations

1000

Ranked List

GSE74706_NSCLCvsNT [29833 names]

Basic fields

Hide

Analysis name

NSCLCvsNT

Enrichment statistic

weighted

Max size: exclude larger sets

500

Min size: exclude smaller sets

15

Save results in this folder

C:\Users\marnal\Documents\CursGSEA_BicoH\GSE74706

Advanced fields

Hide

Normalization mode

meandiv

Alternate delimiter

Create SVG plot images

false

Make detailed gene set report

true

Plot graphs for the top sets of each phenotype

100

Seed for permutation

123

Make a zipped file with all reports

false



GSEA Preranked results

GSEA Report for Dataset GSE74706_NSCLCvsNT

Enrichment in phenotype: *na*

- 970 / 3897 gene sets are upregulated in phenotype *na_pos*
- 569 gene sets are significant at FDR < 25%
- 368 gene sets are significantly enriched at nominal pvalue < 1%
- 469 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



na_pos=NSCLC

Enrichment in phenotype: *na*

- 2927 / 3897 gene sets are upregulated in phenotype *na_neg*
- 1757 gene sets are significantly enriched at FDR < 25%
- 794 gene sets are significantly enriched at nominal pvalue < 1%
- 1272 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



na_neg=NT

Dataset details

- The dataset has 21754 features (genes)
- No probe set => gene symbol collapsing was requested, so all 21754 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 3453 / 7350 gene sets
- The remaining 3897 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

GSEA plots in “Details...”

- Enrichment plot
- Table with genes in the gene set
- Random ES distribution

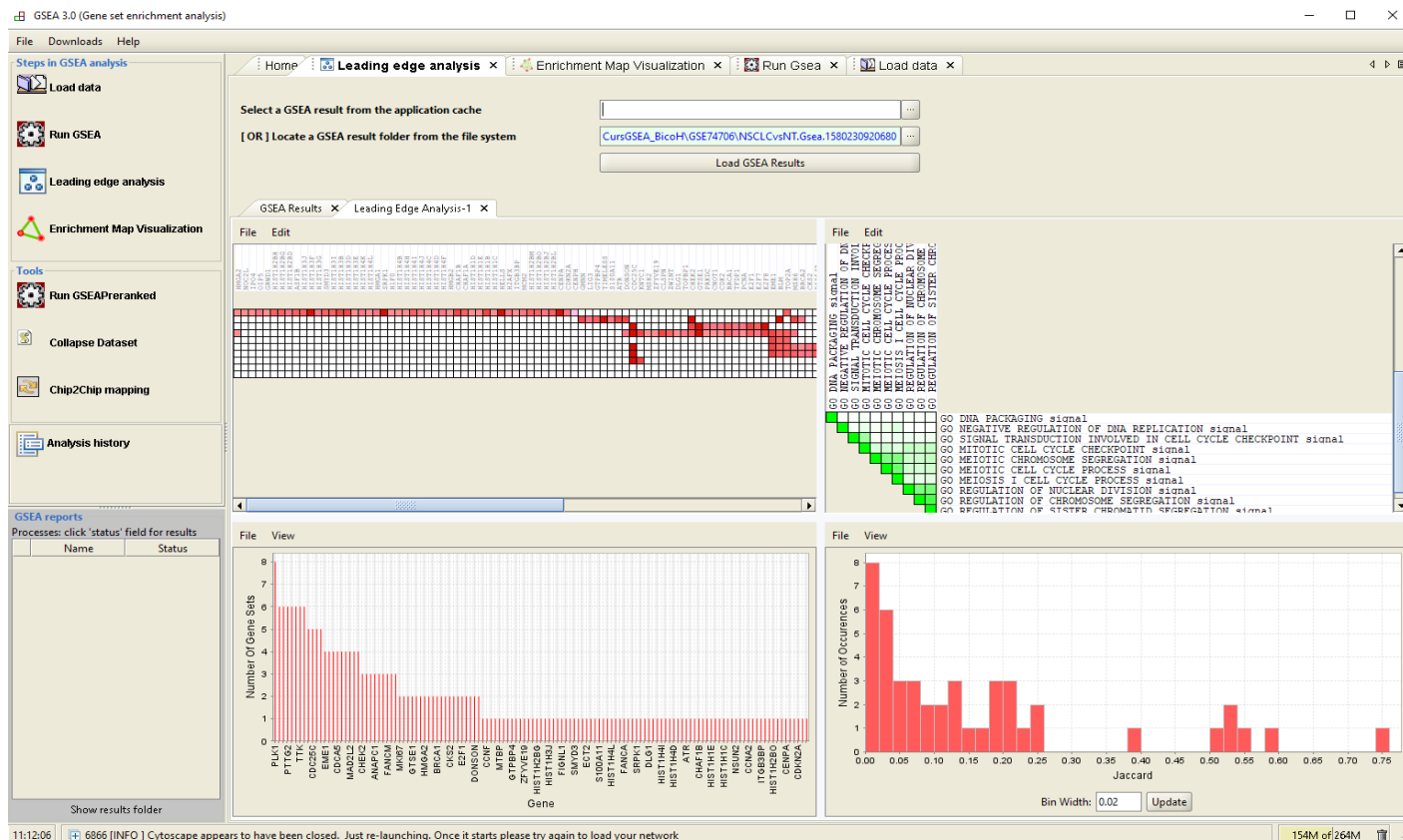
(no heatmap)



GSEA Leading edge analysis

Leading edge analysis to compare two gene sets:

- Load a folder with GSEA results
- Select the gene sets to compare



Practical Exercise

Dataset GSE5600 to perform GSEA analysis:

Classic analysis

1. Download the series matrix from GEO
2. Use Sublime Text to generate the GCT and CLS files
3. Load files in GSEA tool - Load data
4. Perform a GSEA - Run GSEA with the parameters previously described

Pre-Ranked analysis

5. Use the differential expression analysis results from GEOtoR
6. Calculate the RNK score in excel ($=\text{SIGNO}(F2)*-\text{LOG10}(C2)$)
7. Change the extension of the file to .rnk
8. Load .rnk file in GSEA tool - Load data
9. Perform a GSEA - Run GSEAPreranked

