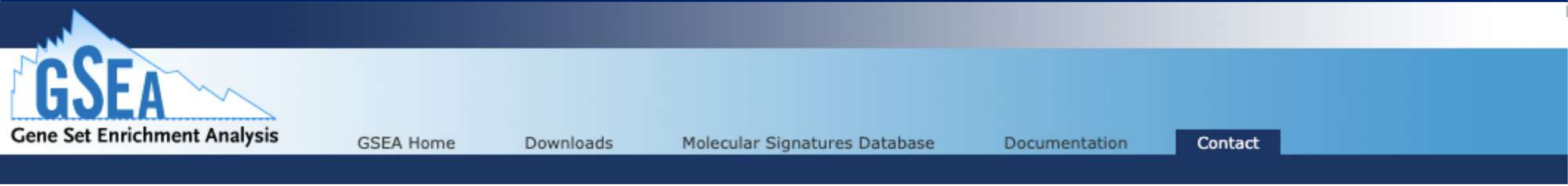
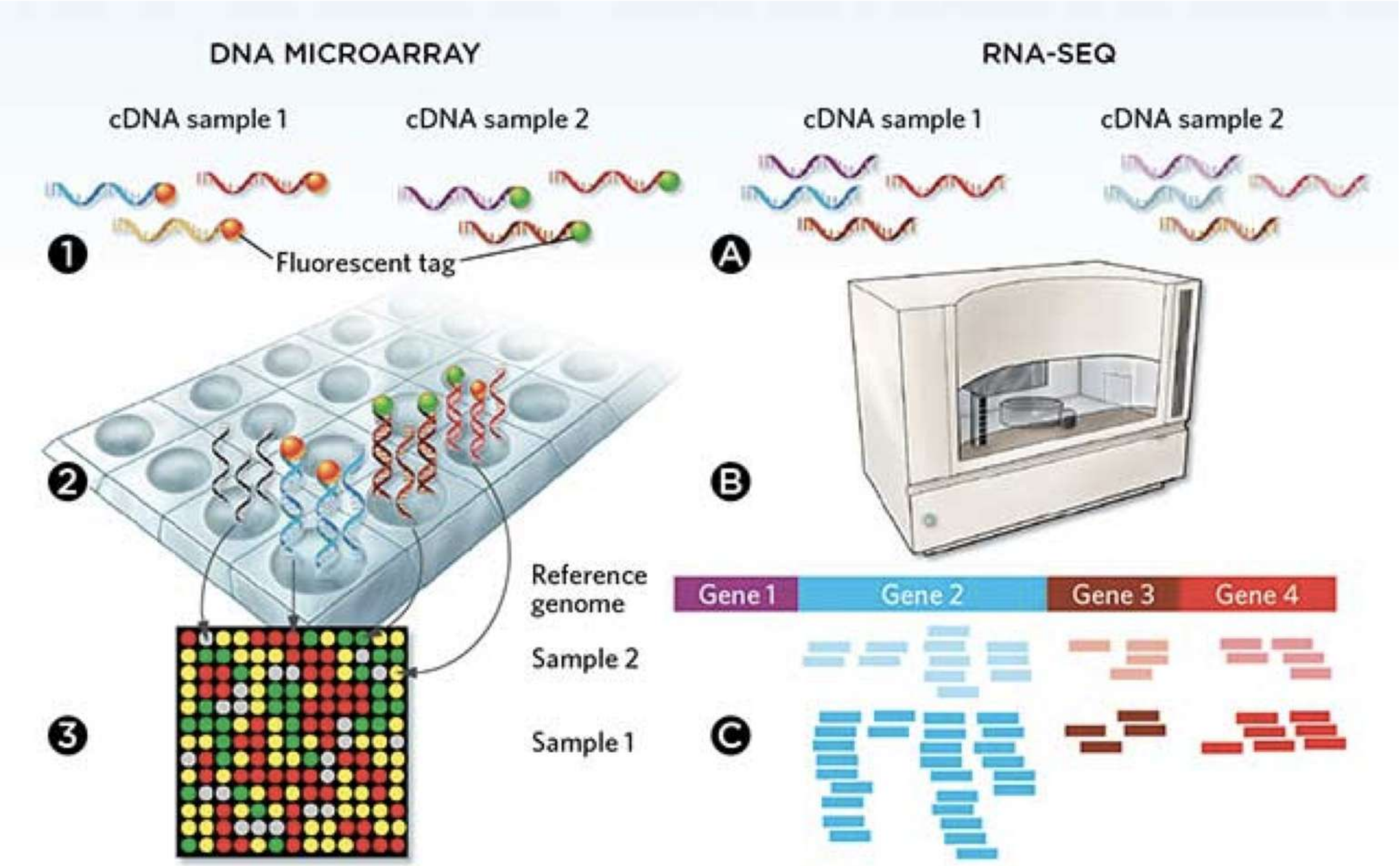


Gene Set Enrichment Analysis

Fernanda Díaz Espinosa

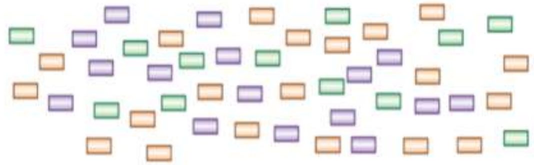

DNA MICROARRAY
RNA-SEQ
cDNA sample 1
cDNA sample 2
cDNA sample 1
cDNA sample 2
1
Fluorescent tag
A
2
B
Reference genome
Gene 1
Gene 2
Gene 3
Gene 4
Sample 2
Sample 1
3
C

What do we do with that MUCH information?

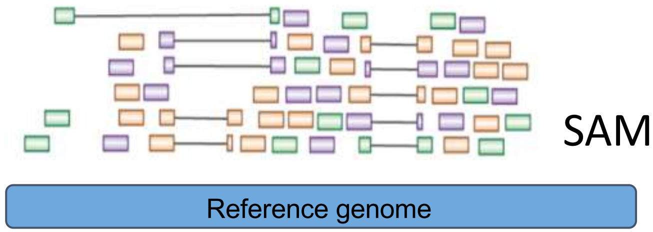
Fastq

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Trimming and QC

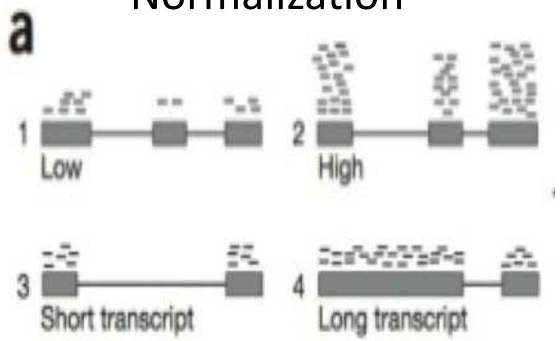


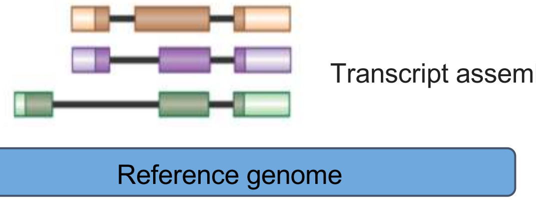
Reading Alignment



SAM/BAM

Normalization





Transcript assembly

Count

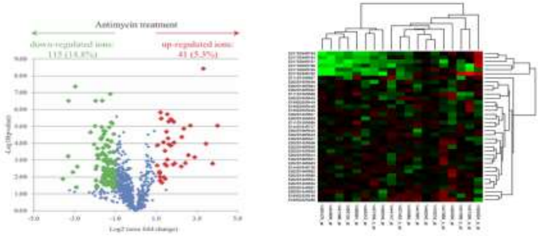
GENE	A	B	C
CTCF	13	1	7
BORIS	0	18	5

Differential Expression

D.E table

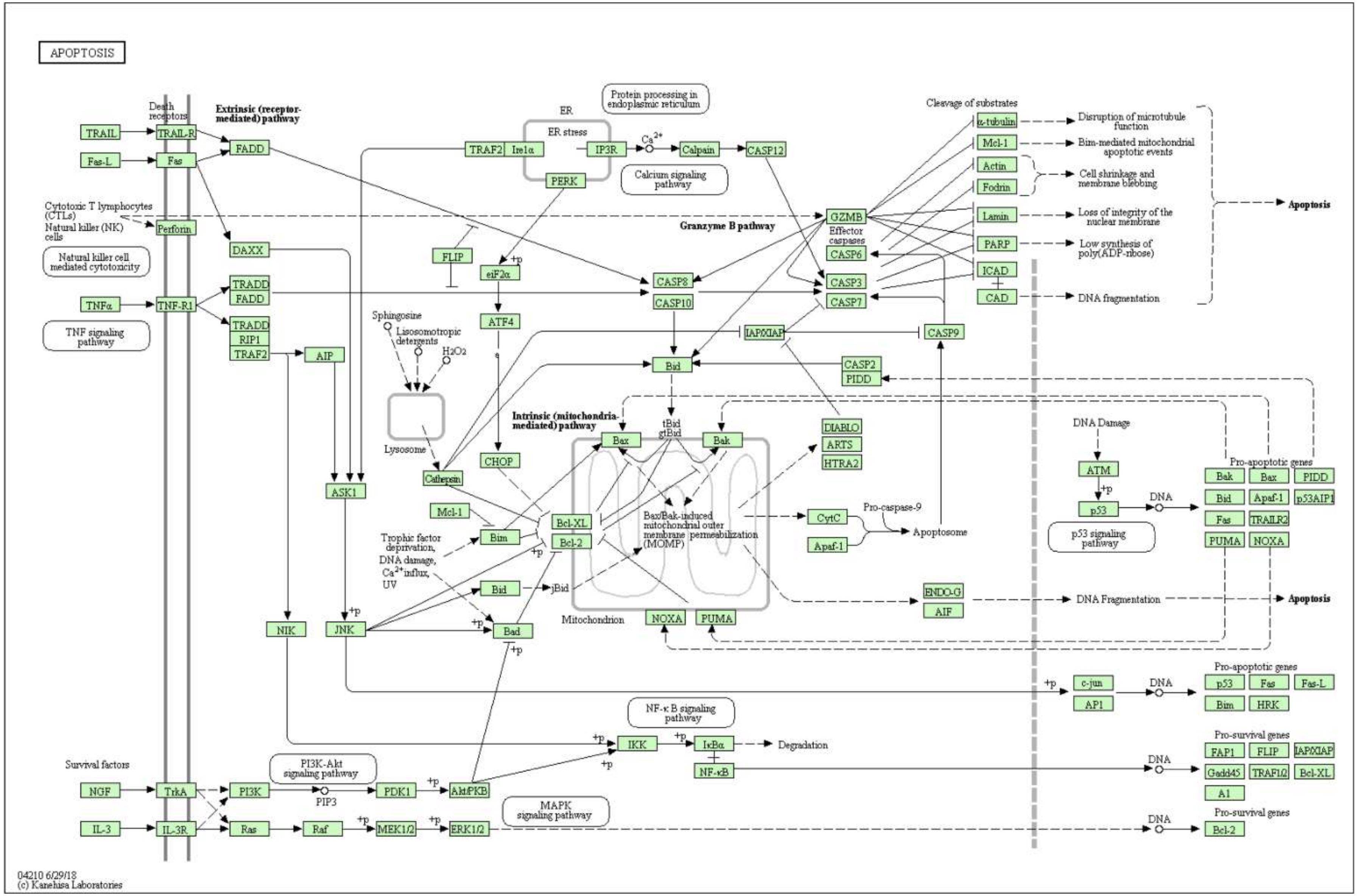
GENE	FC	p-value	FDR
CTCF	-2	0.006	0.00013
BORIS	+2	0.0009	0.0003

Data display

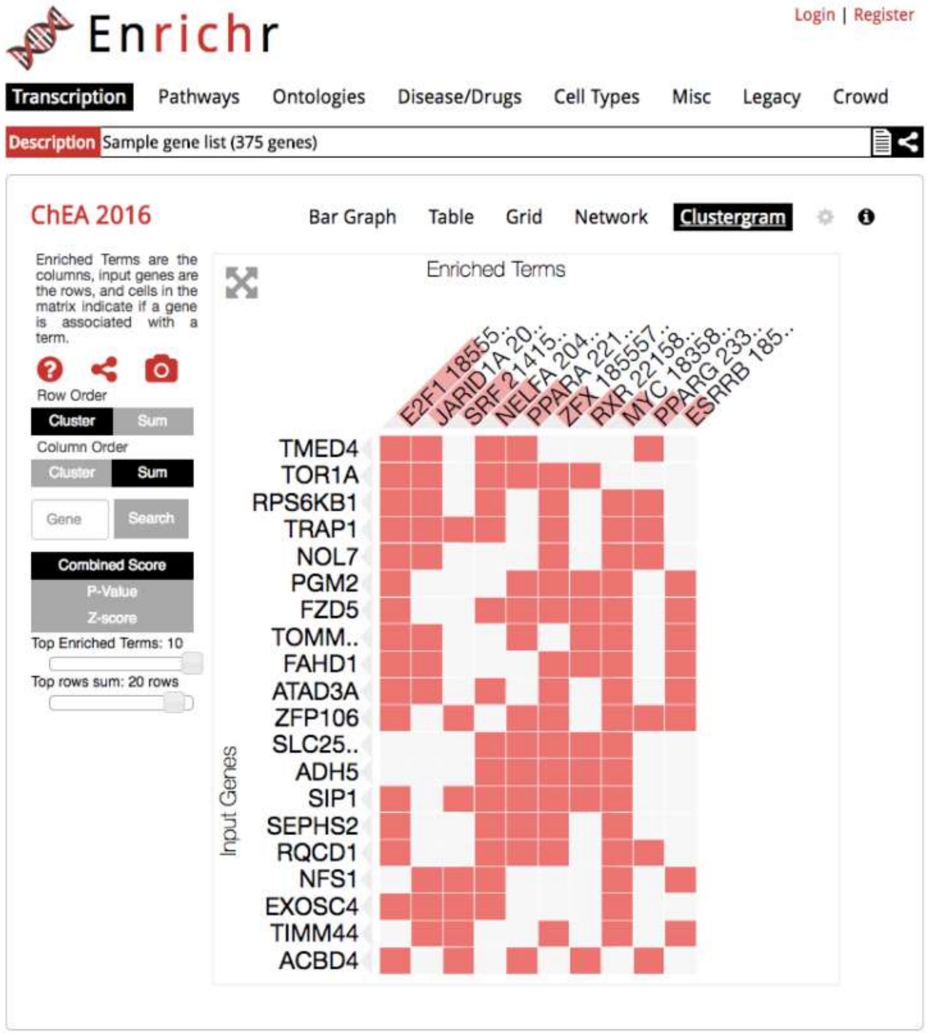


What do we do with that MUCH information?

Pathway analysis



Pathway analysis



Enrichr

Login | Register

Transcription Pathways Ontologies Disease/Drugs Cell Types Misc Legacy Crowd

Description Sample gene list (375 genes)

ChEA 2016

Bar Graph Table Grid Network Clustergram

Enriched Terms are the columns, input genes are the rows, and cells in the matrix indicate if a gene is associated with a term.

Row Order: Cluster Sum

Column Order: Cluster Sum

Gene Search

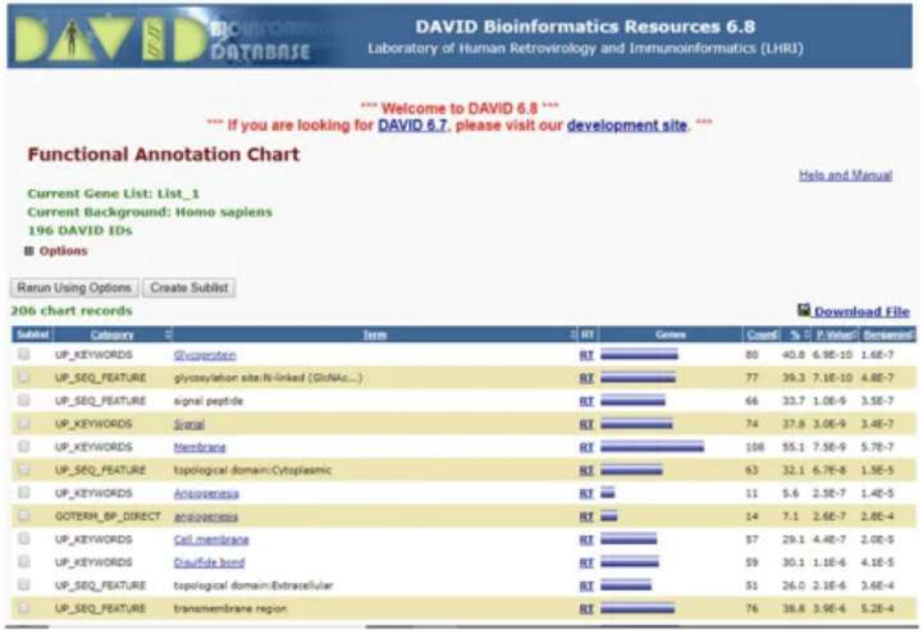
Combined Score

P-Value

Z-score

Top Enriched Terms: 10

Top rows sum: 20 rows



DAVID Bioinformatics Resources 6.8

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

*** Welcome to DAVID 6.8 ***

*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Chart

Help and Manual

Current Gene List: List_1

Current Background: Homo sapiens

196 DAVID IDs

⊞ Options

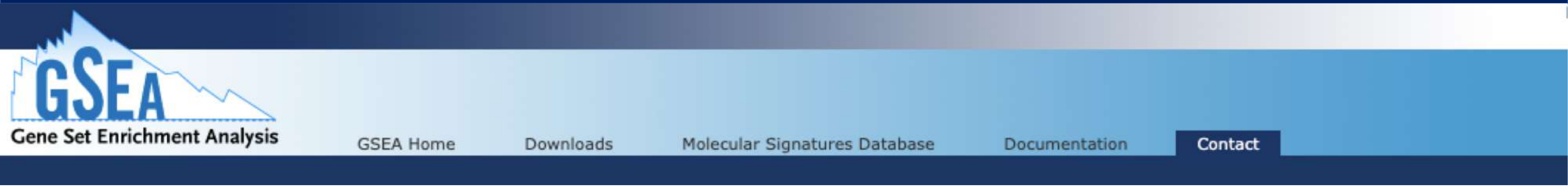
Rerun Using Options Create Sublist

206 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	UP_KEYWORDS	Glycoprotein	RT		80	40.8	6.9E-10	1.6E-7
□	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		77	39.3	7.1E-10	4.8E-7
□	UP_SEQ_FEATURE	signal peptide	RT		66	33.7	1.0E-9	3.5E-7
□	UP_KEYWORDS	Signal	RT		74	37.8	3.0E-9	3.4E-7
□	UP_KEYWORDS	Membrane	RT		108	55.1	7.5E-9	5.7E-7
□	UP_SEQ_FEATURE	topological domain:Cytoplasmic	RT		63	32.1	6.7E-8	1.5E-5
□	UP_KEYWORDS	Angiogenesis	RT		11	5.6	2.5E-7	1.4E-5
□	GOTERM_BP_DIRECT	angiogenesis	RT		14	7.1	2.6E-7	2.8E-4
□	UP_KEYWORDS	Cell membrane	RT		57	29.1	4.4E-7	2.0E-5
□	UP_KEYWORDS	Disulfide bond	RT		59	30.1	1.1E-6	4.1E-5
□	UP_SEQ_FEATURE	topological domain:Extracellular	RT		51	26.0	2.1E-6	3.6E-4
□	UP_SEQ_FEATURE	transmembrane region	RT		76	38.8	3.9E-6	5.2E-4

⬇ And the Fold Change? ⬆

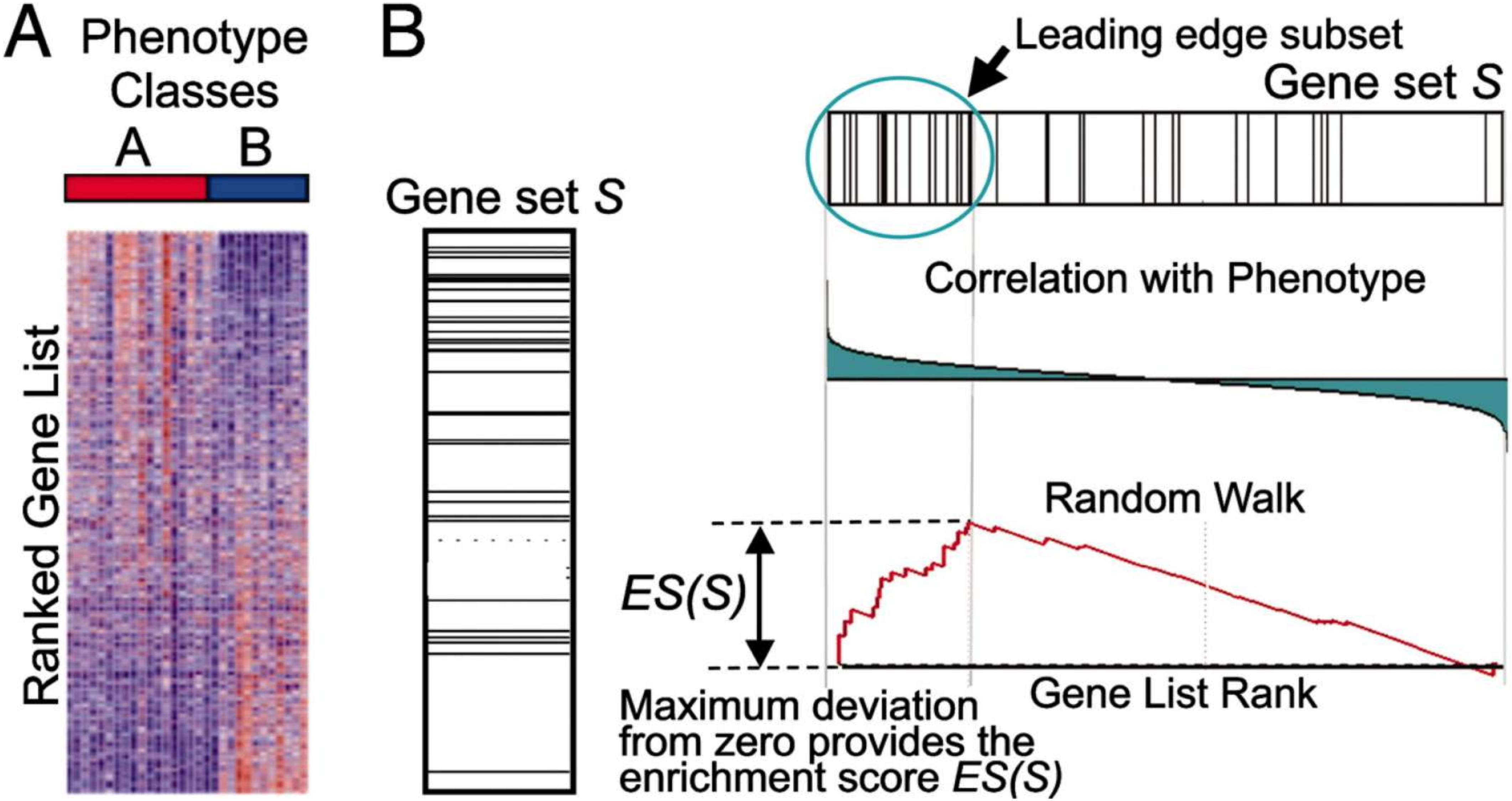


First described by Mootha et al., 2003.

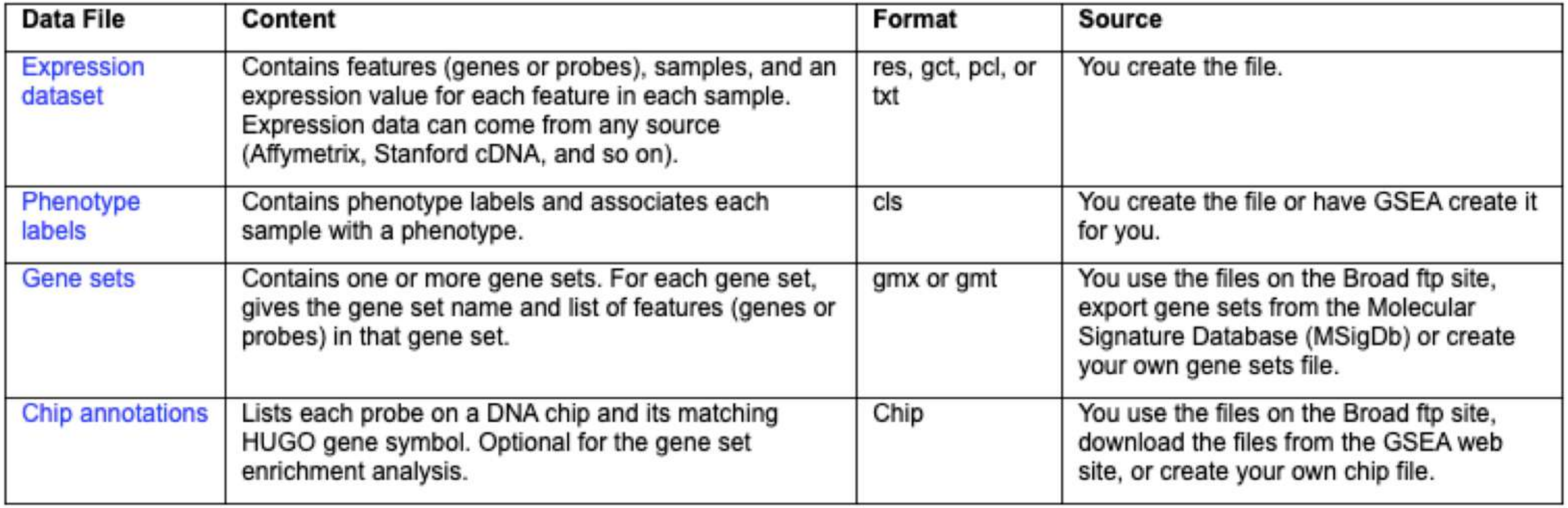
- Analysis of second generation pathways.
- Significance analysis of function and expression (SAFE).

- Four step approach.

A) Calculate the local statistic (at the gene level).
B) Calculate the global statistic (at the level of pathway or set of genes).
C) Determine a significance value.
D) Adjust for multiple comparisons.


A
Phenotype Classes
A
B
Ranked Gene List
B
Gene set *S*
Leading edge subset
Gene set *S*
Correlation with Phenotype
Random Walk
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

Files to start



Data File	Content	Format	Source
Expression dataset	Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).	res, gct, pcl, or txt	You create the file.
Phenotype labels	Contains phenotype labels and associates each sample with a phenotype.	cls	You create the file or have GSEA create it for you.
Gene sets	Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.	gmx or gmt	You use the files on the Broad ftp site, export gene sets from the Molecular Signature Database (MSigDb) or create your own gene sets file.
Chip annotations	Lists each probe on a DNA chip and its matching HUGO gene symbol. Optional for the gene set enrichment analysis.	Chip	You use the files on the Broad ftp site, download the files from the GSEA web site, or create your own chip file.

Choose your database

- Gene set Databases.

  a) Gene Ontology (GO).

  b) Molecular Signatures Database (MSigDB).

- Detailed biochemical pathway databases.

  a) Reactome.

  b) Panther.

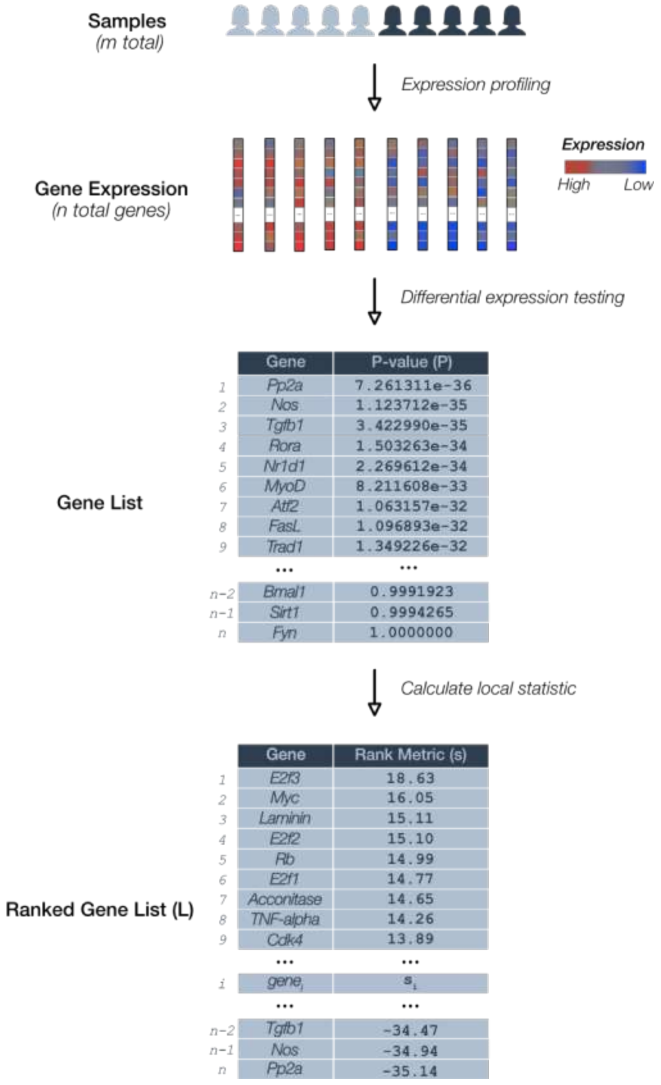
  c) NetPath.

  d) HumanCyc.

  e) National Cancer Institute (NCI) Pathway Interaction Database (PID).

  f) KEGG.

1.Local statistic



	Gene	P-value (P)
1	Pp2a	7.261311e-36
2	Nos	1.123712e-35
3	Tgfb1	3.422990e-35
4	Rora	1.503263e-34
5	Nr1d1	2.269612e-34
6	MyoD	8.211608e-33
7	Atf2	1.063157e-32
8	FasL	1.096893e-32
9	Trad1	1.349226e-32
	...	...
n−2	Bmal1	0.9991923
n−1	Sirt1	0.9994265
n	Fyn	1.0000000

	Gene	Rank Metric (s)
1	E2f3	18.63
2	Myc	16.05
3	Laminin	15.11
4	E2f2	15.10
5	Rb	14.99
6	E2f1	14.77
7	Acconitase	14.65
8	TNF-alpha	14.26
9	Cdk4	13.89
	...	...
i	$gene_i$	s_i
	...	...
n−2	Tgfb1	-34.47
n−1	Nos	-34.94
n	Pp2a	-35.14

The rank file is a list of detected genes and a rank metric score.

- At the top of the list are genes with the "strongest" up-regulation.
- At the bottom of the list are the genes with the "strongest" down-regulation.
- The genes not changing are in the middle.

Metric

- Signal-to-noise ratio (S2N; the default measure in GSEA)
- Absolute value of signal-to-noise ratio (|S2N|)
- Difference of expression means between classes (Ratio)
- Log_2 of Ratio (log_2(Ratio))
- T-test statistic (T-test).

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1674-0

- **Signal2Noise** (default) uses the difference of means scaled by the standard deviation.

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

where μ is the mean and σ is the standard deviation; σ has a minimum value of .2 * absolute(μ), where μ=0 is adjusted to μ=1. The larger the signal-to-noise ratio, the larger the differences of the means (scaled by standard deviations);

- **tTest** uses the difference of means scaled by the standard deviation and number of samples.

$$\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

where μ is the mean, n is the number of samples, and σ is the standard deviation; σ has a minimum value of .2 * absolute(μ), where μ=0 is adjusted to μ=1. The larger the tTest ratio, the more distinct the gene expression is in each phenotype and the more the gene acts as a “class marker.”

- **Ratio_of_Classes** (also referred to as fold change) uses the ratio of class means to calculate fold change for natural scale data:

$$\frac{\mu_A}{\mu_B}$$

where μ is the mean. The larger the fold change, the more distinct the gene expression is in each phenotype and the more the gene acts as a “class marker.”

- **Diff_of_Classe**s uses the difference of class means to calculate fold change for log scale data:

$$\mu_A - \mu_B$$

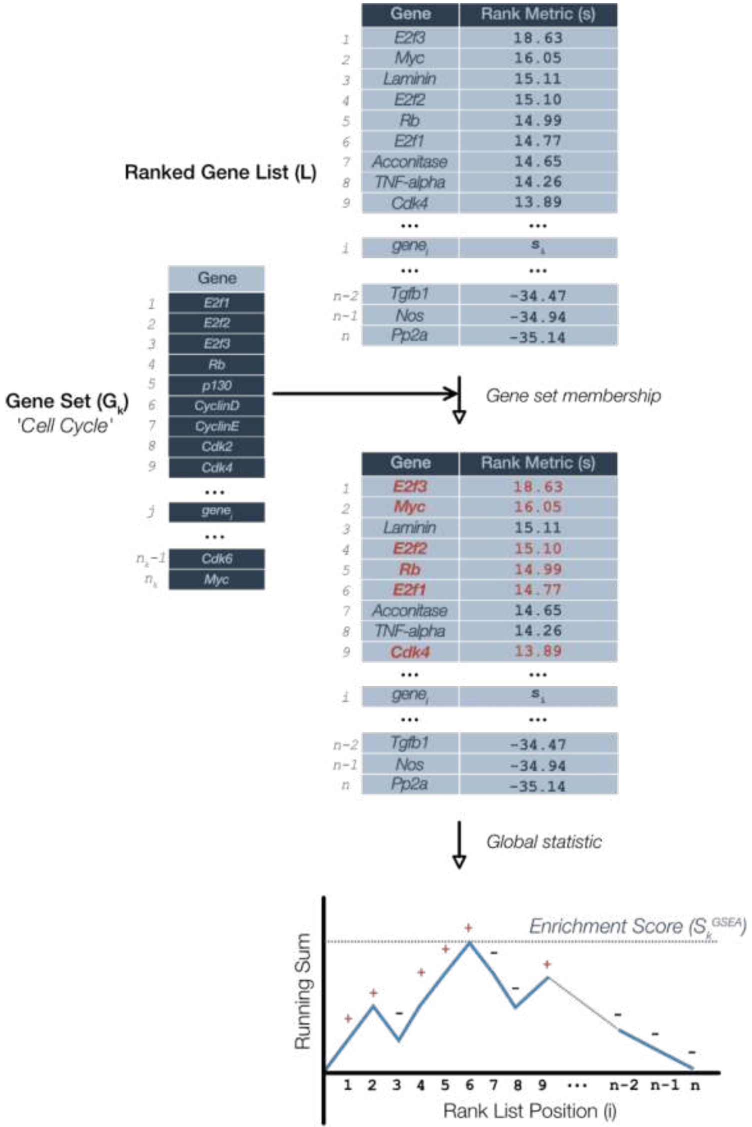
where μ is the mean. The larger the fold change, the more distinct the gene expression is in each phenotype and the more the gene acts as a “class marker.”

- **log2_Ratio_of_Classes** uses the log2 ratio of class means to calculate fold change for natural scale data:

$$\log 2\left(\frac{\mu_A}{\mu_B}\right)$$

where μ is the mean. This is the recommended statistic for calculating fold change for natural scale data.

2. Global statistic



	Gene	Rank Metric (s)
1	E2f3	18.63
2	Myc	16.05
3	Laminin	15.11
4	E2f2	15.10
5	Rb	14.99
6	E2f1	14.77
7	Acconitase	14.65
8	TNF-alpha	14.26
9	Cdk4	13.89
	...	...
i	$gene_i$	s_i
	...	...
n-2	Tgfb1	-34.47
n-1	Nos	-34.94
n	Pp2a	-35.14

	Gene
1	E2f1
2	E2f2
3	E2f3
4	Rb
5	p130
6	CyclinD
7	CyclinE
8	Cdk2
9	Cdk4
	...
j	$gene_j$
	...
n_k-1	Cdk6
n_k	Myc

	Gene	Rank Metric (s)
1	E2f3	18.63
2	Myc	16.05
3	Laminin	15.11
4	E2f2	15.10
5	Rb	14.99
6	E2f1	14.77
7	Acconitase	14.65
8	TNF-alpha	14.26
9	Cdk4	13.89
	...	...
i	$gene_i$	s_i
	...	...
n-2	Tgfb1	-34.47
n-1	Nos	-34.94
n	Pp2a	-35.14

Enrichment Score (ES)

Reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes.

GSEA calculates the ES by walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not.

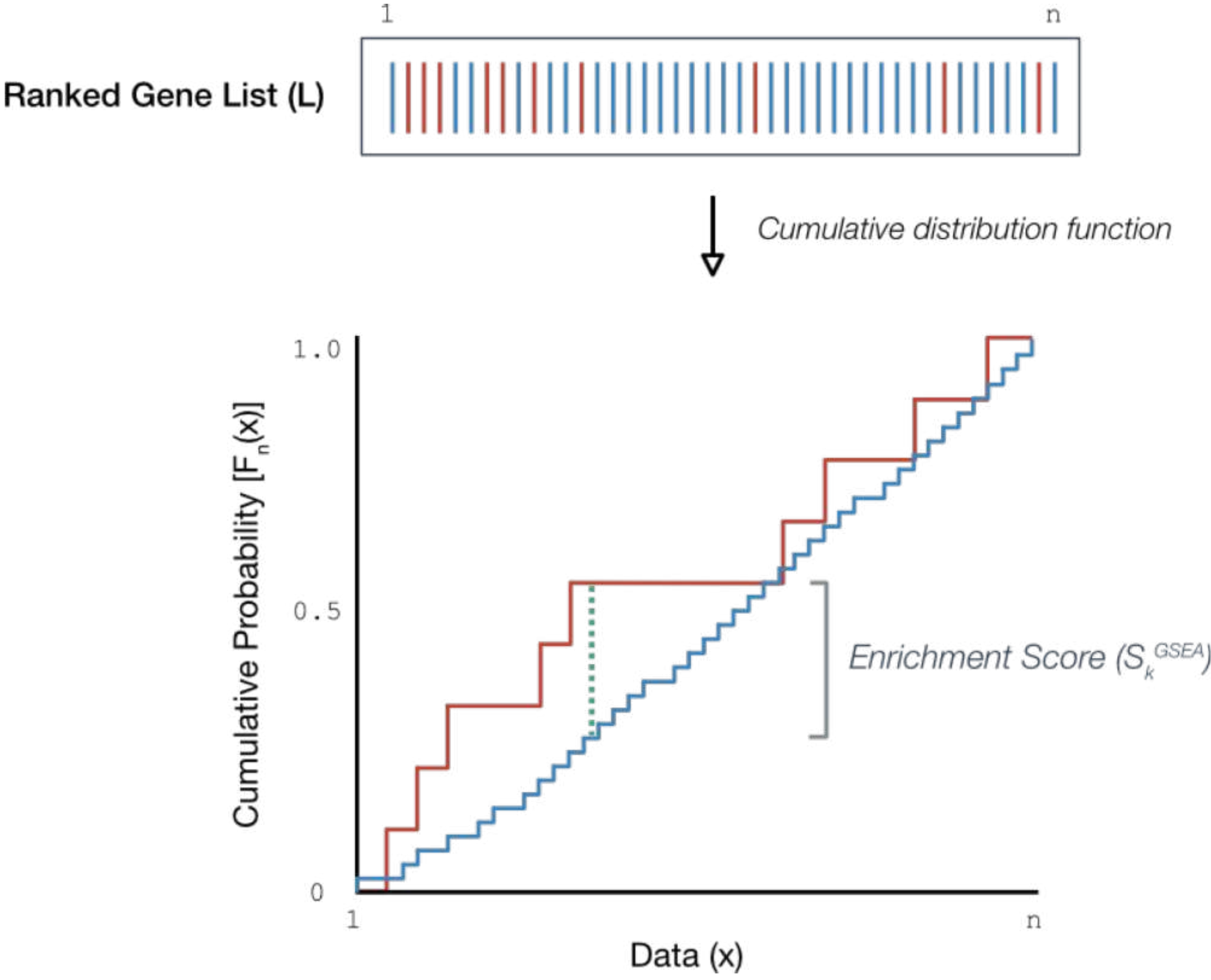
- The magnitude of the increment depends on the correlation of the gene with the phenotype.
- The ES is the maximum deviation from zero encountered in walking the list.
- A positive ES indicates gene set enrichment at the top of the ranked list; a negative
- ES indicates gene set enrichment at the bottom of the ranked list.

Kolmogorov–Smirnov test

The Kolmogorov–Smirnov statistic quantifies
a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution.

…Or between the empirical distribution functions of two samples.

The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution.



3. “Significance” value

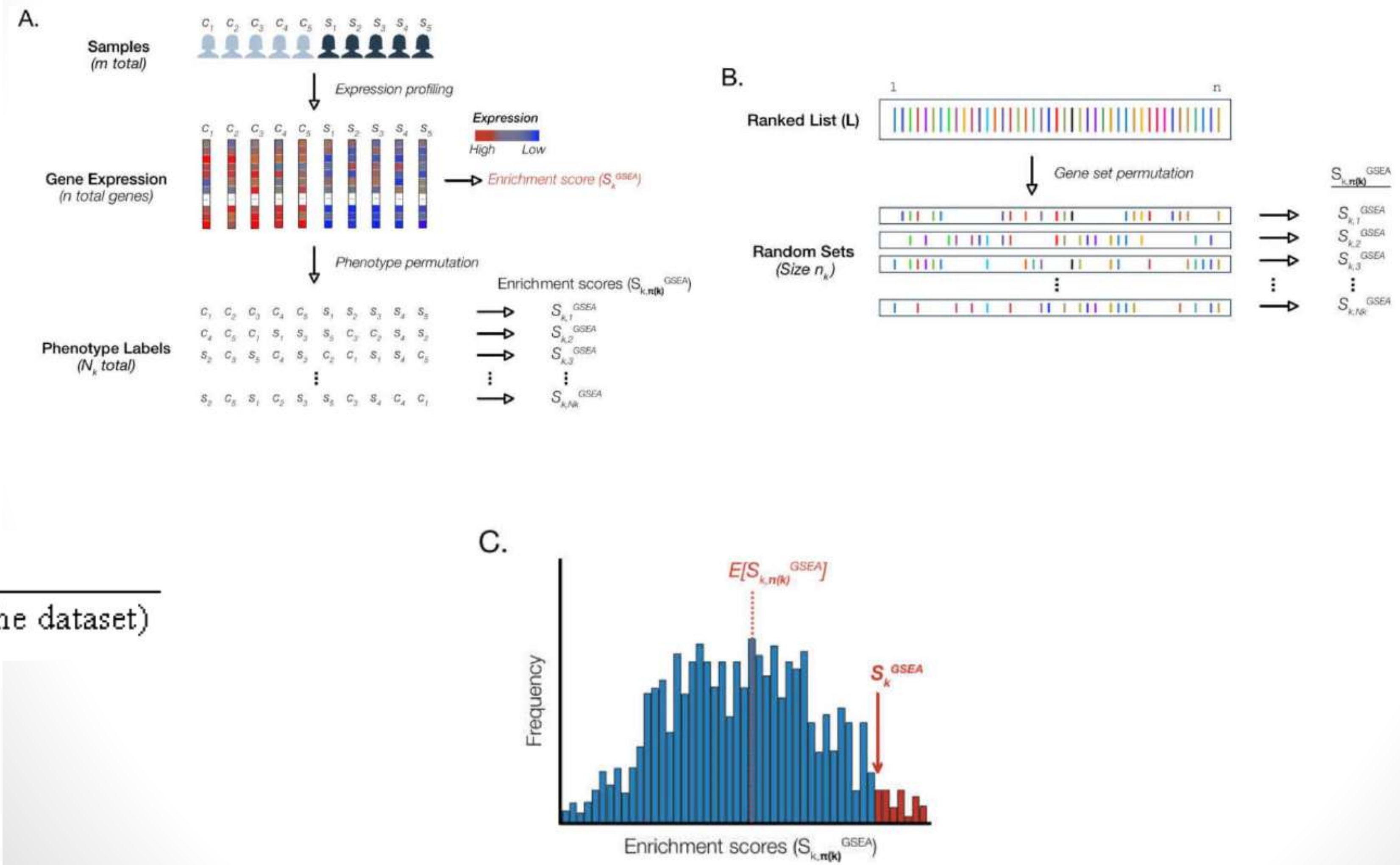
Normalized Enrichment Score (NES)

The normalized enrichment score (NES) is the primary statistic for examining gene set enrichment results.

By normalizing the enrichment score, GSEA accounts for differences in gene set size and in correlations between gene sets and the expression dataset.

Therefore, the normalized enrichment scores (NES) can be used to compare analysis results across gene sets.

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



4. Adjust for multiple comparisons.

False Discovery Rate (FDR)

The false discovery rate (FDR) is the estimated probability that a gene set with a given NES represents a false positive finding.

For example, an FDR of 25% indicates that the result is likely to be valid 3 out of 4 times.

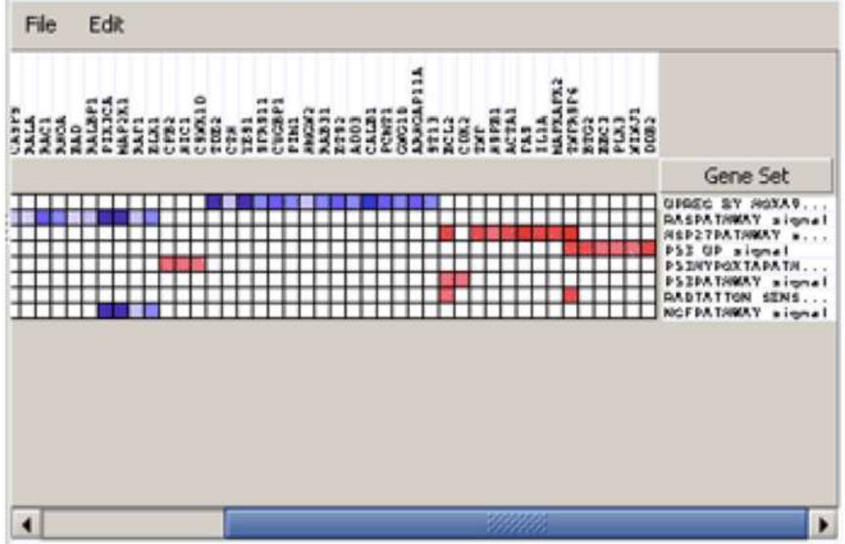
When we test a family of hypotheses, the chance of observing a statistic with a small p-value increases.
When smaller than the significance level, they can be erroneously classified as discoveries or Type I errors.
Multiple testing procedures attempt to quantify and control for these.

$$F\hat{D}R = \frac{s_{null}}{s_{obs}} \cdot \frac{n_{obs}}{n_{null}}$$

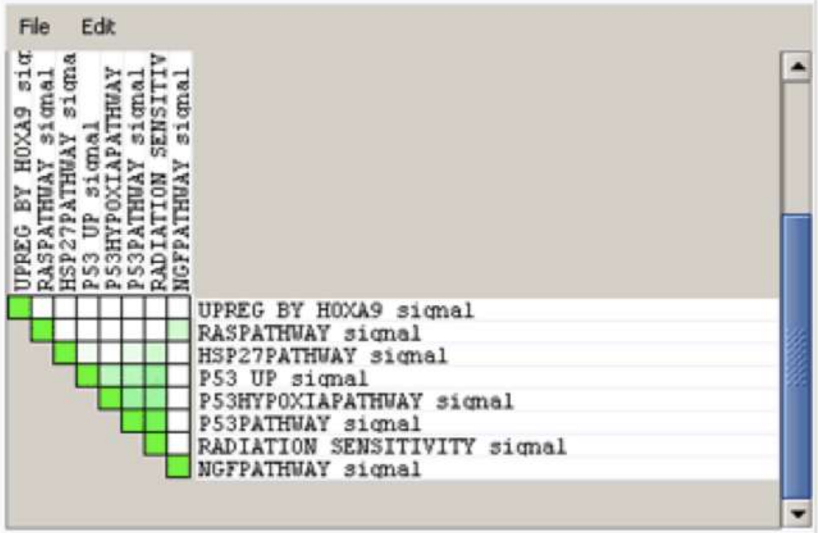
In GSEA, the collection of gene sets interrogated against the observed data are a family of hypotheses.

The recommended procedure for quantifying Type I errors is the false discovery rate (FDR). The FDR is defined as the expected value of the fraction of rejected null hypotheses that are in fact true. In practice, GSEA establishes this proportion empirically.

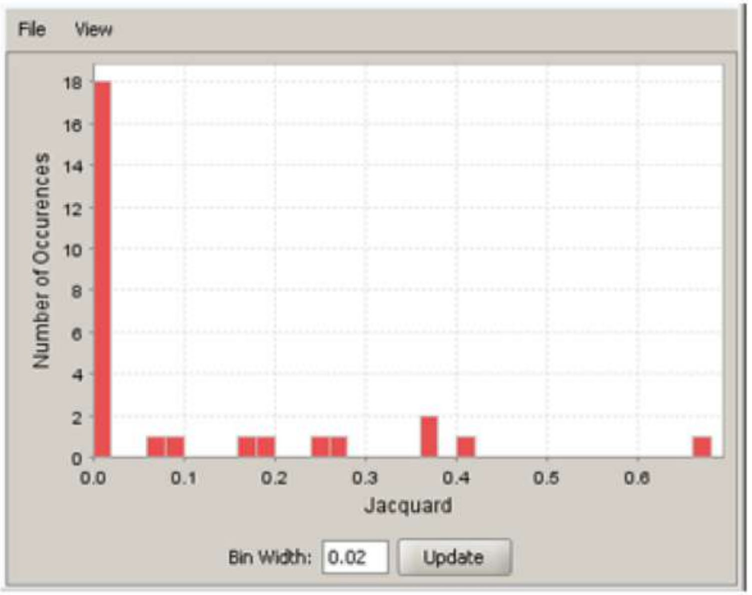
Graphs



HEATMAP

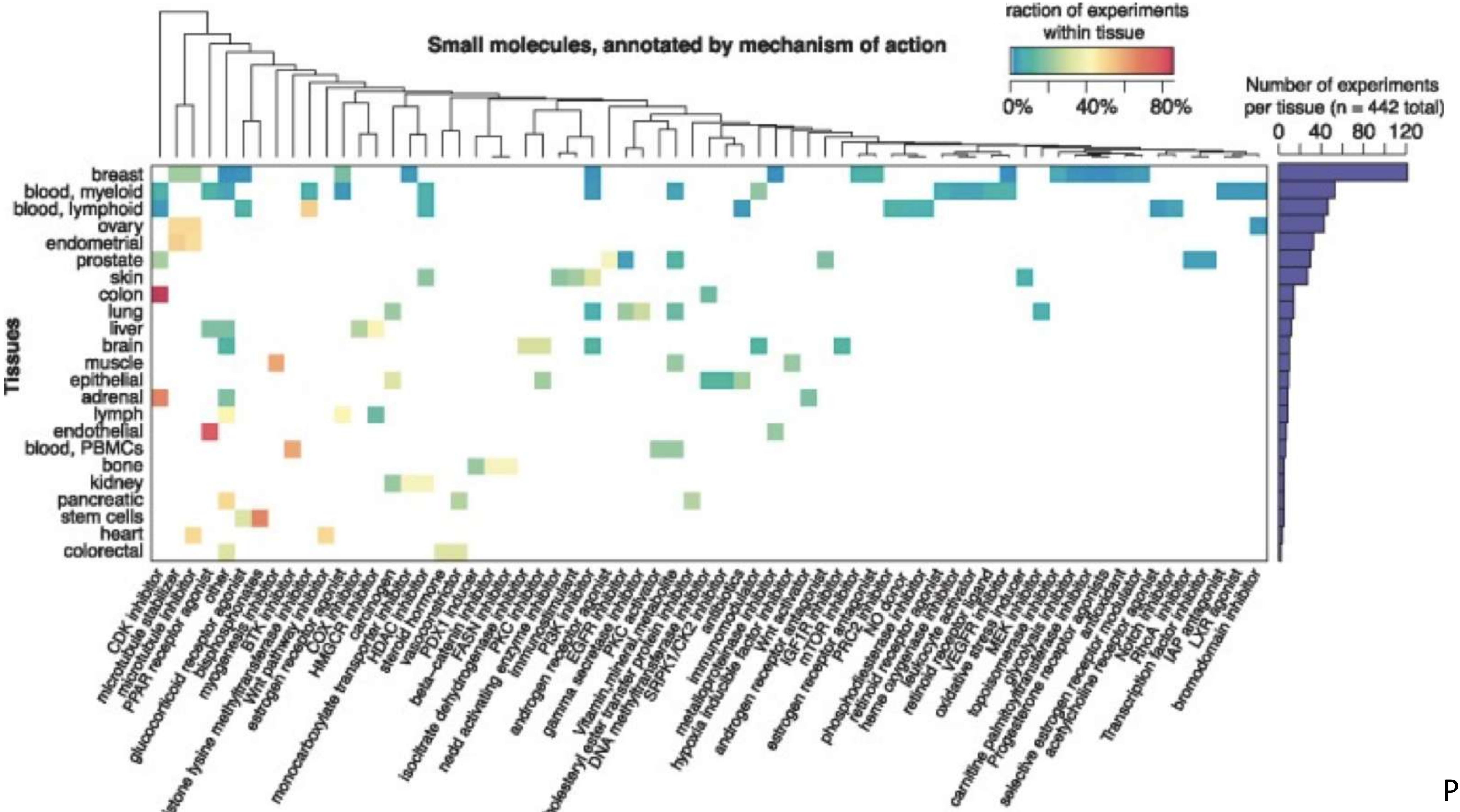


SET TO SET



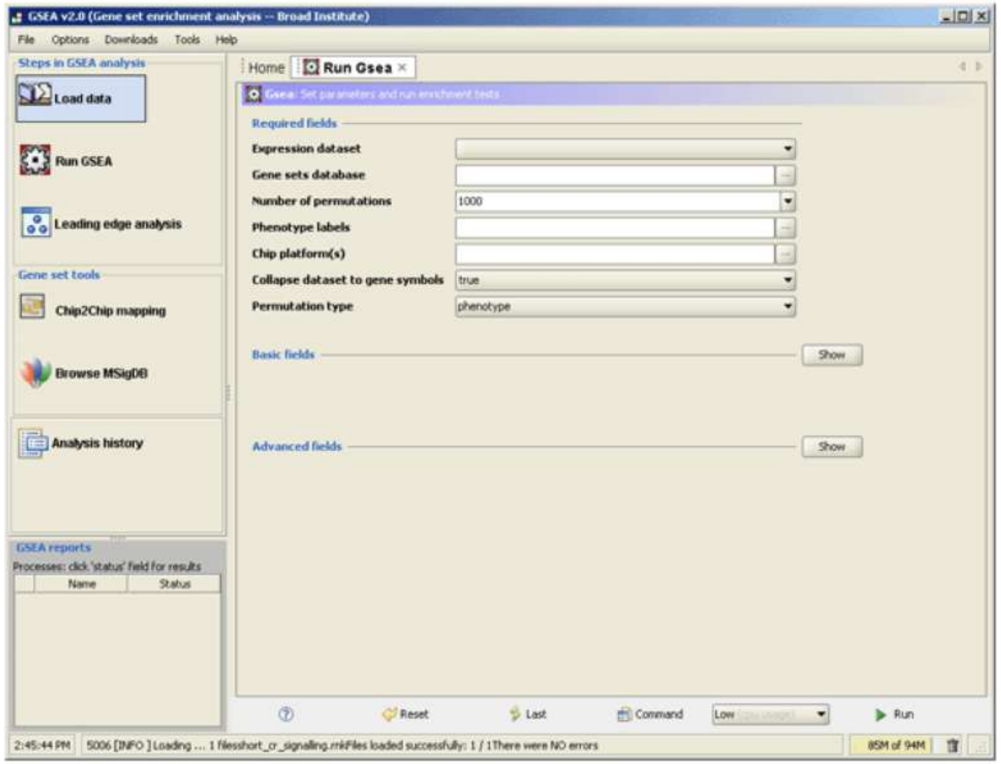
Bin Width: 0.02

Histograms


Small molecules, annotated by mechanism of action
raction of experiments within tissue
0%
40%
80%
Number of experiments per tissue (n = 442 total)
0
40
80
120
Tissues
breast
blood, myeloid
blood, lymphoid
ovary
endometrial
prostate
skin
colon
lung
liver
brain
muscle
epithelial
adrenal
lymph
endothelial
blood, PBMCs
bone
kidney
pancreatic
stem cells
heart
colorectal
CDK inhibitor
microtubule stabilizer
microtubule inhibitor
PPAR receptor agonist
other
glucocorticoid receptor agonist
bisphosphonates
myogenesis inhibitor
BTK inhibitor
istone lysine methyltransferase inhibitor
Wnt pathway inhibitor
estrogen receptor agonist
COX inhibitor
HMGCR inhibitor
carcinogen
monocarboxylate transporter inhibitor
HDAC inhibitor
steroid hormone
vasoconstrictor
PDX1 inducer
beta−catenin inhibitor
FASN inhibitor
isocitrate dehydrogenase inhibitor
PKC inhibitor
nedd activating enzyme inhibitor
immunostimulant
PI3K inhibitor
androgen receptor agonist
EGFR inhibitor
gamma secretase inhibitor
PKC activator
Vitamin,mineral,metabolite
olesteryl ester transfer protein inhibitor
DNA methyltransferase inhibitor
SRPK1/CK2 inhibitor
antibiotics
immunomodulator
metalloproteinase inhibitor
hypoxia inducible factor inhibitor
Wnt activator
androgen receptor antagonist
IGF1R inhibitor
mTOR inhibitor
estrogen receptor antagonist
PRC2 inhibitor
NO donor
phosphodiesterase inhibitor
retinoid receptor agonist
heme oxygenase inhibitor
leukocyte activator
retinoid receptor ligand
VEGFR inhibitor
oxidative stress inducer
MEK inhibitor
topoisomerase inhibitor
glycolysis inhibitor
carnitine palmitoyltransferase inhibitor
Progesterone receptor agonists
antioxidant
selective estrogen receptor modulator
acetylcholine receptor agonist
Notch inhibitor
RhoA inhibitor
Transcription factor inhibitor
IAP antagonist
LXR agonist
bromodomain inhibitor

Powers 2018

Where to run



Expression dataset:

Gene sets database:

Number of permutations: 1000

Phenotype labels:

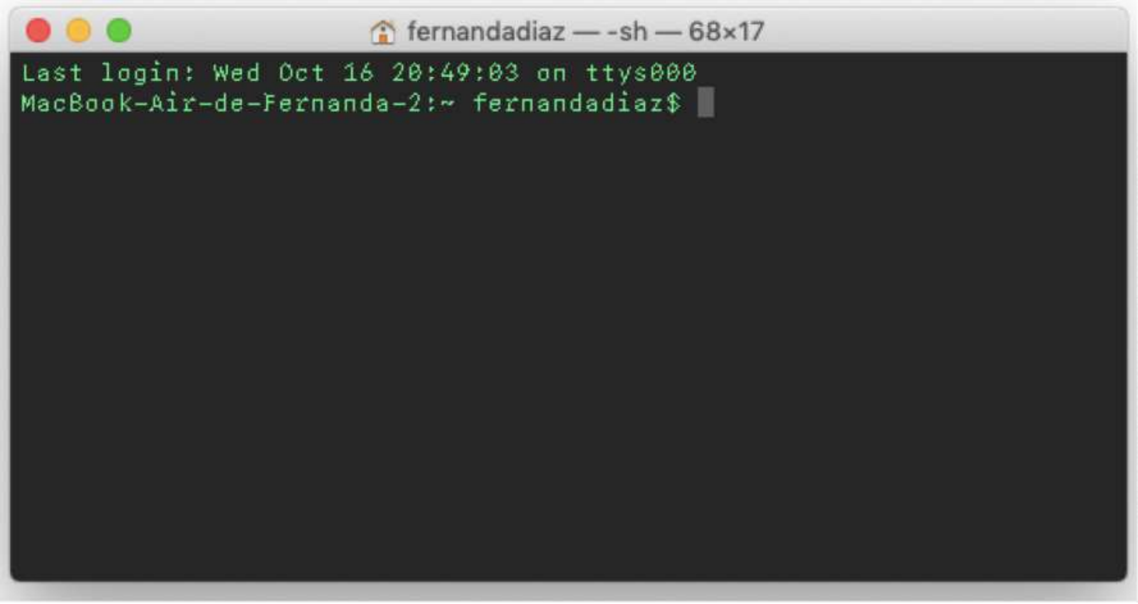
Chip platform(s):

Collapse dataset to gene symbols: true

Permutation type: phenotype

```
2:45:44 PM  5006 [INFO ] Loading ... 1 filesshort_cr_signaling.rnkFiles loaded successfully: 1 / 1There were NO errors
```

Java desktop app



```
Last login: Wed Oct 16 20:49:03 on ttys000
MacBook-Air-de-Fernanda-2:~ fernandadiaz$ 
```

Command line

R package

Alternatives

WebGestalt

WEB-based GEne SeT AnaLysis Toolkit

Translating gene lists into biological insights...

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550.

https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/