RNAseq interpretation – Gene Ontology

- An ontology is a specification of a conceptualization:
  - a hierarchical mapping of concepts within a given frame of reference.
- An ontology is a restricted structured vocabulary of terms that represent domain knowledge.
- An ontology specifies a vocabulary that can be used to exchange queries and assertions.
- A commitment to the use of the ontology is an agreement to use the shared vocabulary in a consistent way.

RNAseq interpretation – Gene Ontology

- The goal of the Gene Ontology (GO) Consortium is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.
  - http://www.geneontology.org/

- For genes and gene products the Gene Ontology Consortium (GO) is an initiative that is designed to address the problem of defining common set of terms and descriptions for basic biological functions.

- GO provides a restricted vocabulary as well as clear indications of the relationships between terms.

RNAseq interpretation – Gene Ontology

- The Gene Ontology (GO) consortium produces three independent ontologies for gene products.
- The three ontologies are:
  - *molecular function* of a gene product which is defined to be biochemical activity or action of the gene product (MF 7220).
  - *biological process* interpreted as a biological objective to which the gene product contributes (BP 9529).
  - *cellular component* is a component of a cell that is part of some larger object or structure (CC 1536).

RNAseq interpretation – Gene Ontology

- The GO ontologies are structured as directed acyclic graphs (DAGs) that represent a network in which each term may be a child of one or more parents.
- GO node is interchangeable with GO term.
- Child terms are more specific than their parents:
  - The term “transmembrane receptor protein-tyrosine kinase” is child of
    - “transmembrane receptor” and “protein tyrosine kinase”.

RNAseq interpretation – Gene Ontology

- The relationship between a child and a parent can be characterized by the relations:
  - is a
  - has a (part of)
  - Positive/negative regulation (BP only)
- "mitotic chromosome" *is a child of* "chromosome" and the relationship is an is a relation.
- "telomere" *is a child of* "chromosome" with the has a relation.

RNAseq interpretation – Gene Ontology – is_a

- The *is_a* relationship is a simple class-subclass relationship, where A *is_a* B means that A is a subclass of B; for example, **nuclear chromosome** *is_a* **chromosome**.

```
GO:0043232 : intracellular non-membrane-bound organelle
[i] GO:0005694 : chromosome
---[i] GO:0000228 : nuclear chromosome
```

RNAseq interpretation – Gene Ontology part_of

- The *part_of* relationship is slightly more complex; C *part_of* D means that whenever C is present, it is always a part of D, but C does not always have to be present. An example would be **periplasmic flagellum** *part_of* **periplasmic space**:

```
GO:0044464 : cell part
[i] GO:0042995 : cell projection
---[i] GO:0019861 : flagellum
------[i] GO:0009288 : flagellin-based flagellum
---------[i] GO:0055040 : periplasmic flagellum
[i] GO:0042597 : periplasmic space
---[p] GO:0055040 : periplasmic flagellum
```

When a **periplasmic flagellum** is present, it is always *part_of* a **periplasmic space**. However, every **periplasmic space** does not necessarily have a **periplasmic flagellum**.

RNAseq interpretation – Gene Ontology - regulates

- The *regulates*, *positively_regulates* and *negatively_regulates* relationships describe interactions between biological processes and other biological processes, molecular functions or biological qualities. When a biological process E regulates a function or a process F, it modulates the occurrence of F. If F is a biological quality, then E modulates the value of F. An example of the regulation of a biological process would be the term regulation of transcription. When regulation of transcription occurs, it always alters the rate, extent or frequency at which a gene is transcribed.

RNAseq interpretation – Gene Ontology – transitivity rule

The *is_a* and *part_of* relationships are transitive, which means that the relationships are propagated from parent terms to child terms. An example of *is_a* transitivity is shown in the nuclear chromosome example previously used:

```
GO:0043232 : intracellular non-membrane-bound organelle
[i] GO:0005694 : chromosome
---[i] GO:0000228 : nuclear chromosome
```

All **nuclear chromosomes** must be **intracellular non-membrane-bound organelles**.
An example of *part_of* transitivity is shown below:

```
GO:0048869 : cellular developmental process
[i] GO:0030154 : cell differentiation
---[p] GO:0048468 : cell development
------[p] GO:0000904 : cellular morphogenesis during differentiation
```

Every occurrence of **cellular morphogenesis during differentiation** must be a part of an occurrence of **cell differentiation**.

RNAseq interpretation – Gene Ontology – transitivity rule

is_a transitivity: If process B exists in the GO biological process ontology and it is an *is_a* child of process A then any process that regulates process B also regulates process A. For example:

```
GO:0016049 : cell growth
[i] GO:0042815 : bipolar cell growth
---[r] GO:0051516 : regulation of bipolar cell growth
```

Due to *is_a* transitivity, we can say that any process that regulates **bipolar cell growth** also regulates **cell growth**.

RNAseq interpretation – Gene Ontology – transitivity rule

The regulates relationships are transitive over *part_of* relationship.

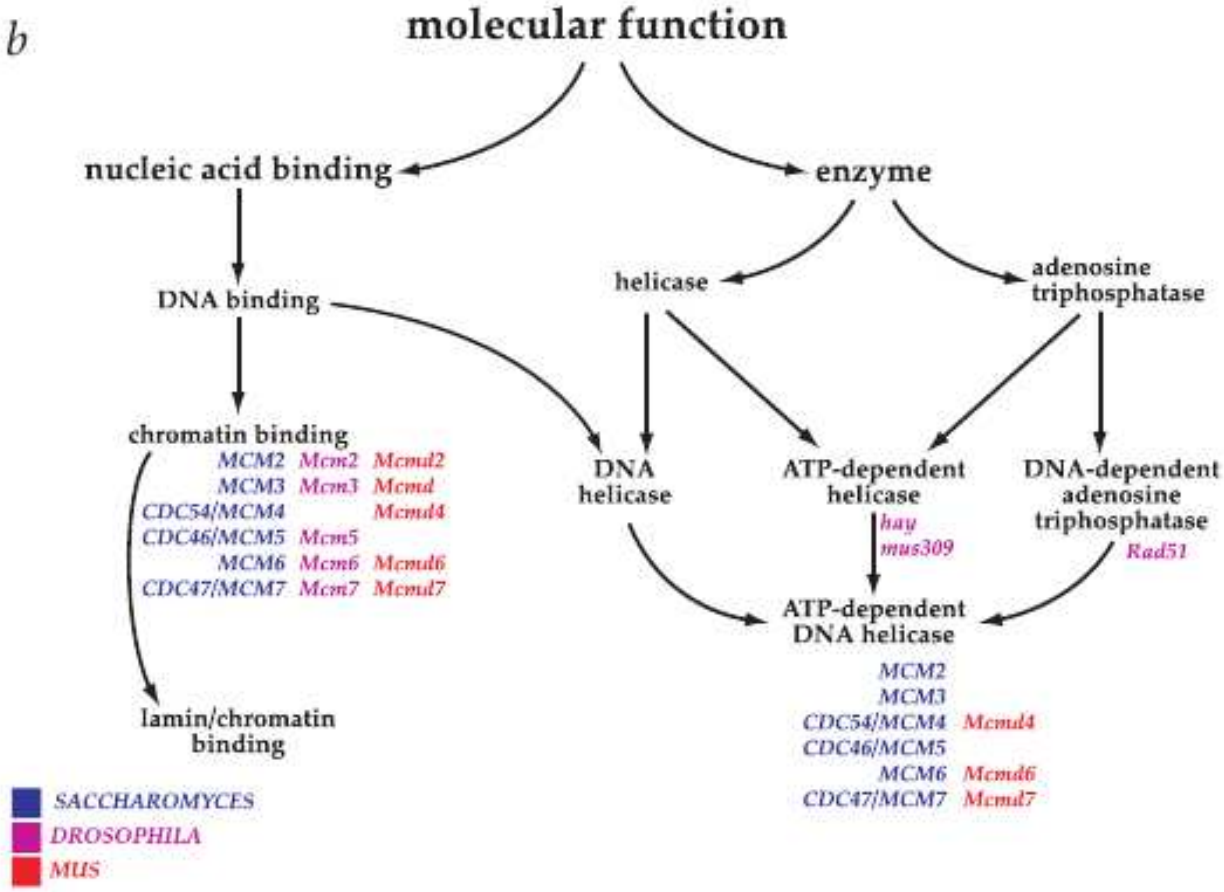
```
GO:0010467 : gene expression
[r] GO:0010468 : regulation of gene expression
---[i] GO:0045449 : regulation of transcription
[p] GO:0006350 : transcription
---[r] GO:0045449 : regulation of transcription
```

part_of transitivity: If process Y exists in the GO biological process ontology and it is a *part_of* child of process X then any process that regulates process Y also regulates process X.

In the example above, **regulation of transcription** regulates **transcription** which is *part_of* **gene expression**. Therefore, **regulation of transcription** also regulates **gene expression**.

RNAseq interpretation – Gene Ontology

layers



RNAseq interpretation – Gene Ontology

GOID	EVIDENCE	ONTOLOGY	ENTREZID	SYMBOL	GENENAME
GO:0030154	IEA	BP	13642	Efnb2	ephrin B2
GO:0030154	IEA	BP	14175	Fgf4	fibroblast growth factor 4
GO:0030154	IEA	BP	14367	Fzd5	frizzled homolog 5 (Drosophila)
GO:0030154	IEA	BP	15482	Hspa1l	heat shock protein 1-like
GO:0030154	IEA	BP	16413	Itgb1bp1	integrin beta 1 binding protein 1
GO:0030154	IMP	BP	16600	Klf4	Kruppel-like factor 4 (gut)
GO:0030154	IMP	BP	16923	Sh2b3	SH2B adaptor protein 3
GO:0030154	IEA	BP	17242	Mdk	midkine
GO:0030154	IEA	BP	17450	Morc1	microrchidia 1

- The *Experimental Evidence codes* are:
  - Inferred from Experiment (EXP)
  - Inferred from Direct Assay (IDA)
  - Inferred from Physical Interaction (IPI)
  - Inferred from Mutant Phenotype (IMP)
  - Inferred from Genetic Interaction (IGI)
  - Inferred from Expression Pattern (IEP)

RNAseq interpretation – Gene Ontology

GOID	EVIDENCE	ONTOLOGY	ENTREZID	SYMBOL	GENENAME
GO:0030154	IEA	BP	13642	Efnb2	ephrin B2
GO:0030154	IEA	BP	14175	Fgf4	fibroblast growth factor 4
GO:0030154	IEA	BP	14367	Fzd5	frizzled homolog 5 (Drosophila)
GO:0030154	IEA	BP	15482	Hspa1l	heat shock protein 1-like
GO:0030154	IEA	BP	16413	Itgb1bp1	integrin beta 1 binding protein 1
GO:0030154	IMP	BP	16600	Klf4	Kruppel-like factor 4 (gut)
GO:0030154	IMP	BP	16923	Sh2b3	SH2B adaptor protein 3
GO:0030154	IEA	BP	17242	Mdk	midkine
GO:0030154	IEA	BP	17450	Morc1	microrchidia 1

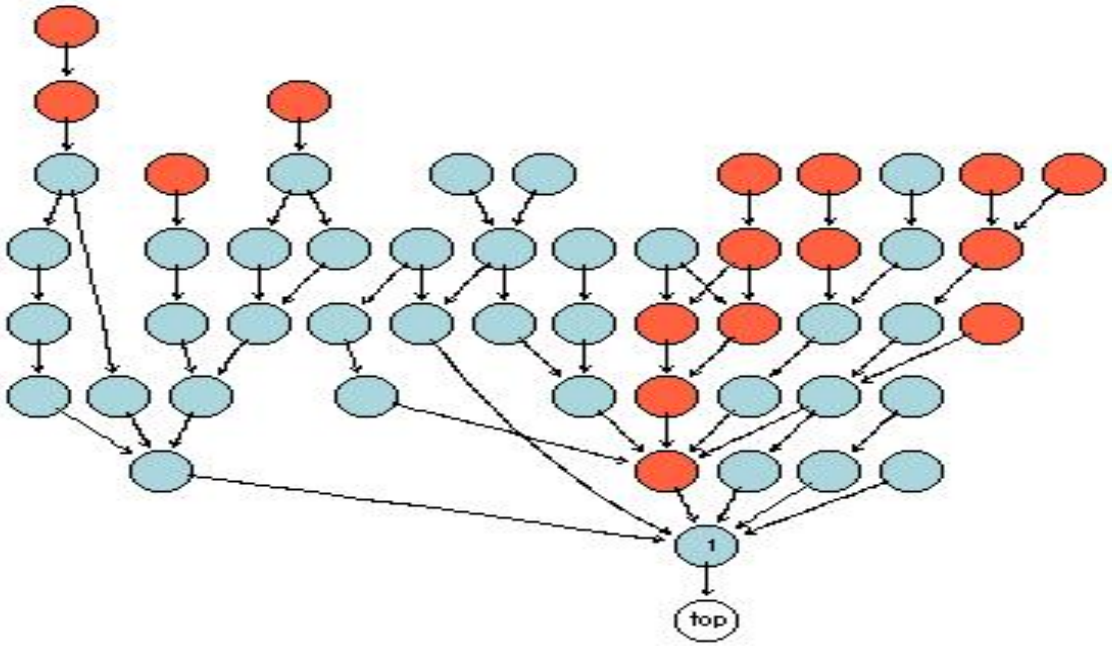
- The *Computational Analysis evidence codes* are:
  - Inferred from Sequence or structural Similarity (ISS)
  - Inferred from Sequence Orthology (ISO)
  - Inferred from Sequence Alignment (ISA)
  - Inferred from Sequence Model (ISM)
  - Inferred from Genomic Context (IGC)
  - Inferred from Biological aspect of Ancestor (IBA)
  - Inferred from Biological aspect of Descendant (IBD)
  - Inferred from Key Residues (IKR)
  - Inferred from Rapid Divergence(IRD)
  - Inferred from Reviewed Computational Analysis (RCA)

RNAseq interpretation – Gene Ontology

GOID	EVIDENCE	ONTOLOGY	ENTREZID	SYMBOL	GENENAME
GO:0030154	IEA	BP	13642	Efnb2	ephrin B2
GO:0030154	IEA	BP	14175	Fgf4	fibroblast growth factor 4
GO:0030154	IEA	BP	14367	Fzd5	frizzled homolog 5 (Drosophila)
GO:0030154	IEA	BP	15482	Hspa1l	heat shock protein 1-like
GO:0030154	IEA	BP	16413	Itgb1bp1	integrin beta 1 binding protein 1
GO:0030154	IMP	BP	16600	Klf4	Kruppel-like factor 4 (gut)
GO:0030154	IMP	BP	16923	Sh2b3	SH2B adaptor protein 3
GO:0030154	IEA	BP	17242	Mdk	midkine
GO:0030154	IEA	BP	17450	Morc1	microrchidia 1

- The *Author Statement evidence codes* used by GO are:
  - Traceable Author Statement (TAS)
  - Non-traceable Author Statement (NAS)
- The *Curatorial Statement codes* are:
  - Inferred by Curator (IC)
  - No biological Data available (ND) evidence code
- The *Automatically-Assigned evidence code* is:
  - Inferred from Electronic Annotation (IEA)

RNAseq interpretation – Gene Ontology



Top node

The induced GO graph colored according to unadjusted hypergeometric p-value≤0.01

GO can be used to link differentially expressed genes to specific functional classes.

Enrichment analysis

We consider a total population of genes, e.g. the genes expressed in a high-throughput experiment, and we are interested in the property of a **gene to belong to a specific GO category**. The aim is to establish whether the class of the DE genes presents an **enrichment and/or a depletion of the GO category of interest with respect to the total gene population.**

The **null hypothesis** that the property for a gene to belong to the GO category of interest and that to be DE are **independent**, or **equivalently that the DE genes are picked at random from the total gene population**

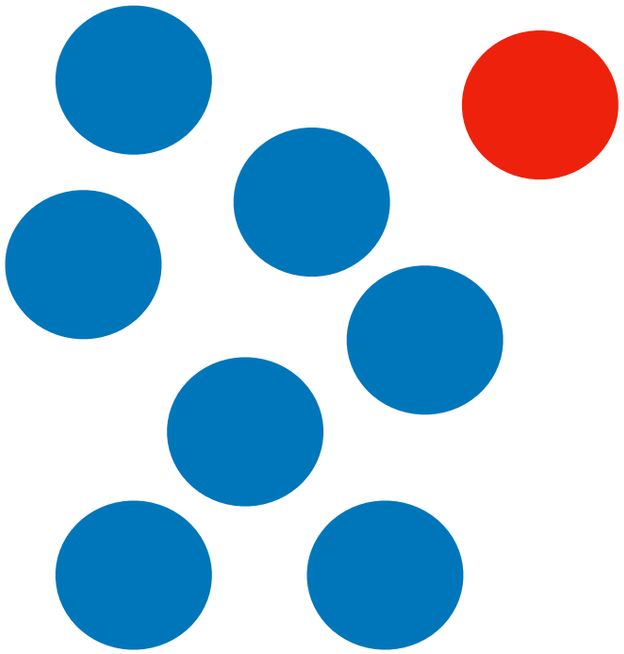
Hypergeometric distribution and Fisher's test

The hypergeometric distribution is a discrete probability distribution that describes the probability of *k* successes (random draws for which the object drawn has a specified feature) in *n* draws, **without replacement**, from a **finite population** of size *N* that **contains exactly *K* objects with that feature,** wherein each **draw is either a success or a failure.**

Fisher's exact test to determine if something is enriched or not.

Hypergeometric distribution and Fisher's test

Bag of balls



I extract 7 blue balls and 1 red

What does that say about the distributions of colours in the bag?

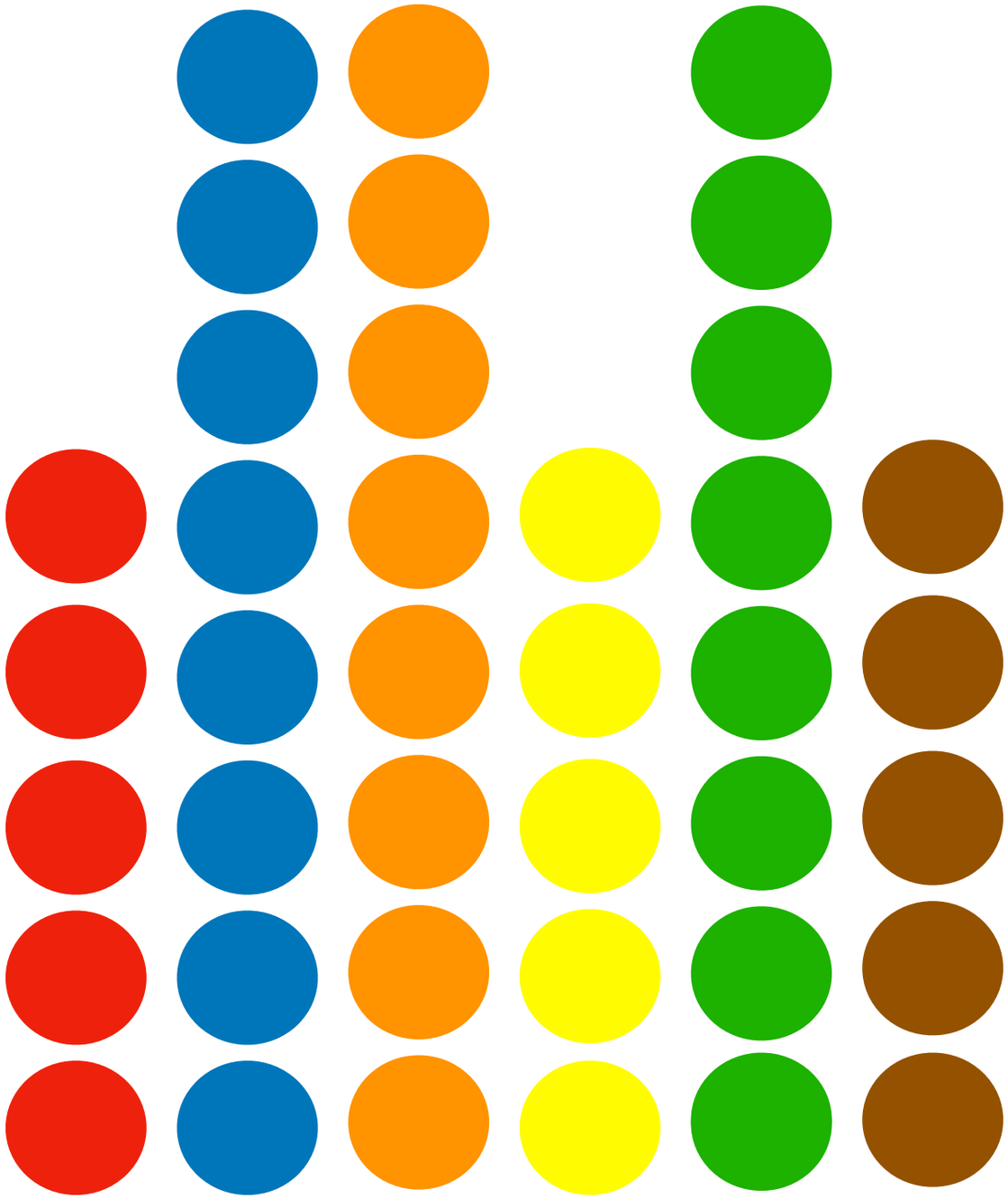
Do I have more blues than normal?

Can I calculate a p-value from this sample?

Hypergeometric distribution and Fisher's test

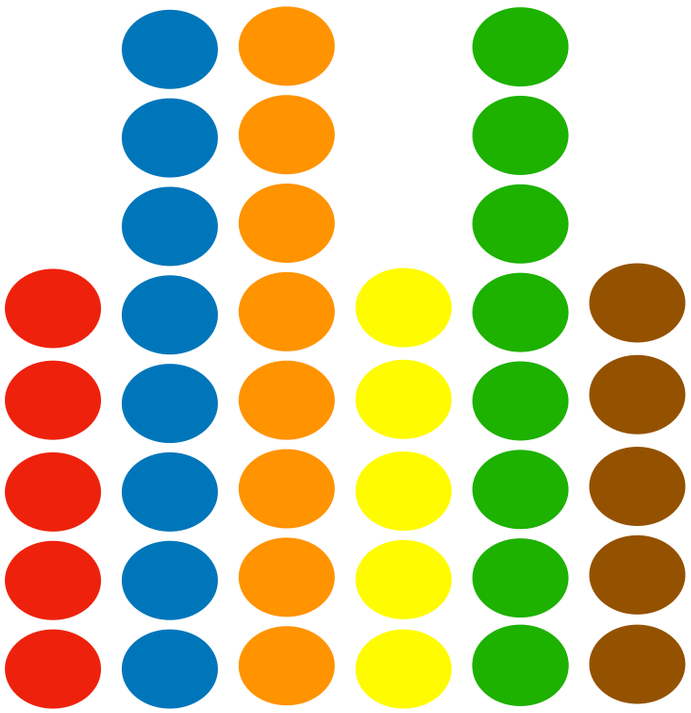
Bag of balls

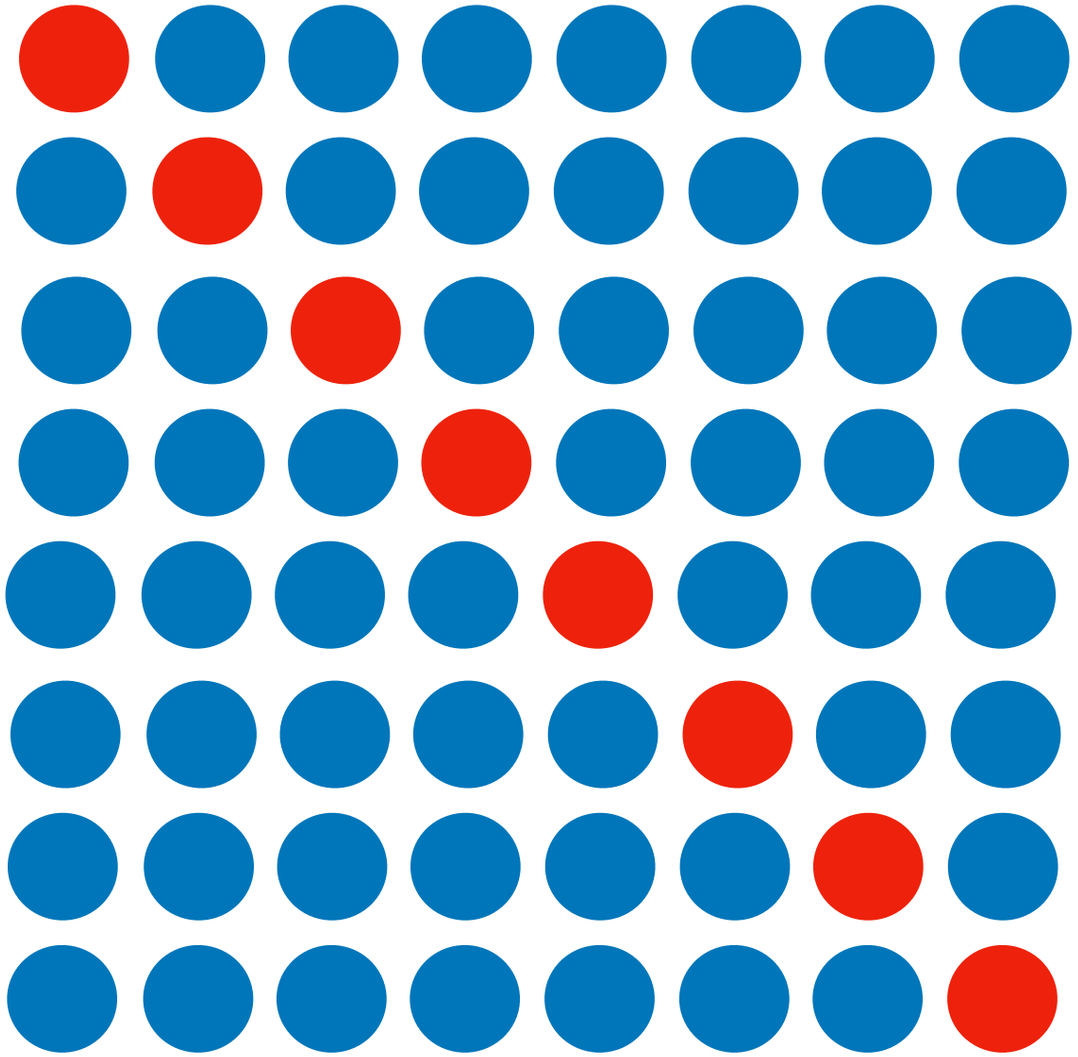
Red	=	13%
Yellow	=	14%
Orange	=	21%
Green	=	20%
Brown	=	12%
Blue	=	21%



Hypergeometric distribution and Fisher's test

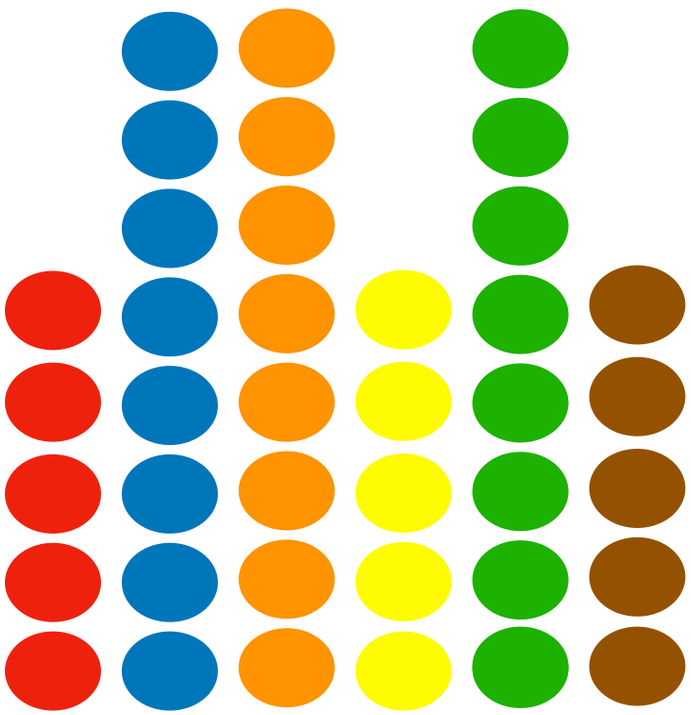
Determine if the set of balls of this sample is special or not?

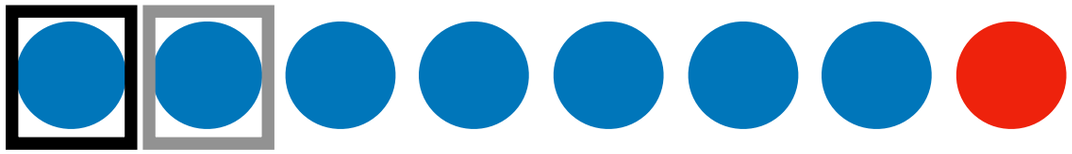




The order of how the balls are extracted is not important, then consider all possible ordering of the 7 blue and 1 red as legit

Hypergeometric distribution and Fisher's test





Let's start by calculating the probability of getting 7 blues balls followed by a single red

The probability that the first ball blue is 8/40 ,

Where:

8 because there are 8 blues

40 is the total number of balls

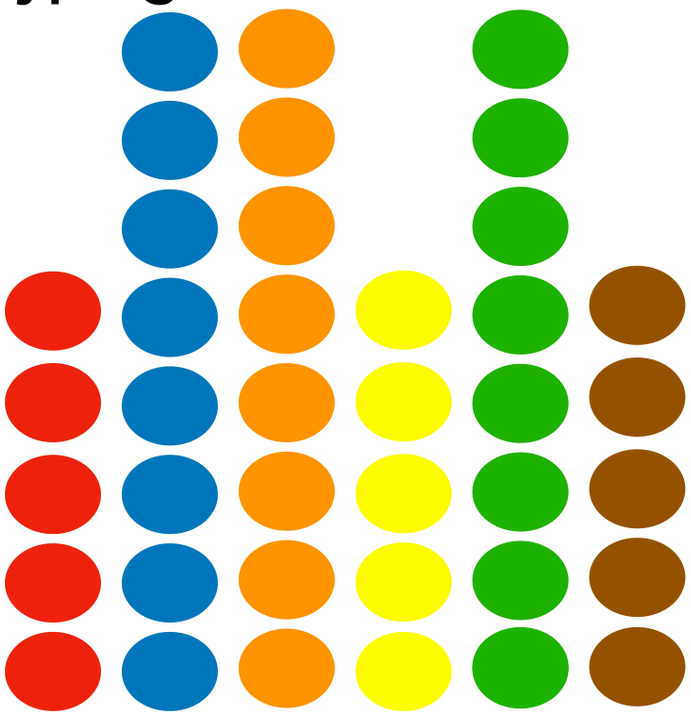
The probability that the second ball blue is 7/39 ,

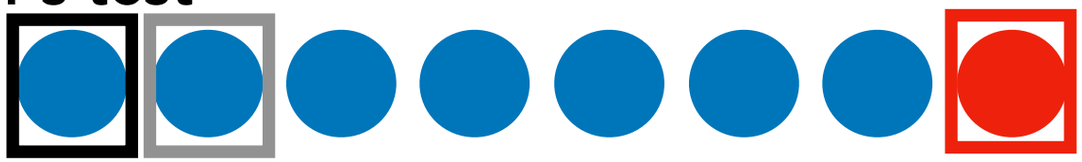
Where:

7 because there are 8 blues

39 is the total number of balls

Hypergeometric distribution and Fisher's test





Let's start by calculating the probability of getting 7 blues balls followed by a single red

The probability that the first ball blue is 8/40 ,

Where:

8 because there are 8 blues

40 is the total number of balls

The probability that the first ball blue is 7/39 ,

Where:

7 because there are 7 blues

39 is the total number of balls

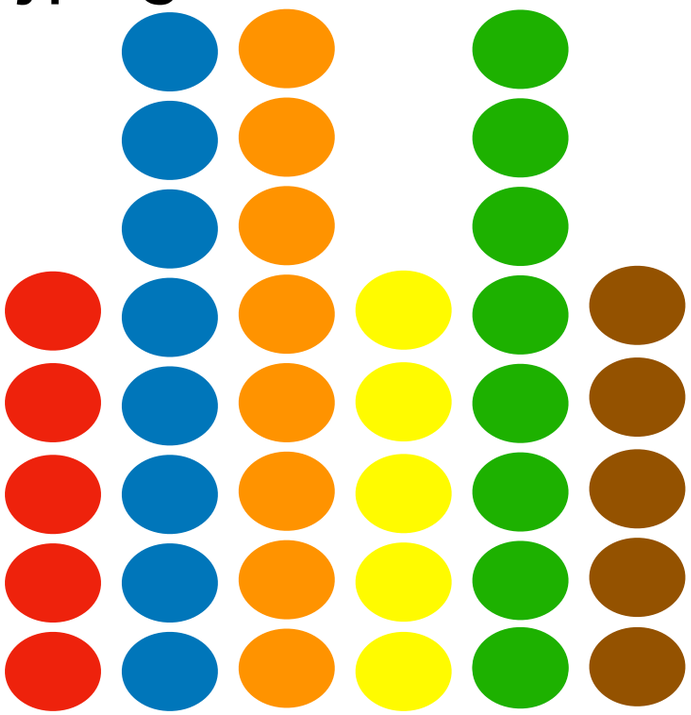
The probability that the first ball red is 5/33 ,

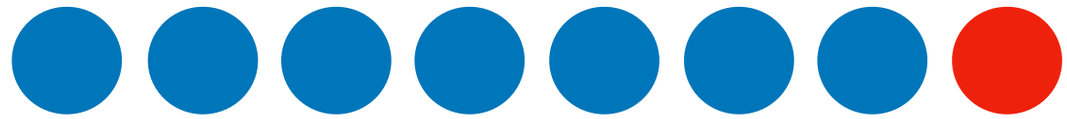
Where:

5 because there are 5 reds

33 is the total number of balls

Hypergeometric distribution and Fisher's test





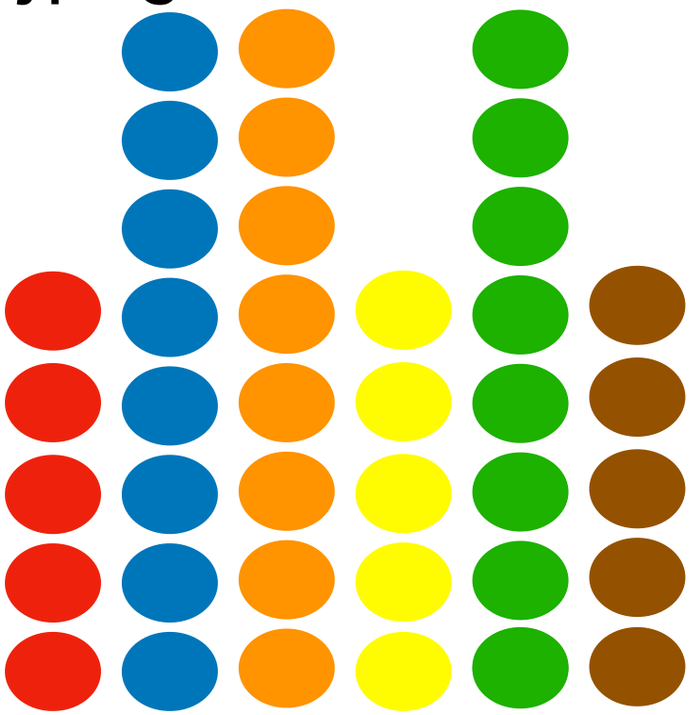
Let's start by calculating the probability of getting 7 blues balls followed by a single red

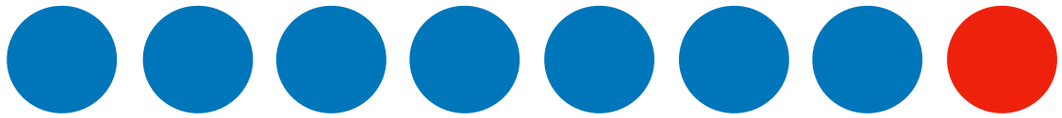
Multiply all those probabilities together to get the probability of getting 7 blues followed by one red is 0.000000065

The probability to obtain 7 blues and 1 red not depend by the order then, to calculate the probability of getting 7 blues and 1 red we need to consider all the probabilities of each possible ordering.

We repeat the computation of the probability considering any order and we obtain: 0.00000053

Hypergeometric distribution and Fisher's test

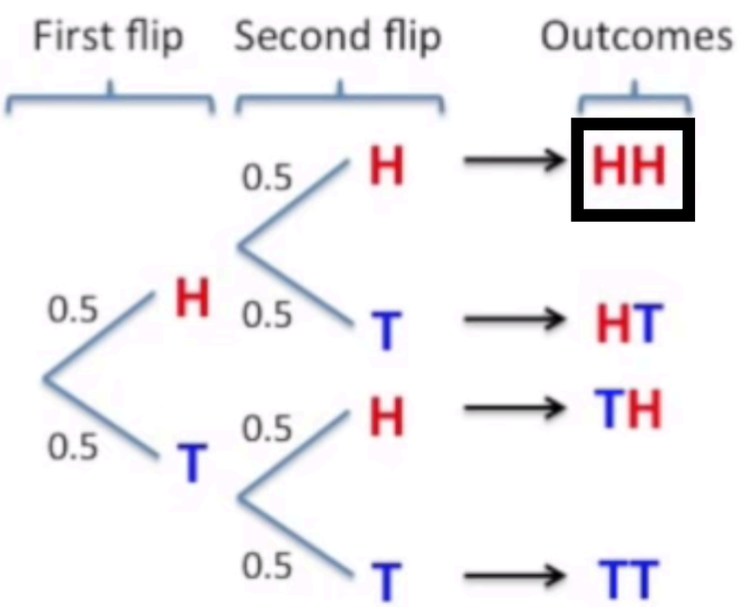




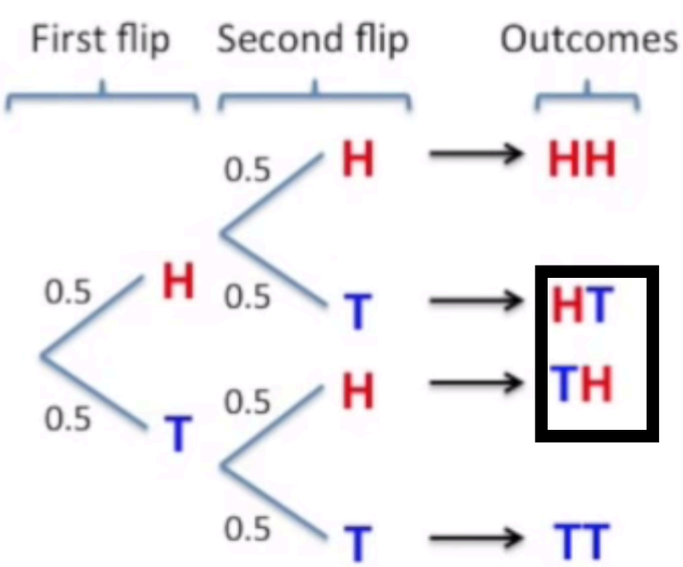
We repeat the computation of the probability considering any order and we obtain:
0.00000053

Compute the p-value

Probability versus p-value



$$\frac{\text{The number of times HH occurred}}{\text{The total \# of outcomes}} = 1/4$$



The probability of getting one H and one T (regardless of order) is...

$$\frac{\text{The number of times HT or TH occurred}}{\text{The total \# of outcomes}} = \frac{2}{4}$$

The order of the elements does not matter.

Probability versus p-value

A p-value is the probability that random chance generated the data, or something else that is equal or rarer.

We've already taken care of the first part...

$$\frac{HH}{HH, HT, TH, TT} = \frac{1}{4} = 0.25$$

+

$$\frac{TT}{HH, HT, TH, TT} = \frac{1}{4} = 0.25$$

+

Since nothing is rarer, this part is equal to zero.

The **probability** of getting **HH** is **0.25**

The **p-value** for getting **HH** is **0.5**

Probability versus p-value

Pr(4 **heads** and 1 **tails**) =

$$\frac{5}{32} = 0.15625$$

Outcomes

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

What's the p-value?

Pr(4 **heads** and 1 **tails**)

+

Pr(1 **heads** and 4 **tails**)

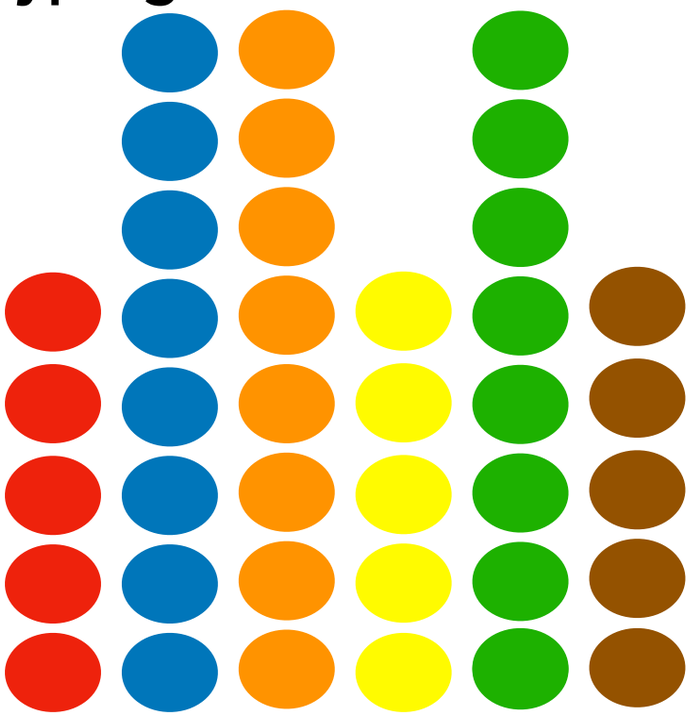
+

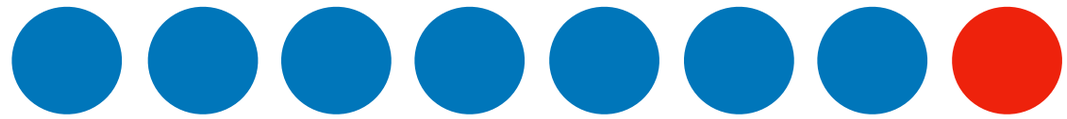
Pr(5 **heads**) + Pr(5 **tails**)

= 0.375

A p-value is the probability that random chance generated the data, or something else that is equal or rarer.

Hypergeometric distribution and Fisher's test





We repeat the computation of the probability considering any order and we obtain:
0.00000053

The p-value is the sum of the probabilities of all things equally rare or rarer. Then compete the probability for 7 blues and 1 orange, 8 blues (as the rarer) etc.

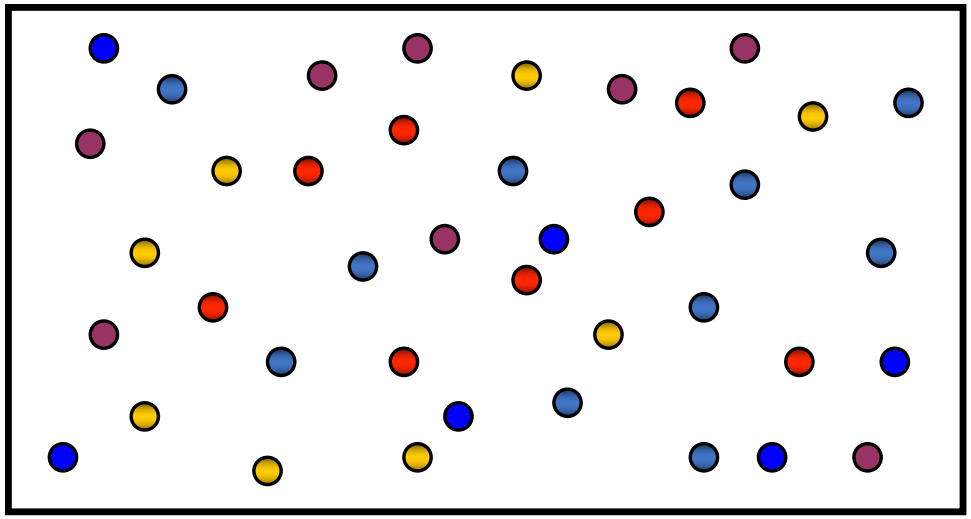
Finally the p-values is 0.01.

This is call Fisher's exact test.

Enrichment for other things, "does this list of genes have more involved in metabolism than normal" can be answered following the same way.

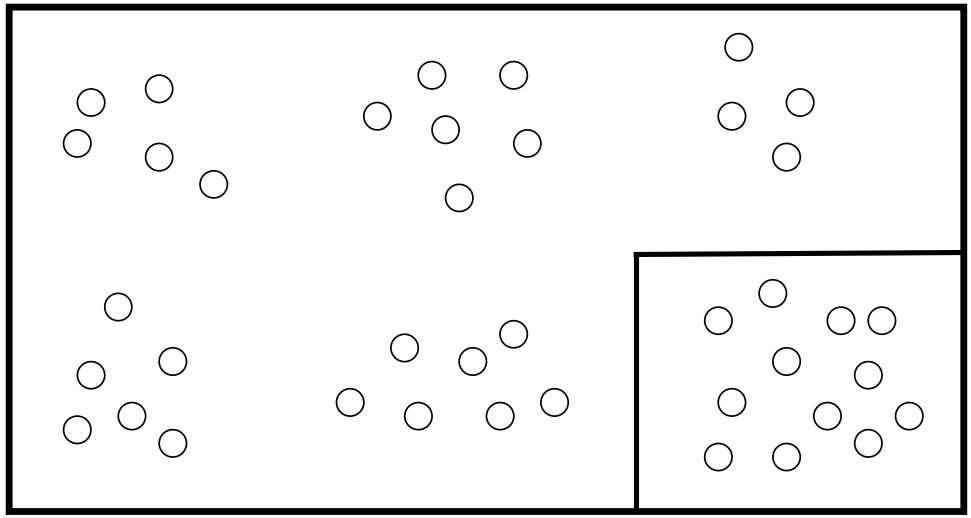
RNAseq interpretation – Gene Ontology, enrichment

Consider a population of genes representing a diverse set of GO terms shown below as different colors.



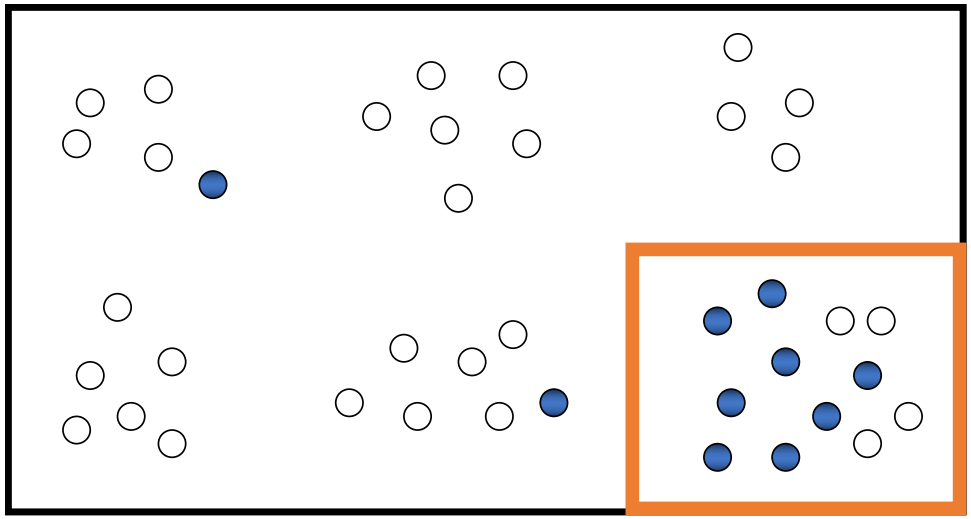
RNAseq interpretation – Gene Ontology, enrichment

Many methods can be used to identify a set of differentially expressed genes



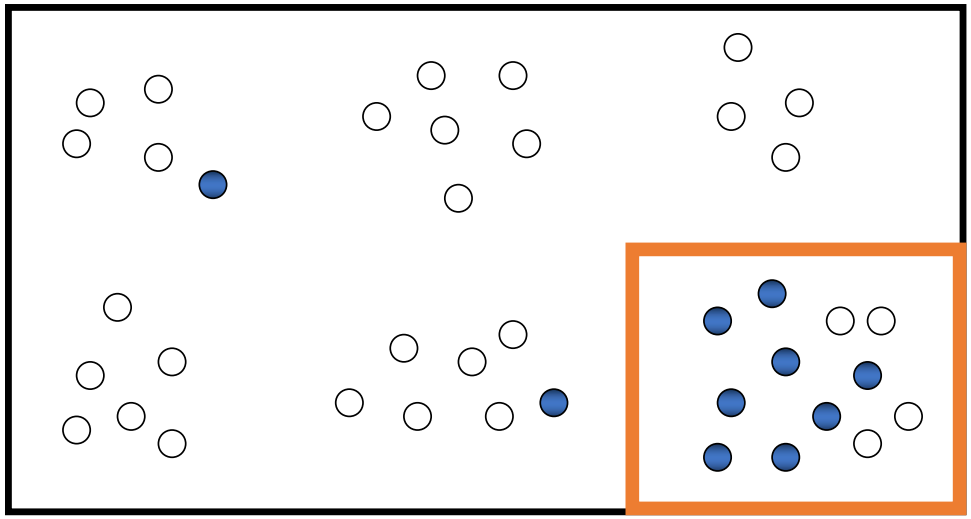
RNAseq interpretation – Gene Ontology, enrichment

What are the some of the predominant GO terms represented in the set of differentially expressed genes and how should significance be assigned to a discovered GO term?



RNAseq interpretation – Gene Ontology, enrichment

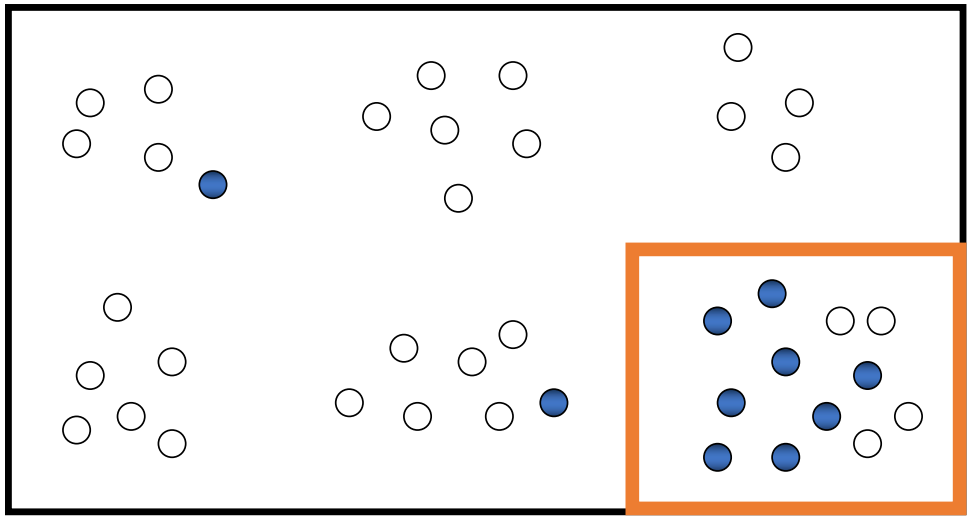
A 2x2 contingency matrix is typically used to capture the relationships between differentially expressed membership and membership to a GO term.



RNAseq interpretation – Gene Ontology, enrichment

Contingency Matrix

GO term \ Subset	in	out
in	8	2
out	4	26



RNAseq interpretation – Gene Ontology, enrichment

Hypergeometric Distribution

a	b	a+b
c	d	c+d
a+c	b+d	

The probability of any **particular** matrix occurring by random selection, given no association between the two variables, is given by the **hypergeometric rule**.

$$\frac{\dfrac{(a+c)!}{a!c!} \times \dfrac{(b+d)!}{b!d!}}{\dfrac{n!}{(a+b)!(c+d)!}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

RNAseq interpretation – Gene Ontology, enrichment

Assigning Significance to the Findings

The **HyperGeometric Test** permits us to determine if there are non-random associations between the two variables, differential expression membership and membership to a particular Gene Ontology term.

		Subset	
		in	out
GO term	in	8	2
	out	4	26

(2x2 contingency matrix)

p ≈ .0002

RNAseq interpretation – Gene Ontology, criticalities

The GO is a consistent descriptions of genes in different data sources. The **annotations** can use also to measure **the functional similarities** (SS) of genes.

Different types of **SS have been proposed: based only on GO structure, or based on the information content of a term derived form the corpus statistics**.
SS's measure is based on both

(i) the location in the GO graph
(ii) the GO term's semantics that are inherited from all its ancestor terms.

Based on human perspectives, if two terms sharing the same parent are near the root of the ontology (terms are more general), they should have larger semantic difference than two terms having the same parent and being far away from the root of the ontology because the later are more specific terms.

RNAseq interpretation – Gene Ontology, criticalities

Every GO term must obey **the true path rule**: if the child term describes the gene product, then all its parent terms must also apply to that gene product. Let consider how chitin metabolism is represented in the process ontology. Chitin metabolism is a part of cuticle synthesis in the fly and is also part of cell wall organization in plants. This was once represented in the process ontology as follows:

```
cuticle synthesis
[i] chitin metabolism
cell wall biosynthesis
[i] chitin metabolism
---[i] chitin biosynthesis
---[i] chitin catabolism
```

The problem with this organization becomes apparent when one tries to annotate a specific gene product from one species. A fly chitin synthase could be annotated to chitin biosynthesis, and appear in a query for genes annotated to cell wall biosynthesis (and its children), which makes no sense because flies don't have cell walls.

RNAseq interpretation – Gene Ontology, criticalities

This is the revised ontology structure which ensures that the true path rule is not broken:

```
chitin metabolism
[i] chitin biosynthesis
[i] chitin catabolism
[i] cuticle chitin metabolism
---[i] cuticle chitin biosynthesis
---[i] cuticle chitin catabolism
[i] cell wall chitin metabolism
---[i] cell wall chitin biosynthesis
---[i] cell wall chitin catabolism
```

RNAseq interpretation – Gene Ontology, criticalities

GO is marked by flaws due to a failure to address basic ontological principles.

- the existing annotation databases are incomplete;

- the quality of an association among GO terms and genes depends upon the source of the annotation, some information are imprecise or incorrect;

- the GO is an ongoing project in which new GO terms are added continuously and this can lead to a re-classification of all tagged gene products;

- genes involved in several biological process, all the biological process is weight equally, it is not possible single out the more relevant one.

RNAseq interpretation – Gene Ontology, criticalities

Optimize the organization of the GO to optimize the distribution of the information. Particularly used by enrichment web tools.

The quantification of information contained in the terms ontology is computed considering the amount of annotations available for a given term. With this measure Alterovitz et al demostrate some structural inefficiency:

1.the variability of the information content among the terms within a given ontology level. For example, pilus retraction is at the same level of cell cycle and cell development.

2.in some area of GO the mean information content decrease from one level to the next creating the bottle-neck → problem in the use of enrichment tools.

3.the closer a topological structure is to uniform, the greater is the information that experiments can derive from it.

RNAseq interpretation – Gene set enrichment

- Interpreting the results to gain insights into biological mechanisms remains a major challenge

- For a typical two group comparison, e.g., tumor vs. normal, treated vs. control, **a standard approach has been to produce a list of differentially expressed genes (DEGs)**

- One also might obtain a list of "Distinguished Genes" from examining correlation of gene expression with a pertinent clinical variable, or from differences in methylation

RNAseq interpretation – Gene set enrichment

Criteria for Differential Expression of a Gene

- **Statistically significant differential expression**
  - by t-test, multi-way ANOVA, etc.
  - P-value cut-off: require, e.g., $p \leq 0.01$, but see FDR (which will impose more stringent requirement for p-values)

- **Satisfactory false discovery rate (FDR)**
  - What fraction of the DEG list is false positives?
  - Benjamini-Hochberg procedure for estimating the FDR is a common choice (e.g., require FDR ≤ 0.1 or 0.2).

- **Sufficient level of fold change (FC)**
  - require $|FC| \geq 1.5$ or 2; common convention: groups A, B, gene g with average expression levels μA, μB; $FC \equiv \mu A / \mu B$

RNAseq interpretation – Gene set enrichment

Challenges in Interpreting Gene Microarray/Seq Data

- Even with DEG lists of upregulated and of down-regulated genes, **still need to accurately extract valid biological inferences**. Cutoff for inclusion in DEG lists is somewhat arbitrary.
- May obtain a long list of statistically significant genes without any obvious unifying biological theme
- May have few individual genes meeting the threshold for statistical significance

RNAseq interpretation – Gene set enrichment

- These methods formulate a statistic for the ensemble of genes in each gene set using a selected metric for each gene. Increases statistical power.
  - T-score for group A vs. group B comparison
  - Fold Change for group A vs. group B
  - Pearson correlation of gene expression with a pertinent clinical variable

- The expression data for all the genes in the dataset is used.Can be applied to many types of gene sets
  - pathways from BioCarta & KEGG
  - genes changed in response to some disease or experimental condition
  - GO categories
  - genes co-located in cytobands
  - genes having common transcription factor motifs

- But note: results depend on the collection of gene sets examined, and still must address multiple testing error control (though much less severe than for all probes on a large array). Run different types of gene set collections separately.

Overview of GSEA

- Take gene expression data from two different conditions and rank according to the differential expression across the conditions
- Take a test set of genes and determine whether they are collectively differentially expressed
- Randomly swap the class labels of the data and repeat the test many times as a gauge of significance

RNAseq interpretation – Gene set enrichment

GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

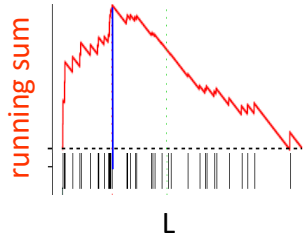
text and figure from the Broad Institute web pages for GSEA : http://www.broad.mit.edu/gsea/index.html
the current version of the figure at the Broad site is slightly different from the one above

RNAseq interpretation – Gene set enrichment

Get ranked list L of all the genes on the chip based on a chosen measure, e.g., FC or Tscore, of the difference of their expression levels between the phenotypes A & B under study, e.g., tumor vs. normal

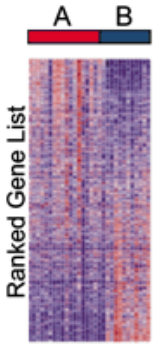
For each gene set S: find the location of each gene s in S within L

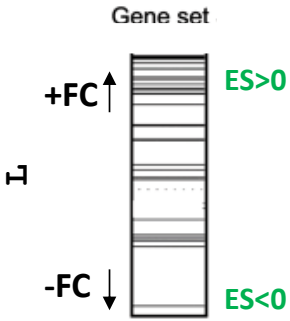
Generate enrichment score ES for S based on running-sum statistic: "reward" presence of s toward top or bottom of L



Analyze significance of this Kolmogorov-Smirnov type statistic by permutation testing

Multiple hypothesis testing (MHT) error control for multiple S's using the estimated false discovery rate (FDR)





bands are locations in L of genes from S

RNAseq interpretation **Enrichment Score (ES) Calculation**

Start with ranked list (L) of genes that are in (Hit) or not in (Miss) a gene set (S), using fold change (FC) as example metric

Ranked List (L)	FC		Contribution to running sum for ES	Hits $+\lvert FC\rvert / \Sigma$	Misses $-1/(N-N_H)$	Running sum for ES
	15	Hit	+0.15	+0.15		0.15
	12	Hit	+0.12	+0.12		0.27
	10	Miss	-0.001		-0.001	0.269
	9	Hit	+0.09	+0.09		0.359
	8	Hit	+0.08	+0.08		0.439
	6	Miss	-0.001		-0.001	0.438
…	…	…				…

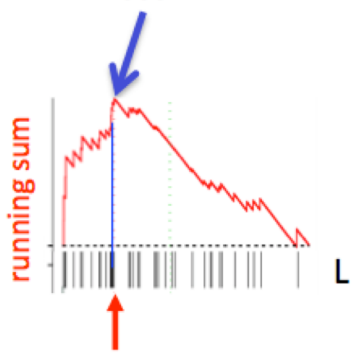
Hits: Genes $\in$ S $\quad +\lvert FC\rvert / \Sigma$

Misses: Genes $\notin$ S $\quad -1/(N-N_H)$

Σ = sum of fold changes for genes in gene set (S) (e.g., 100)

N = no. of genes in the array (e.g., 1020)

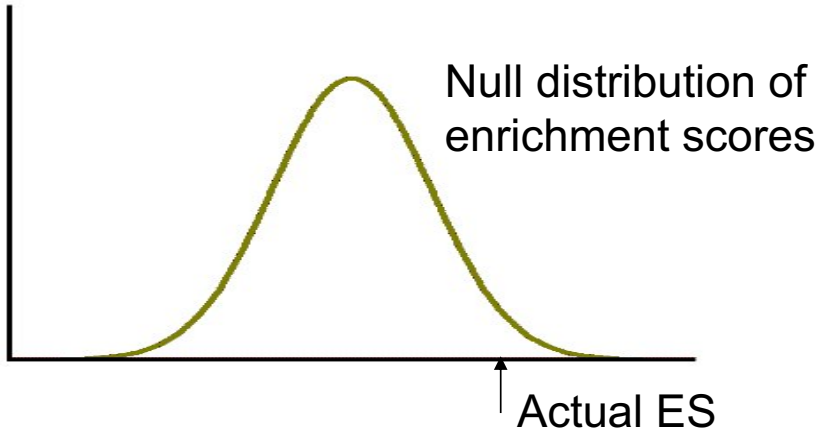
N_H = no. of genes in the gene set (S) (e.g., 20)



ES(S) ≡ value of maximum deviation from 0 of the running sum

RNAseq interpretation – Gene set enrichment

- Randomise data (groups), rank genes again and repeat test 1000 times
- Null distribution of 1000 ES for geneset



- FDR q-value computed – corrected for gene set size and testing multiple gene sets

RNAseq interpretation – Gene set enrichment

Testing the Significance of ES using Sample Label Permutations:

gene expression matrix, sample labels indicate phenotype group

gene \ sample	T1	T2	T3	T4	N1	N2	N3	N4
CASP4	7.82	7.87	8.15	7.81	7.96	7.92	7.90	7.96
BAX	8.01	7.85	7.82	7.95	8.05	7.91	7.78	7.96
CASP8	7.73	7.82	7.92	8.13	8.18	8.01	7.90	7.86
CD40	8.12	8.15	8.32	8.21	8.06	8.02	8.00	8.08
BIRC3	7.87	8.01	7.99	7.84	7.99	7.89	8.01	7.96
GADD45A	7.84	7.77	7.99	7.94	7.93	7.99	7.75	7.69
BIRC2	8.07	8.01	7.88	8.01	7.94	7.86	8.06	7.92
ATM	9.40	9.54	9.32	9.60	9.11	9.45	9.42	9.34
...	...	...	...	...	...	...	...	...

→ compute the differential expression value for each gene (DE(g)), and then the ES(S) values for all the gene sets

do ≈ 1000 sample label permutations π * - for each permutation π_i randomly shuffle the labels of which sample is in which group while **leaving the rest of the expression matrix fixed, and recalculate {DE(g)} and then the enrichment score for each S**

permutation number

1	T4	N3	N4	T3	T1	T2	N1	N2
2	T3	N2	T1	N3	T4	N4	T2	N1
3	N4	T4	N1	N3	T3	T2	N2	T1
4	N2	T4	N3	T1	T2	N1	T3	N4
...	...	...	...	...	...	...	...	...

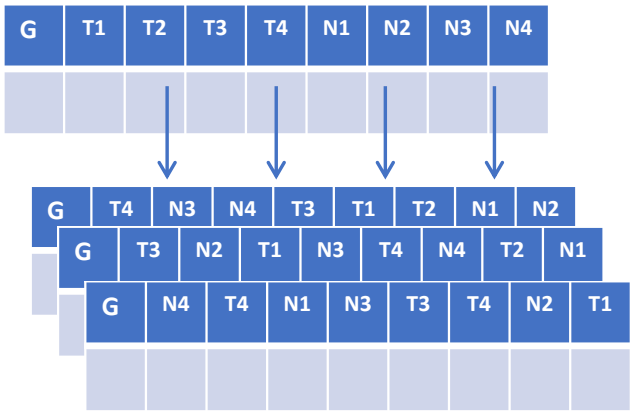
→ $\{ES(S,\pi_1)\}$
→ $\{ES(S,\pi_2)\}$
→ $\{ES(S,\pi_3)\}$
→ $\{ES(S,\pi_4)\}$

*actually want at least 7 samples in each group for sample label permutation, else do gene permutation

RNAseq interpretation – Gene set enrichment

Testing the Significance of ES

Significance of the observed ES(S) is compared with the set of empirical null distribution scores ES(S,π) computed with the *randomly assigned phenotypes or random gene sets*.



:

1000 x

ES(S)

ES(S,π_1)

ES(S,π_2)

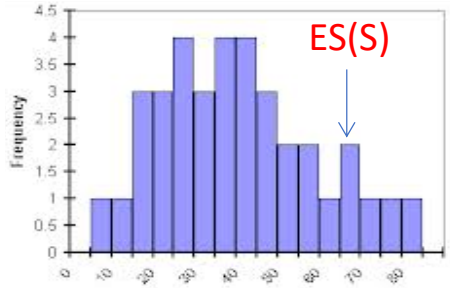
ES(S,π_3)

:

ES(S,π_{1000})

ES_{NULL}(S):null distribution for ES(S)

Histogram of 1000 ES(S,π) Scores



The empirical, nominal *p-value* for each ES(S) is then calculated relative to the null distribution for ES(S):

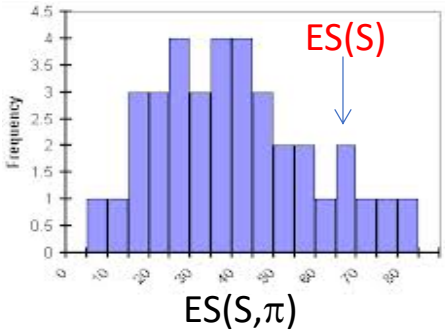
***p* = fraction of ES(S,π) values ≥ ES(S)**

RNAseq interpretation – Gene set enrichment

How normalized enrichment scores (NES) are calculated from ES

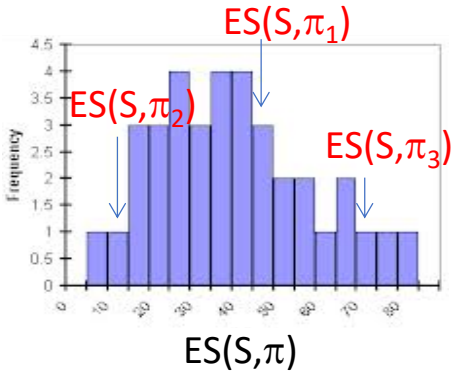
Using the NES helps normalize out effect of different gene set sizes

Histogram of the ES(S,π) values for a given S from the permutations



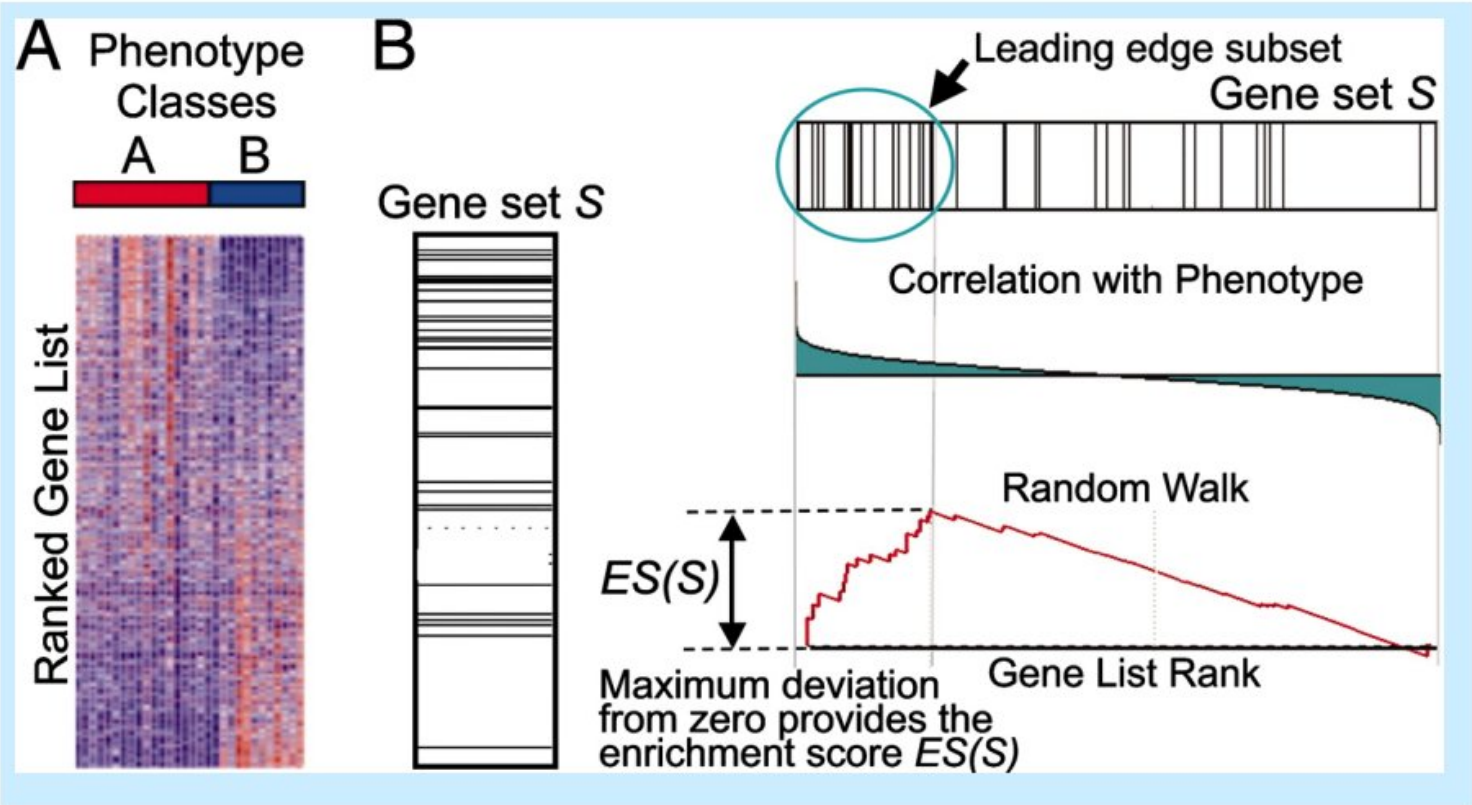
$$\mathbf{NES(S)} \equiv \frac{\textbf{original ES(S)}}{\textbf{mean}_{\pi}\textbf{\{ES(S, }\pi\textbf{) values ES(S)\}}}$$

For each permutation π and gene set S, compute NES(S, π) to use in computing the FDR:



$$\mathbf{NES(S, \pi_k)} \equiv \frac{\textbf{ES(S, }\pi_k\textbf{)}}{\textbf{mean}_{\pi}\textbf{\{ES(S, }\pi\textbf{) values ES(S, }\pi_k\textbf{)\}}}$$

RNAseq interpretation – Gene set enrichment

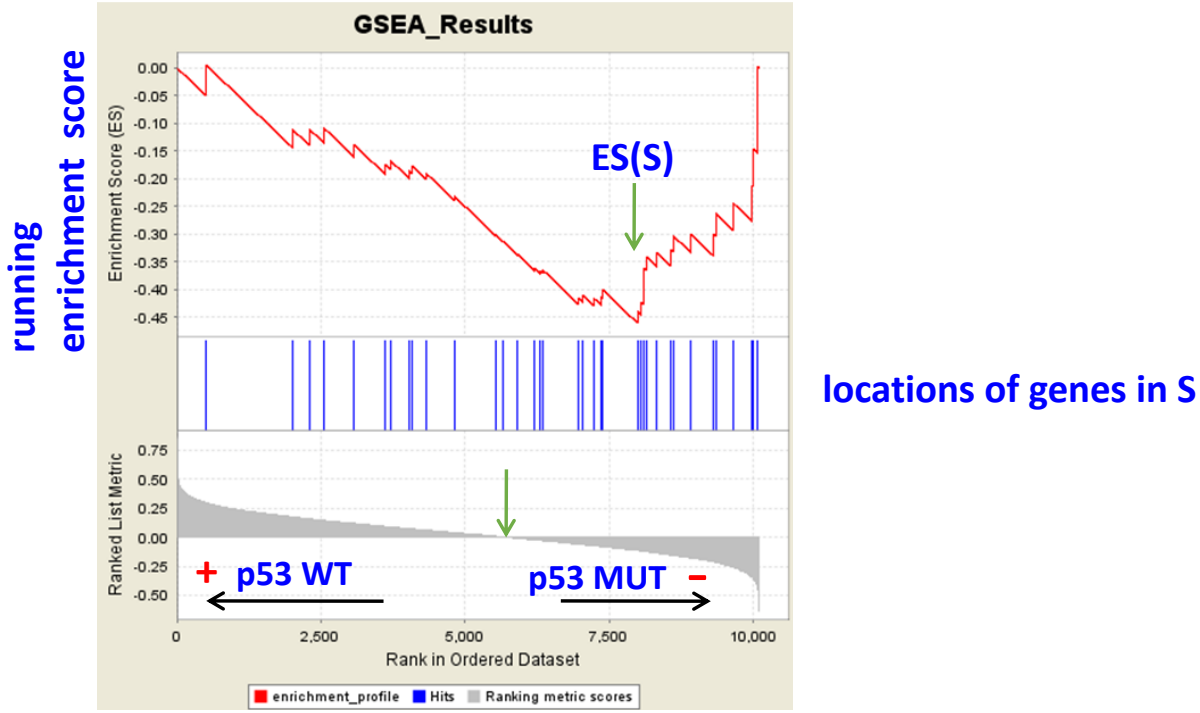


Genes in expression matrix are sorted based on correlation to phenotype classes (red and blue at the top of D, panel A). The positions of genes in S are noted with black bars to the right of D. ES(S) is calculated based on both the correlations and the positions in L.

RNAseq interpretation – Gene set enrichment

The running enrichment score for a negative ES gene set

from the P53 GSEA example data set



Zero crossing of ranking metric values

running enrichment score figure copied from
http://www.broadinstitute.org/gsea/datasets.jsp
p53 dataset (gene set is BRCA_UP)

RNAseq interpretation – Gene set enrichment

GSEA returns two lists of gene sets: {S with NES > 0} and {S with NES < 0} (sorted by NES value)

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val
1	EXTRACELLULAR_SPACE	Details ...	229	0.58	2.07	0.000	0.000
2	PROTEASE_INHIBITOR_ACTIVITY	Details ...	40	0.72	2.00	0.000	0.010
3	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	Details ...	67	0.66	1.94	0.000	0.037
4	REACTOME_SPHINGOLIPID_METABOLISM	Details ...	62	0.64	1.94	0.000	0.028
5	KEGG_LYSOSOME	Details ...	117	0.60	1.94	0.000	0.022

NES > 0, descending order

		GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val
1	REACTOME_MITOTIC_M_M_G1_PHASES	Details ...	164	-0.77	-2.67	0.000	0.000
2	REACTOME_DNA_REPLICATION	Details ...	184	-0.77	-2.66	0.000	0.000
3	REACTOME_M_G1_TRANSITION	Details ...	77	-0.80	-2.57	0.000	0.000
4	REACTOME_G1_S_TRANSITION	Details ...	105	-0.77	-2.55	0.000	0.000

NES < 0, ascending order

Conclusions of GSEA

- GSEA is a statistical test which can identify sets of genes, belonging to a particular biological category, which play an important role in distinguishing between two classes of gene expression data.
- The test is particularly sensitive as small changes which are coordinated across the set can be detected.
- The test helps reveal the biological mechanisms responsible for the difference between the two classes because the test set has an *a priori* biological theme.