




# Canadian Bioinformatics Workshops


[www.bioinformatics.ca](http://www.bioinformatics.ca)

Creative Commons


This page is available in the following languages:  
 Afrikaans (Suidaposa) Català Dansk Deutsch English (CA) English (GB) English (US) Esperanto  
 Español Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)  
 Eesti keel Suomi Suomi (Finland) Français (CA) Galego (Irish) Hrvatski Magyar Italiano 日本語 한국어 Maori Maori (New Zealand)  
 Nederlands Norsk Sesotho sa Leboa (South) Português português (Brasil) português (Portugal) português (Timor) português (Timor-Leste)  
 Slovenščina Slovenian Slovenian (Slovenia) Slovenian (Slovenia) Slovenian (Slovenia) Slovenian (Slovenia) Slovenian (Slovenia)  
 中文 普通话 (台湾) isiZulu



**You are free:**

-  **to Share** — to copy, distribute and transmit the work
-  **to Remix** — to adapt the work

**Under the following conditions:**

-  **Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
-  **Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

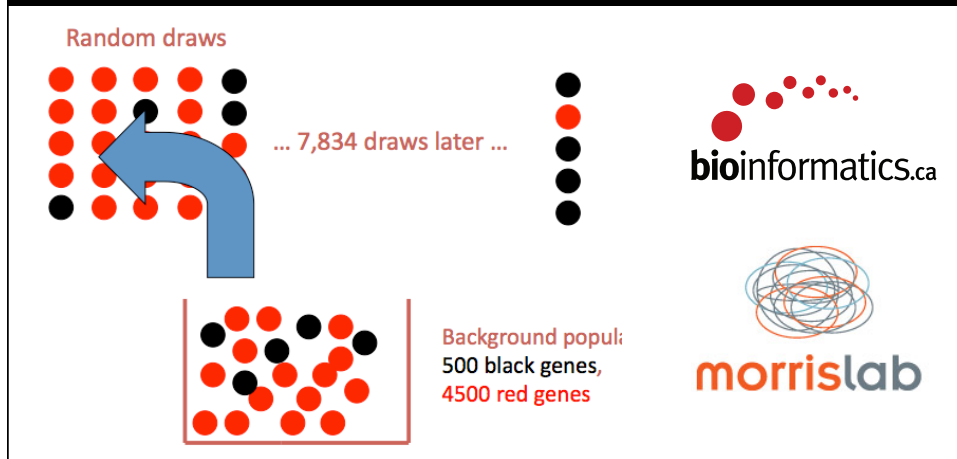
Your fair dealing and other rights are in no way affected by the above.  
 This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
 English French

[Learn how to distribute your work using this licence](#)

## Module 2

### Finding over-represented pathways in gene lists

Quaid Morris  
Pathway and network analysis of -omic data  
June 13-14, 2016



## Learning Objectives of Module 2

- **Be able** to select the appropriate enrichment test for your data.
- **Be able** to determine the appropriate background gene list when running Fisher's Exact Test (aka Hypergeometric test).
- **Be able** to compute a minimum hypergeometric test on a ranked list
- **Be able** to determine when you need a multiple test correction.
- **Be able** to select whether to use a Bonferroni corrected P-value or a false discovery rate.
- **Be able** to explain, in plain language, how you calculate each correction.

## Outline

- Introduction to enrichment analysis
- Hypergeometric Test, aka Fisher's Exact Test
- GSEA and minimum hypergeometric test for ranked lists.
- Multiple test corrections:
  - Bonferroni correction
  - False Discovery Rate computation using Benjamini-Hochberg procedure

Module 2

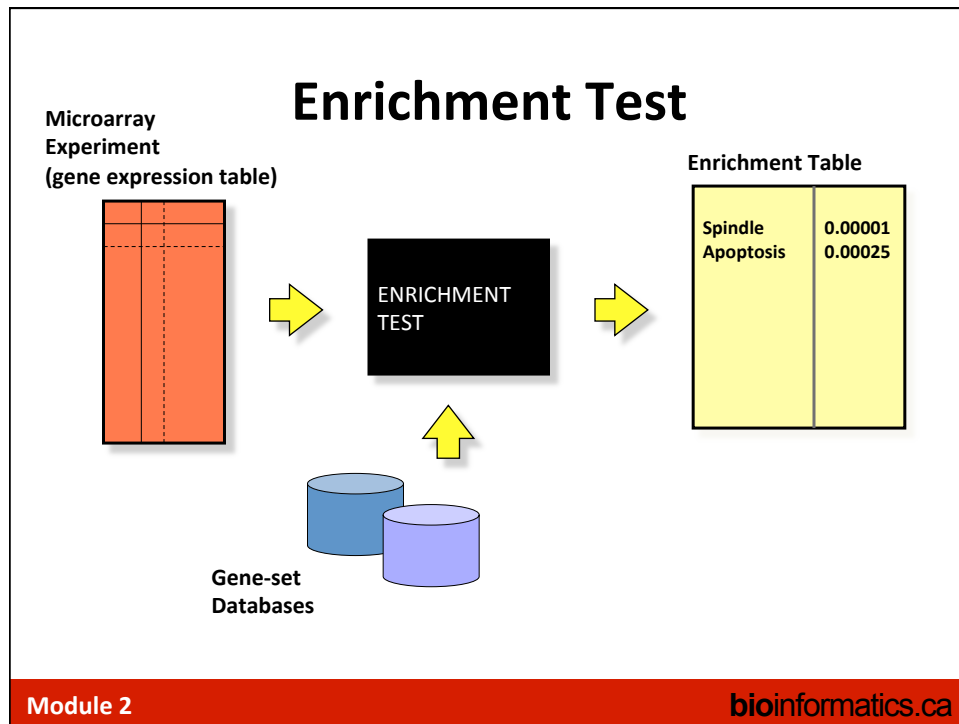
bioinformatics.ca

## Types of enrichment analysis

- **Gene list** (e.g. expression change > 2-fold)
  - Answers the question: **Are any gene sets surprisingly enriched (or depleted) in my gene list?**
  - Statistical test: Fisher's Exact Test (aka Hypergeometric test)
- **Ranked list** (e.g. by differential expression)
  - Answers the question: **Are any gene set ranked surprisingly high or low in my ranked list of genes?**
  - Statistical test: minimum hypergeometric test, GSEA (+ others we won't discuss)

Module 2

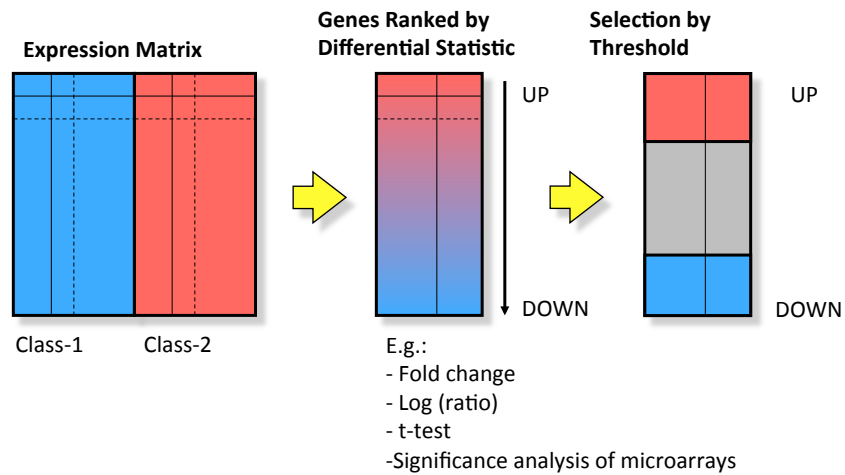
bioinformatics.ca



## Gene list enrichment analysis

- Given:
  1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
  2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- Question: *Are any of the gene annotations surprisingly enriched in the gene list?*
- Details:
  - Where do the gene lists come from?
  - How to assess “surprisingly” (statistics)
  - How to correct for repeating the tests

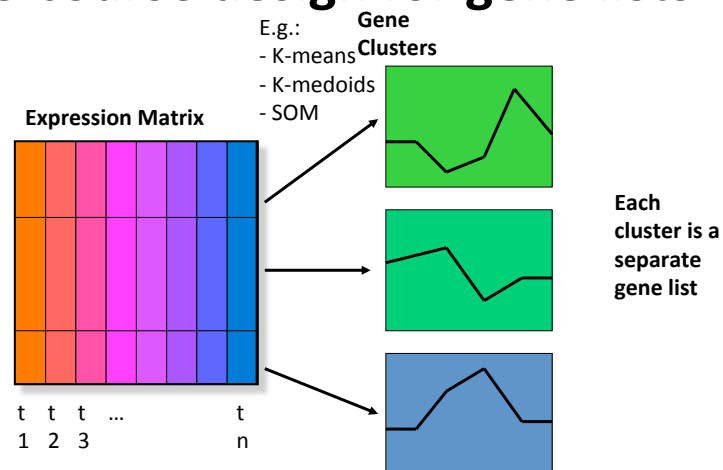
## Two-class design for gene lists



Module 2

bioinformatics.ca

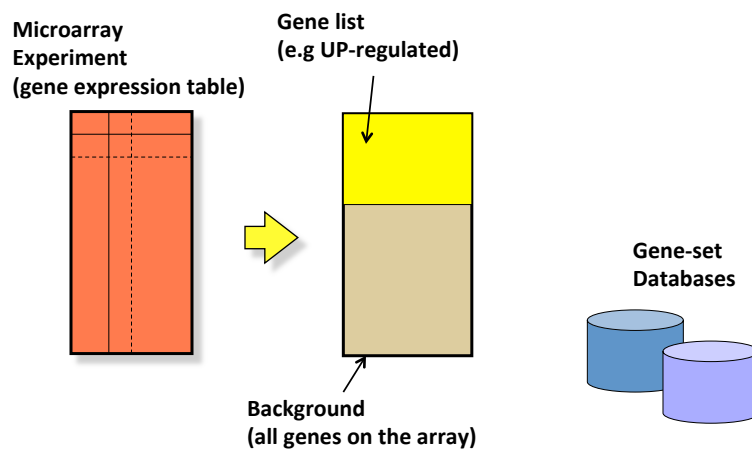
## Time-course design for gene lists



Module 2

bioinformatics.ca

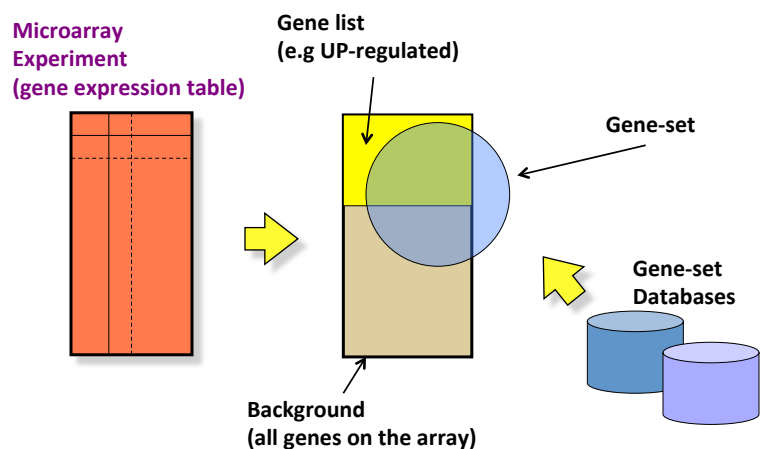
## Example gene list enrichment test



Module 2

bioinformatics.ca

## Example gene list enrichment test

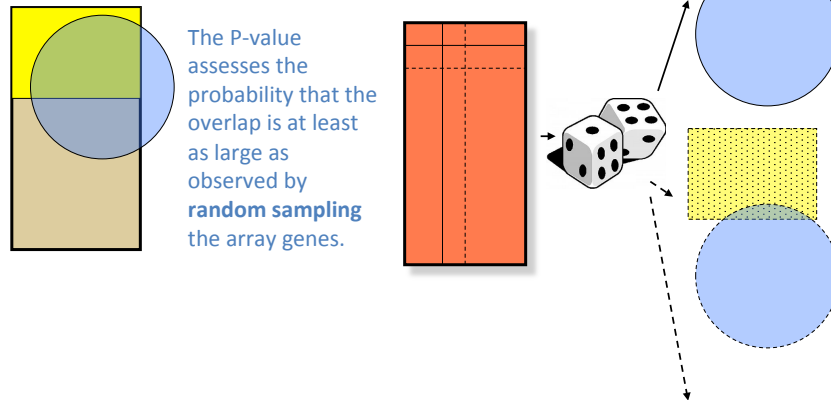


Module 2

bioinformatics.ca

## Enrichment Test

The output of an enrichment test is a *P-value*



Module 2

bioinformatics.ca

## Recipe for gene list enrichment test

- **Step 1:** Define your gene list and your background list,
- **Step 2:** Select your gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

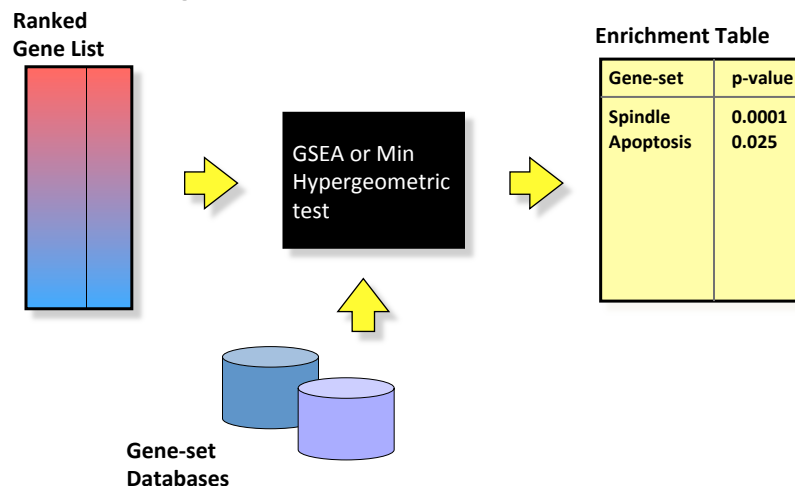
Module 2

bioinformatics.ca

## Why test enrichment in ranked lists?

- Possible problems with gene list test
  - No “natural” value for the threshold
  - Different results at different threshold settings
  - Possible loss of statistical power due to thresholding
    - No resolution between significant signals with different strengths
    - Weak signals neglected

## Example ranked list enrichment test





## Recipe for **ranked** list enrichment test

- **Step 1:** Rank your genes,
- **Step 2:** Select your gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Module 2

bioinformatics.ca

## Outline of theory component

- Hypergeometric test for calculating enrichment P-values for gene lists
- GSEA and minimum hypergeometric (mHG) test for computing enrichment P-values for ranked lists
- Multiple test corrections:
  - Bonferroni
  - Benjamini-Hochberg FDR

Module 2

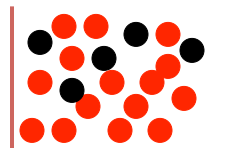
bioinformatics.ca

# The hypergeometric test

a.k.a., Fisher's exact test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



**Null hypothesis:** List is a random sample from population

**Alternative hypothesis:** More black genes than expected

Background population:

500 black genes,

4500 red genes

Module 2

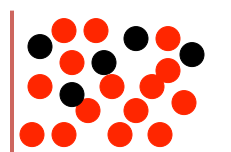
bioinformatics.ca

# The hypergeometric test

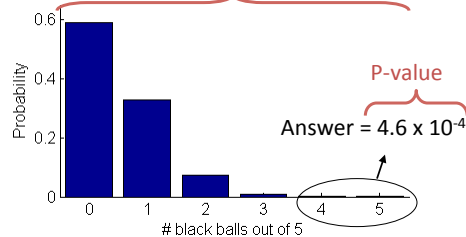
a.k.a., Fisher's exact test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Null distribution



Background population:

500 black genes,

4500 red genes

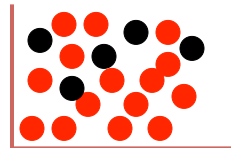
Module 2

bioinformatics.ca

## 2x2 contingency table for Fisher's Exact Test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Background population:  
500 black genes,  
4500 red genes

	In gene list	Not in gene list
In gene set	4	496
Not in gene set	1	4499

Module 2

bioinformatics.ca

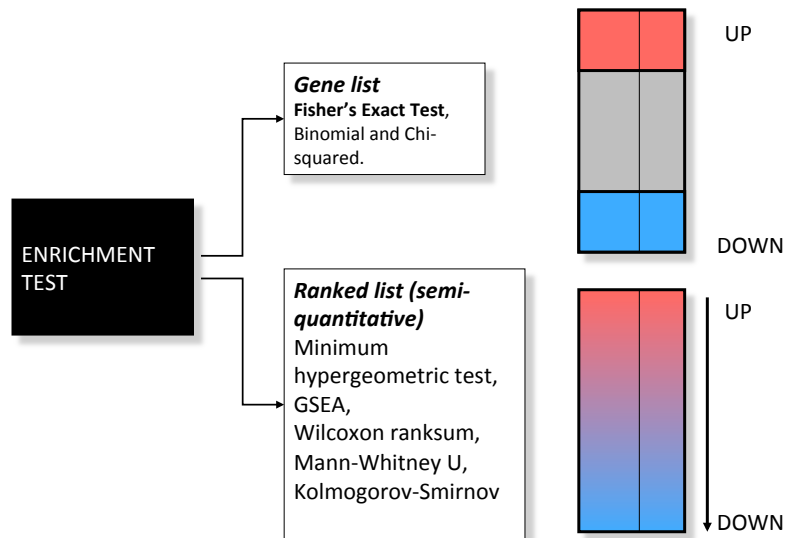
## Important details

- To test for *under-enrichment* of “black”, test for *over-enrichment* of “red”.
- Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.
- To test for enrichment of more than one independent types of annotation (red vs black and circle vs square), apply Fisher's exact test separately for each type. \*\*\*More on this later\*\*\*

Module 2

bioinformatics.ca

## Other enrichment tests



Module 2

bioinformatics.ca

## Minimum hypergeometric test (mHG)

### Steps

1. Calculate P-value at multiple thresholds
2. Correct for multiple testing (or compute empirical P-values using permutations)

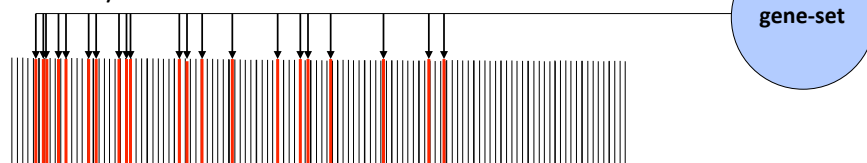
Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol.* 2007 Mar 23;3(3):e39

Module 2

bioinformatics.ca

## GSEA/mHG: Method

GSEA/mHG score calculation



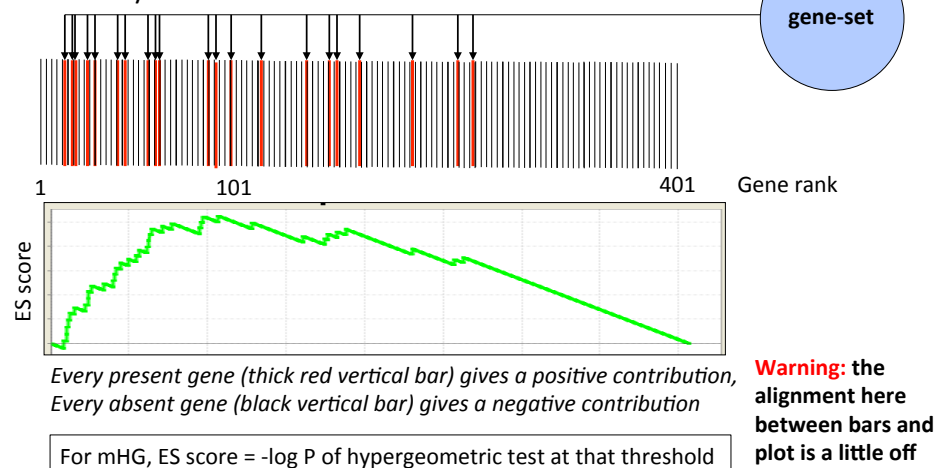
*Where are the gene-set genes located in the ranked list?  
Is there distribution random, or is there an enrichment in either end?*

Module 2

bioinformatics.ca

## GSEA/mHG: Method

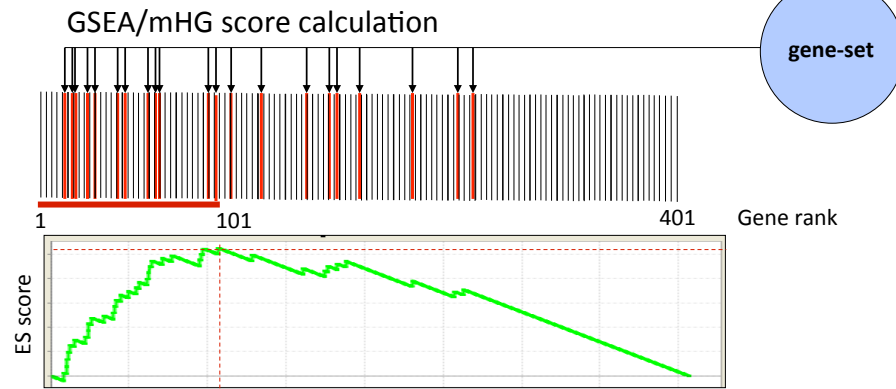
GSEA/mHG score calculation



Module 2

bioinformatics.ca

## GSEA/mHG: Method



1. Maximum (or minimum) ES score is the final **ES score** for the gene set
2. Can define "leading edge subset" as all those genes ranked as least as high as the enriched set.

Module 2

bioinformatics.ca

## Going from ES score → P-value

### Two options

1. For GSEA and mHG, can compute empirical P-values using permutations (see following slides)
2. For mHG, you have another option, you can use a multiple test correction.

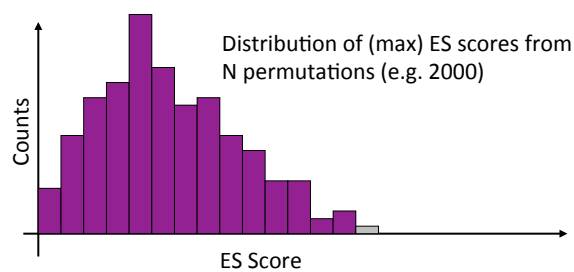
Module 2

bioinformatics.ca

## Permutation-based P-values

Empirical p-value estimation (for every gene-set)

1. Generate null-hypothesis distribution from randomized data (see permutation settings)

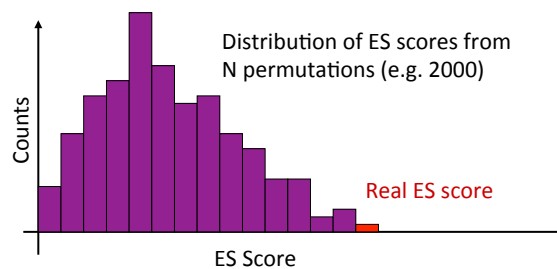


Module 2

bioinformatics.ca

## Permutation-based P-values

Estimate empirical p-value by comparing observed max ES score to null-hypothesis distribution from randomized data (for every gene-set)

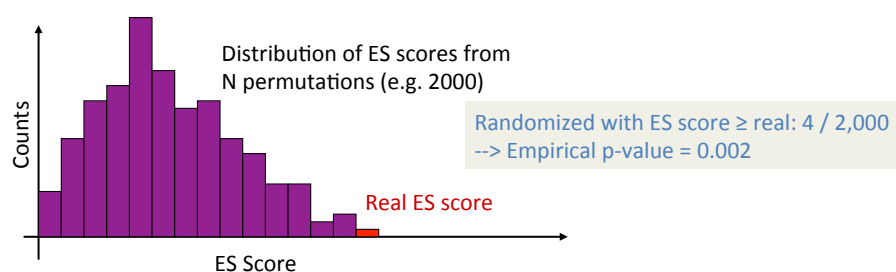


Module 2

bioinformatics.ca

## Permutation-based P-values

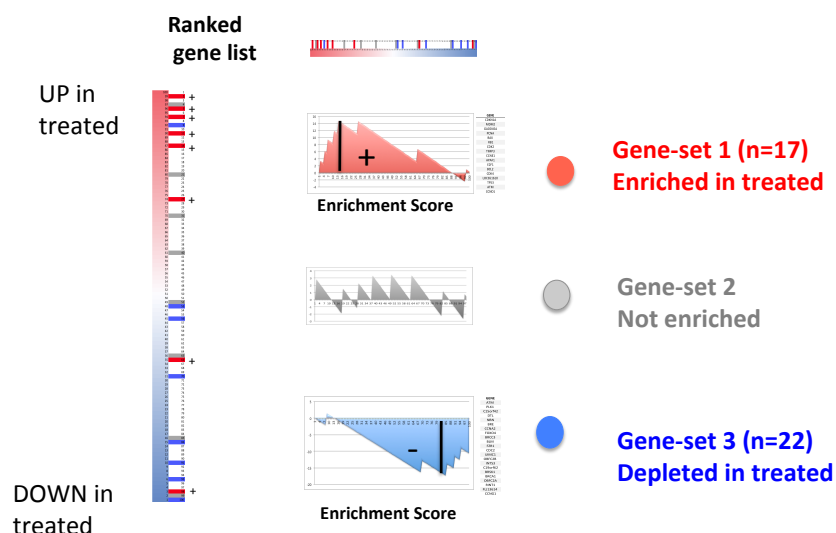
Estimate empirical p-value by comparing observed max ES score to null-hypothesis distribution from randomized data (for every gene-set)



Module 2

bioinformatics.ca

## More GSEA examples



Module 2

bioinformatics.ca

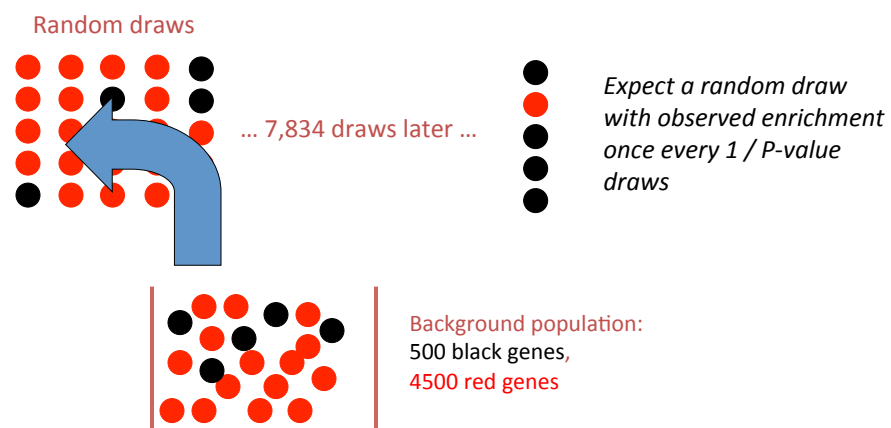


## Multiple test corrections

Module 2

bioinformatics.ca

### How to win the P-value lottery, part 1



Module 2

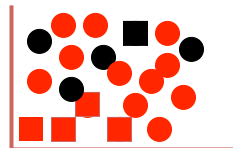
bioinformatics.ca

## How to win the P-value lottery, part 2

Keep the gene list the same, evaluate different annotations

### Observed draw

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



### Different annotation

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



## Simple P-value correction: Bonferroni

If  $M$  = # of annotations tested:

Corrected P-value =  $M \times$  original P-value

Corrected P-value is greater than or equal to the probability that **one or more** of the observed enrichments could be due to random draws. The jargon for this correction is “**controlling for the Family-Wise Error Rate (FWER)**”

## Bonferroni correction caveats

- Bonferroni correction is very stringent and can “wash away” real enrichments leading to false negatives,
- Often one is willing to accept a less stringent condition, the “false discovery rate” (FDR), which leads to a gentler correction when there are real enrichments.

## False discovery rate (FDR)

- FDR is *the expected **proportion** of the observed enrichments due to random chance.*
- Compare to Bonferroni correction which is a bound on *the probability that **any one** of the observed enrichments could be due to random chance.*
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the “q-value”

## Benjamini-Hochberg example I

Rank	Category	(Nominal) P-value
1	<i>Transcriptional regulation</i>	0.001
2	<i>Transcription factor</i>	0.002
3	<i>Initiation of transcription</i>	0.003
4	<i>Nuclear localization</i>	0.0031
5	<i>Chromatin modification</i>	0.005
...	...	...
52	<i>Cytoplasmic localization</i>	0.97
53	<i>Translation</i>	0.99

Sort P-values of all tests in increasing order

## Benjamini-Hochberg example II

Rank	Category	(Nominal) P-value	Adjusted P-value
1	<i>Transcriptional regulation</i>	0.001	$0.001 \times 53/1 = 0.053$
2	<i>Transcription factor</i>	0.002	$0.002 \times 53/2 = 0.053$
3	<i>Initiation of transcription</i>	0.003	$0.003 \times 53/3 = 0.053$
4	<i>Nuclear localization</i>	0.0031	$0.0031 \times 53/4 = 0.040$
5	<i>Chromatin modification</i>	0.005	$0.005 \times 53/5 = 0.053$
...	...	...	...
52	<i>Cytoplasmic localization</i>	0.97	$0.985 \times 53/52 = 1.004$
53	<i>Translation</i>	0.99	$0.99 \times 53/53 = 0.99$

Adjusted P-value is “nominal” P-value times # of tests divided by the rank of the P-value in sorted list

Adjusted P-value = P-value X [# of tests] / Rank

## Benjamini-Hochberg example III

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	<i>Transcriptional regulation</i>	0.001	$0.001 \times 53/1 = 0.053$	0.040
2	<i>Transcription factor</i>	0.002	$0.002 \times 53/2 = 0.053$	0.040
3	<i>Initiation of transcription</i>	0.003	$0.003 \times 53/3 = 0.053$	0.040
4	<i>Nuclear localization</i>	0.0031	$0.0031 \times 53/4 = 0.040$	0.040
5	<i>Chromatin modification</i>	0.005	$0.005 \times 53/5 = 0.053$	0.053
...	...	...	...	...
52	<i>Cytoplasmic localization</i>	0.97	$0.985 \times 53/52 = 1.004$	0.99
53	<i>Translation</i>	0.99	$0.99 \times 53/53 = 0.99$	0.99

**Q-value (or FDR) corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.**

Module 2

bioinformatics.ca

## Benjamini-Hochberg example III

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	<i>Transcriptional regulation</i>	0.001	$0.001 \times 53/1 = 0.053$	0.040
2	<i>Transcription factor</i>	0.002	$0.002 \times 53/2 = 0.053$	0.040
3	<i>Initiation of transcription</i>	0.003	$0.003 \times 53/3 = 0.053$	0.040
4	<i>Nuclear localization</i>	0.0031	$0.0031 \times 53/4 = 0.040$	0.040
5	<i>Chromatin modification</i>	0.005	$0.005 \times 53/5 = 0.053$	0.053
...	...	...	...	...
52	<i>Cytoplasmic localization</i>	0.97	$0.985 \times 53/52 = 1.004$	0.99
53	<i>Translation</i>	0.99	$0.99 \times 53/53 = 0.99$	0.99

P-value threshold for FDR < 0.05

Red: non-significant

Green: significant at FDR < 0.05

**P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold**

Module 2

bioinformatics.ca

## Reducing multiple test correction stringency

- The correction to the P-value threshold  $\alpha$  depends on the # of tests that you do, so, no matter what, the more tests you do, the more sensitive the test needs to be
- Can control the stringency by reducing the number of tests: e.g. use GO slim; restrict testing to the appropriate GO annotations; or filter gene sets by size.

## Summary

- Enrichment analysis:
  - Statistical tests
    - Gene list: **Fisher's Exact Test**
    - Ranked list: **mHG, GSEA**, also see Wilcoxon ranksum, Mann-Whitney U-test, Kolmogorov-Smirnov test
  - Multiple test correction
    - **Bonferroni**: stringent, controls probability of at least one false positive\*
    - **FDR**: more forgiving, controls expected proportion of false positives\* -- typically uses Benjamini-Hochberg

\* Type 1 error, aka probability that observed enrichment if no association

## Learning Objectives of Module 2

- **Be able** to select the appropriate enrichment test for your data.
- **Be able** to determine the appropriate background gene list when running Fisher's Exact Test (aka Hypergeometric test).
- **Be able** to compute a minimum hypergeometric test on a ranked list
- **Be able** to determine when you need a multiple test correction.
- **Be able** to select whether to use a Bonferroni corrected P-value or a false discovery rate.
- **Be able** to explain, in plain language, how you calculate each correction.