Canadian Bioinformatics Workshops

www.bioinformatics.ca

In collaboration with

Cold Spring Harbor Laboratory

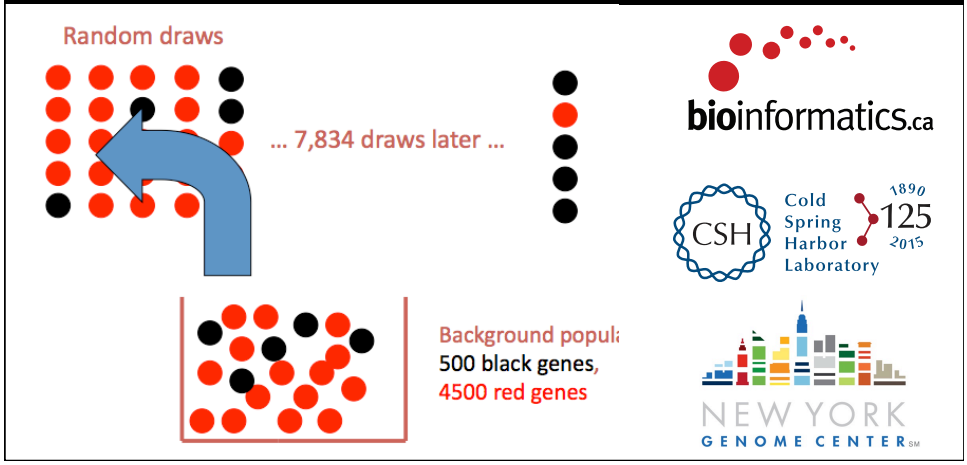
&

New York Genome Center

Module 2
Finding over-represented pathways in gene lists

Quaid Morris

High-throughput Biology: From Sequence to Networks
April 27-May 3, 2015



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.

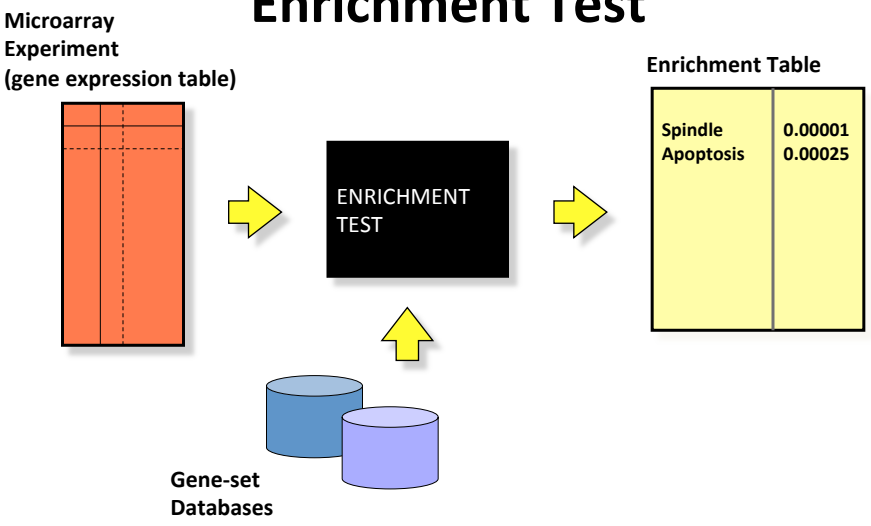
Outline

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

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 (+ others we won't discuss)

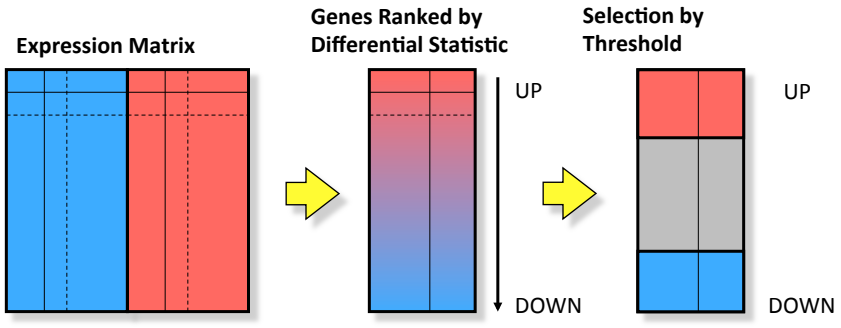
Enrichment Test



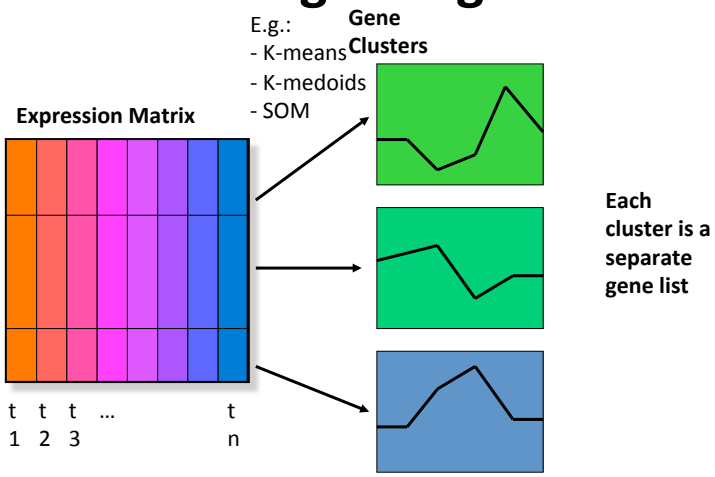
Gene list enrichment analysis

- 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

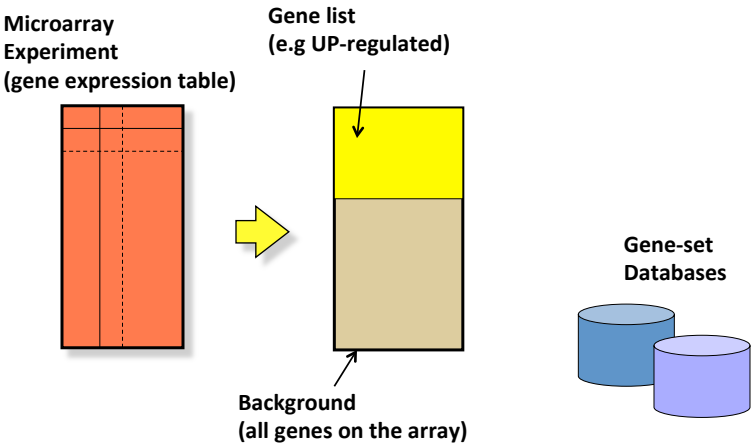
Two-class design for gene lists



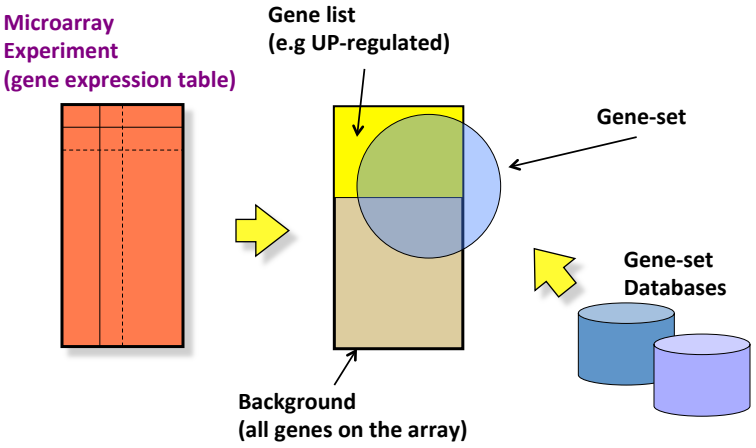
Time-course design for gene lists



Example gene list enrichment test

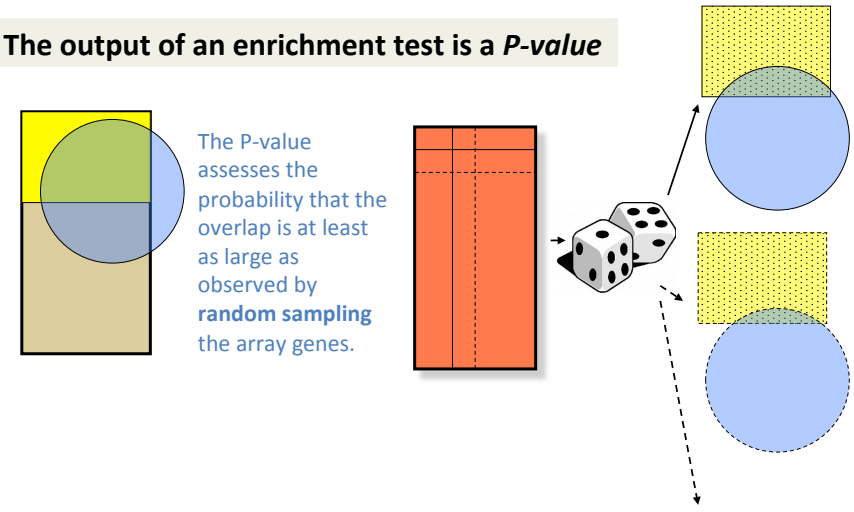


Example gene list enrichment test



Enrichment Test

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



The P-value assesses the probability that the overlap is at least as large as observed by **random sampling** the array genes.

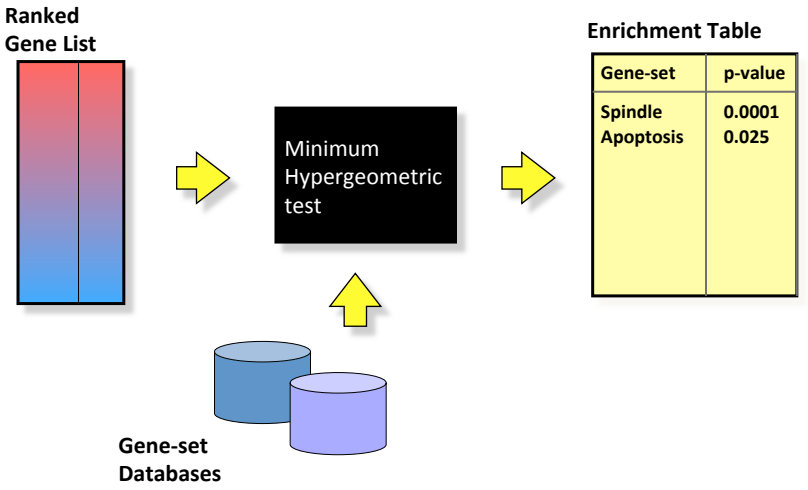
Recipe for gene list enrichment test

- **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! ;)

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! ;)

Outline of theory component

- Hypergeometric test for calculating enrichment P-values for gene lists
- Minimum hypergeometric test for computing enrichment P-values for ranked lists
- Multiple test corrections:
  - Bonferroni
  - Benjamini-Hochberg FDR

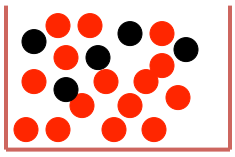
The hypergeometric test

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



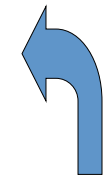
Background population:
500 black genes,
4500 red genes

The hypergeometric test

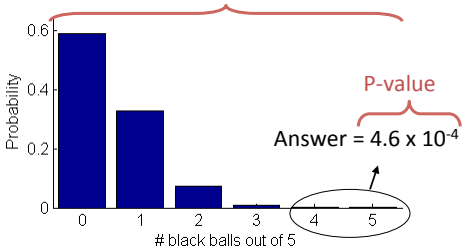
a.k.a., Fisher's exact test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

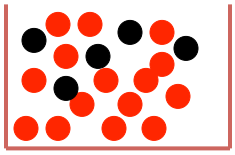


Null distribution



P-value

Answer = 4.6×10^{-4}

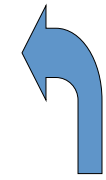


Background population:
500 black genes,
4500 red genes

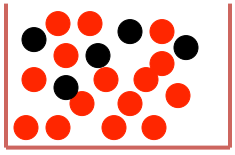
2x2 contingency table for Fisher's Exact Test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



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

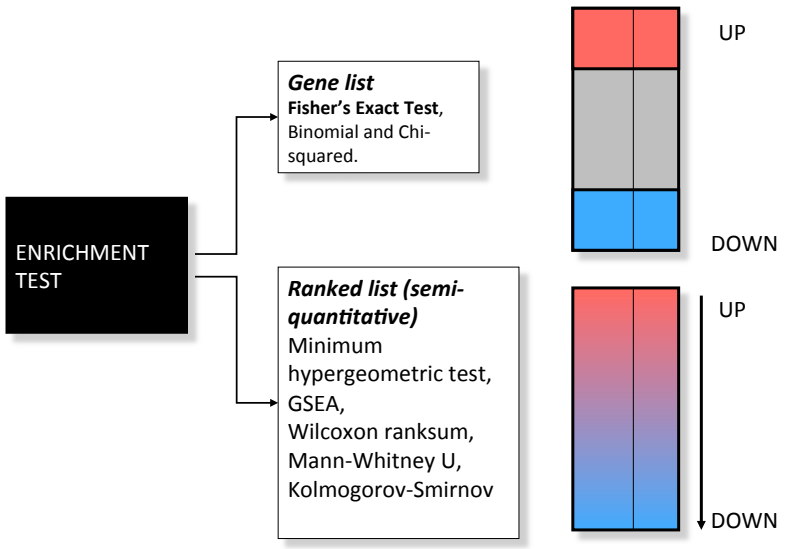


Background population:
500 black genes,
4500 red genes

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***

Other enrichment tests



Minimum hypergeometric test (mHG)

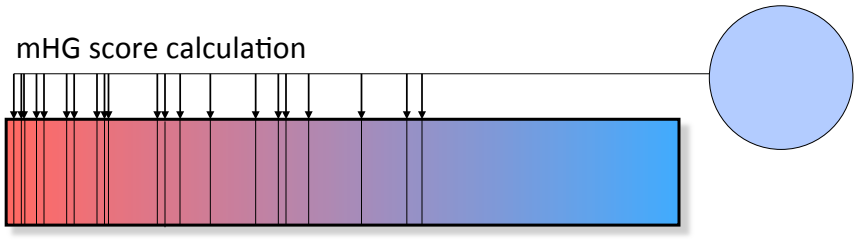
Steps

1. Calculate P-value at multiple thresholds

2. Correct for multiple testing (or compute empirical P-values using permutations)

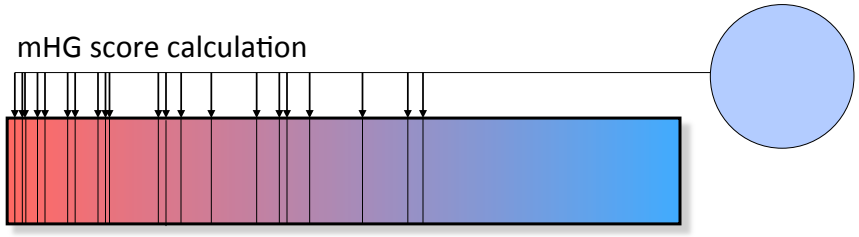
Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 2007 Mar 23;3(3):e39

mHG: Method



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

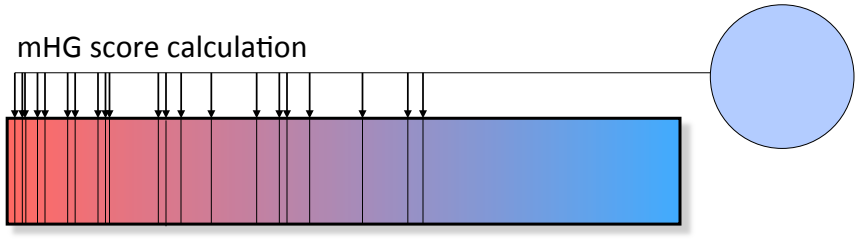
mHG: Method



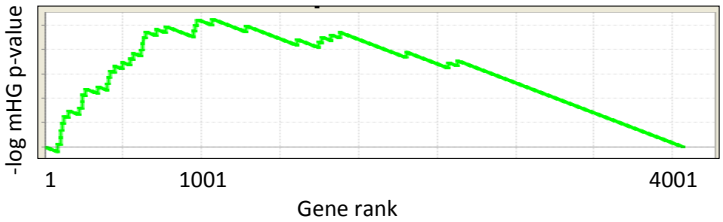
Where are the gene-set genes located in the ranked list?
Is there distribution random, or is there an enrichment in either end?

In next slide, I'm showing "e = $-\log_{10}$ P-value",
e.g. if p = 0.001, then e = 3

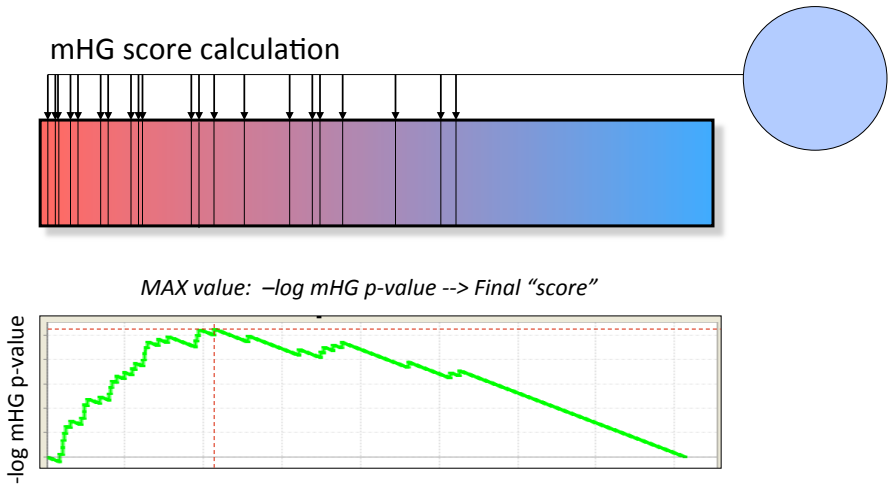
mHG: Method



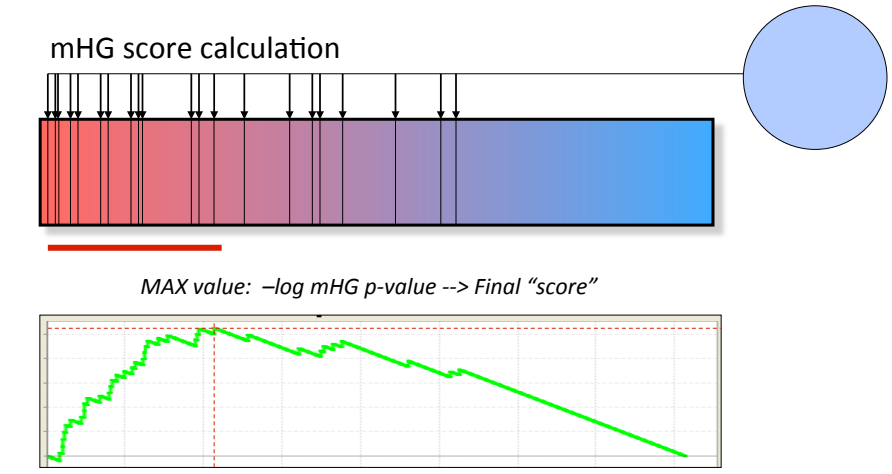
Every present gene (black vertical bar) gives a positive contribution,
every absent gene (no vertical bar) gives a negative contribution



mHG: Method



mHG: Method



Correcting mHG "score" for multiple testing

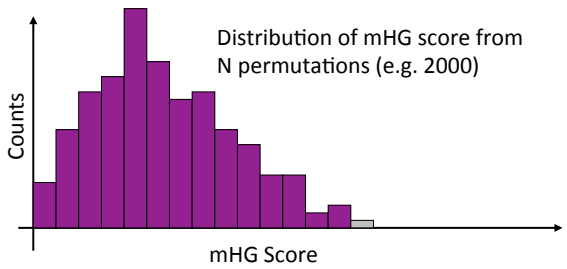
Two options

1. Use a multiple test correction (see next section)
2. Compute empirical P-values using permutations (see following slides)

mHG: permutation correction

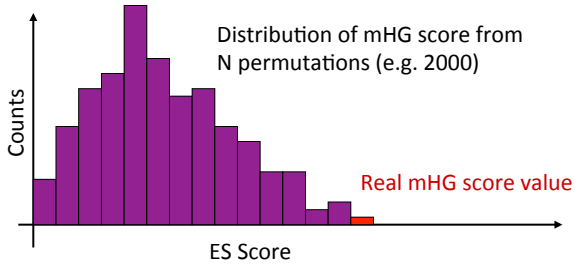
Empirical p-value estimation (for every gene-set)

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



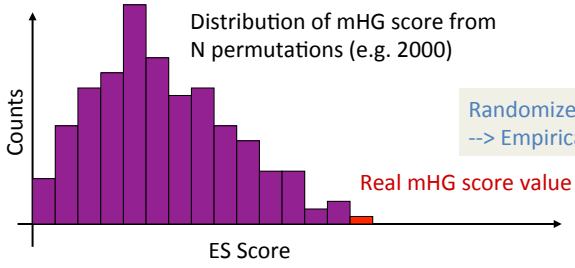
mHG: permutation correction

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



mHG: permutation correction

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

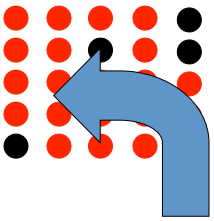


Randomized with mHG ≥ real: 4 / 2,000
--> Empirical p-value = 0.002

Multiple test corrections

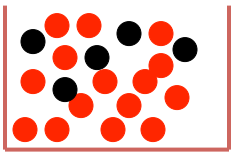
How to win the P-value lottery, part 1

Random draws



… 7,834 draws later …

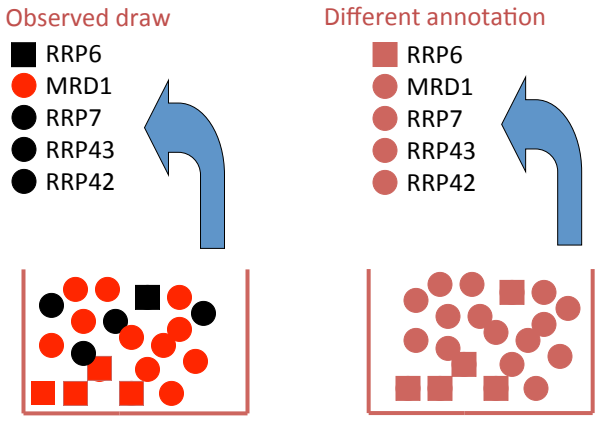
Expect a random draw with observed enrichment once every 1 / P-value draws



Background population:
500 black genes,
4500 red genes

How to win the P-value lottery, part 2

Keep the gene list the same, evaluate different annotations



Simple P-value correction: Bonferroni

If *M* = # of annotations tested:

Corrected P-value = *M* x 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	*Transcriptional regulation*	0.001
2	*Transcription factor*	0.002
3	*Initiation of transcription*	0.003
4	*Nuclear localization*	0.0031
5	*Chromatin modification*	0.005
…	…	…
52	*Cytoplasmic localization*	0.97
53	*Translation*	0.99

Sort P-values of all tests in decreasing order

Benjamini-Hochberg example II

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

Benjamini-Hochberg example III

P-value threshold for FDR < 0.05

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

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

Reducing multiple test correction stringency

- The correction to the P-value threshold α 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**, also see GSEA, 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.

We are on a Coffee Break & Networking Session

bioinformatics.ca

CSH Cold Spring Harbor Laboratory 1890 125 2015

NEW YORK GENOME CENTER