Lecture 6: Contingency Tables

BMI 713
November 7, 2017
Peter J Park

Contingency Table

- Previously, we tested the null hypothesis that the two proportions p_1 and p_2 from two populations were equal.
- Example: Gene expression profiling study was conducted to find genes that are differentially expressed between tumor and normal cells from an individual. Out of 20,000 genes total, 1000 were differentially expressed. In the p53 pathway, there are 100 genes, 10 of which were among the list of differentially expressed genes. I would like to test the hypothesis that the p53 pathway is involved.

Contingency Table

- Alternatively, we can apply a different approach

	DE genes	Not DE genes	Total
p53 pathway	10	90	100
other genes	990	18910	19900
total	1000	19000	20000

- With this method, data are arranged in the form of a contingency table
- This is a 2 x 2 table for two dichotomous random variables
- Row and column assignments are arbitrary
- The subjects in the two columns are independent

Chi-square Test

- To carry out the test, we first calculate the expected counts for the table assuming that $\boldsymbol{H_0}$ **is true** and $\boldsymbol{p_1 = p_2}$
- The chi-square test compares the observed frequencies in each category with the expected frequencies given that H_0 is true
- Are the deviations between observed (O_{ij}) and expected values (E_{ij}) too large to be attributed to chance?
- To determine this, deviations from all 4 cells must be combined to form the statistic

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Chi-square Test

Observed:

	col 1	col 2	Total
row 1	O_{11}	O_{12}	O_{1-}
row 2	O_{21}	O_{22}	O_{2-}
total	O_{-1}	O_{-2}	Total

Expected:

	col 1	col 2	Total
row 1	E_{11}	E_{12}	E_{1-}
row 2	E_{21}	E_{22}	E_{2-}
total	E_{-1}	E_{-2}	Total

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$$

$$= \frac{\left(O_{11}-E_{11}\right)^2}{E_{11}} + \frac{\left(O_{12}-E_{12}\right)^2}{E_{12}} + \frac{\left(O_{21}-E_{21}\right)^2}{E_{21}} + \frac{\left(O_{22}-E_{22}\right)^2}{E_{22}}$$

Example

	DE genes	Not DE genes	Total
p53 pathway	10	90	100
other genes	990	18910	19900
total	1000	19000	20000

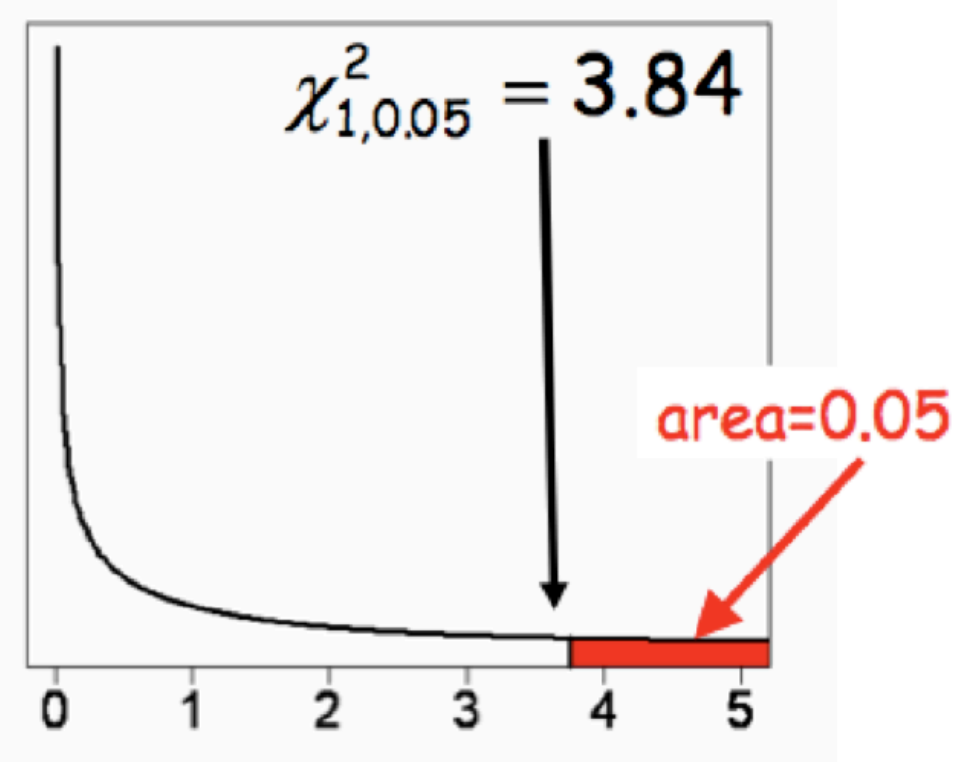
	DE genes	Not DE genes	Total
p53 pathway	5	95	100
other genes	995	18905	19900
total	1000	19000	20000

$$X^2 = \frac{(10-5)^2}{5} + \frac{(90-95)^2}{95} + \frac{(990-995)^2}{995} + \frac{(18910-18905)^2}{18905} = 5.29$$

- If the null hypothesis is true, the distribution of X_2 is approximated by the **chi-square** distribution with 1 df

Chi-square Test

- The null hypothesis is rejected at the α level if X^2 is too large
- If $\alpha = 0.05$, we would reject H_0 for X^2 greater than $X^2_{1,.05} = 3.84$



Chi-square Test

- Like the *t* and *F* distributions, the chi-square is a family of distributions indexed by the degrees of freedom
- The alternative is always two-sided
- In order for the approximation to be valid, **no cell in the table should have an expected count less than 5** (a fairly conservative criterion)
- The chi-square test is also a **test of association** between two categorical variables
- Marginals are assumed to be fixed

Yate's correction

- For 2 x 2 tables, a better approximation may be achieved for small samples (smallest expected frequencies of ~5 or 10) by using the test sta

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|O_{ij} - E_{ij}\right| - 1/2\right)^2}{E_{ij}}$$

- It is to prevent overestimation of statistical significance for small data, but it could also over-correct
- Same idea as before: we are using a continuous function to approximate a discrete one:
  - Discrete: P(X=5); Continuous: P(4.5 < X < 5.5)

R Code

- help.search('chi')
- chisq.test() takes a matrix
- > matrix(c(10,90,990,18910),2,2)
- ```
       [,1]  [,2]
  [1,]  10   990
  [2,]  90  18910
  ```
- chisq.test(matrix(c(10,90,990,18910),2,2))
- (10-5)^2/5+(90-95)^2/95+(990-995)^2/995+(18910-18905)^2/18905
- chisq.test(matrix(c(10,90,990,18910),2,2),correct=F)
- prop.test(c(10,90),c(1000,19000),correct=F)
  ```

Fisher's Exact Test

- What happens if the expected cell counts are too small to use the chi-square test as described?
- In the one-sample binomial case, when the sample was too small to use the normal approximation, we used an **exact method** to get the p-value
- Here too we would use an exact method. For a 2x2 table, the method used is **Fisher's exact test**
- Basic idea: enumerate all possible tables (with column sun and row sums fixed) and **count the fraction of tables that are as 'extreme' as or more extreme than the observed table**
- Sum the probabilities associated with these tables to obtain the p-value of the test

- Observed:

	col 1	col 2	Total
row 1	1	4	5
row 2	5	3	8
total	6	7	13

- All tables that could have been observed with fixed marginals

	col 1	col 2	Total
row 1	0	5	5
row 2	6	2	8
total	6	7	13

	col 1	col 2	Total
row 1	1	4	5
row 2	5	3	8
total	6	7	13

	col 1	col 2	Total
row 1	2	3	5
row 2	4	4	8
total	6	7	13

	col 1	col 2	Total
row 1	3	2	5
row 2	3	5	8
total	6	7	13

	col 1	col 2	Total
row 1	4	1	5
row 2	2	6	8
total	6	7	13

	col 1	col 2	Total
row 1	5	0	5
row 2	1	7	8
total	6	7	13

Hypergeometric Distribution

- Calculate the probability associated with each table
- Use the hypergeometric distribution

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

- Give the fixed margins, the probability of obtaining the specific table which was observed is

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

- The p-value of Fisher's exact test is the sum the probabilities associated with the tables as 'extreme' as or more extreme than the observed table

- What if we are interested in a variable that has more than two categories?
- Example: Test for association between eye color and presence or absence of a mutant allele at some genetic locus
- Eye color categories: blue, green, brown, hazel, gray
- Genetic categories: 0 copies mutant allele; ≥ 1 copy mutant allele

	blue	green	brown	hazel	gray	Total
Mutant allele absent						
Mutant allele present						
Total						

R = # rows, C = # columns

R x C Contingency Table

- Chi-square test for R x C table is similar to the one for 2x2 table

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

- Rule of thumb:
  - No more than 1/5 of cells should have expected count <5
  - No cell should have expected count <1
- **What is the "degrees of freedom"?**
- Under H_0, the X^2 test statistics follows a chi-square distribution on *(R-1)(C-1)* degrees of freedom

Pathway Enrichment Analysis

- Given a set of well-annotated pathways, a standard analysis in a genomewide experiment is to detect relevant pathways
- More generally, any “gene set” can be used, e.g.,
  - gene ontology (GO) categories
  - differentially expressed genes, co-expressed genes, genes with the same upstream motif, gene regulated by the same miRNA, etc.
  - GWAS loci, genetic/CRISPR/drug screens, any cluster from cluster analysis, protein-protein interaction data, etc.
- The simple version involves going through all pathways and perform the Chi-square or the Fisher’s exact test

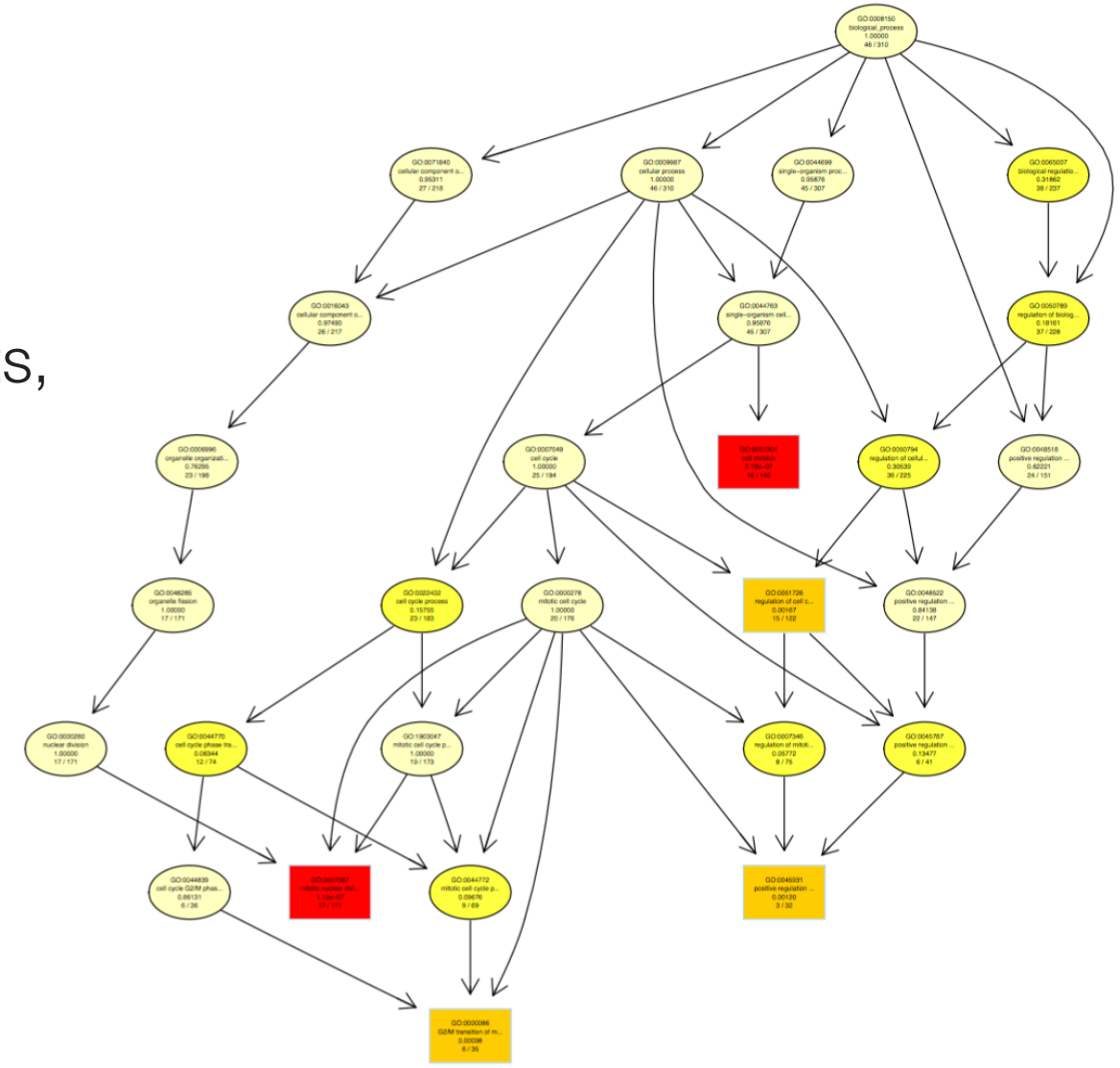
Pathway Enrichment Analysis

- By borrowing strength across the similar set of genes, potential for increased statistical power
- More robust to biological and/or technical variability
- More advanced methods account for the gene order on the list of differentially expressed genes
- There are issues with multiple hypothesis testing

Gene Ontology

- 40,000 biological concepts, multiple organisms

H	**hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
C1	**positional gene sets** for each human chromosome and cytogenetic band.
C2	**curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
C3	**motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
C4	**computational gene sets** defined by mining large collections of cancer-oriented microarray data.
C5	**GO gene sets** consist of genes annotated by the same GO terms.
C6	**oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
C7	**immunologic signatures** defined directly from microarray gene expression data from immunologic studies.



MSigDB
Molecular Signatures Database

- Other sources: Reactome, KEGG, etc.

Which test?

- Proportion test
- Chi-square test
- Fisher’s exact

- Ordered or unordered?
- Must you applied a threshold to define a list?

Gene Set Enrichment Analysis (GSEA)

- Ordered or unordered?
- Must you applied a threshold to define a list?
- Solution: use a ranked list.

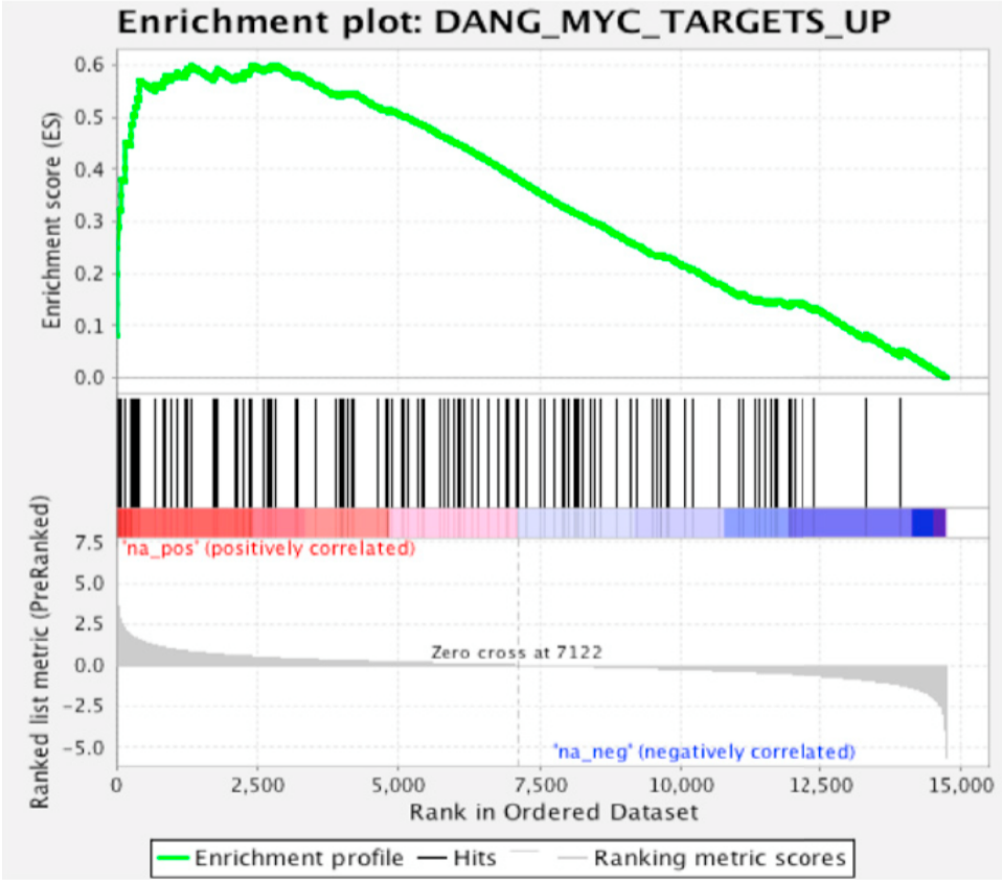
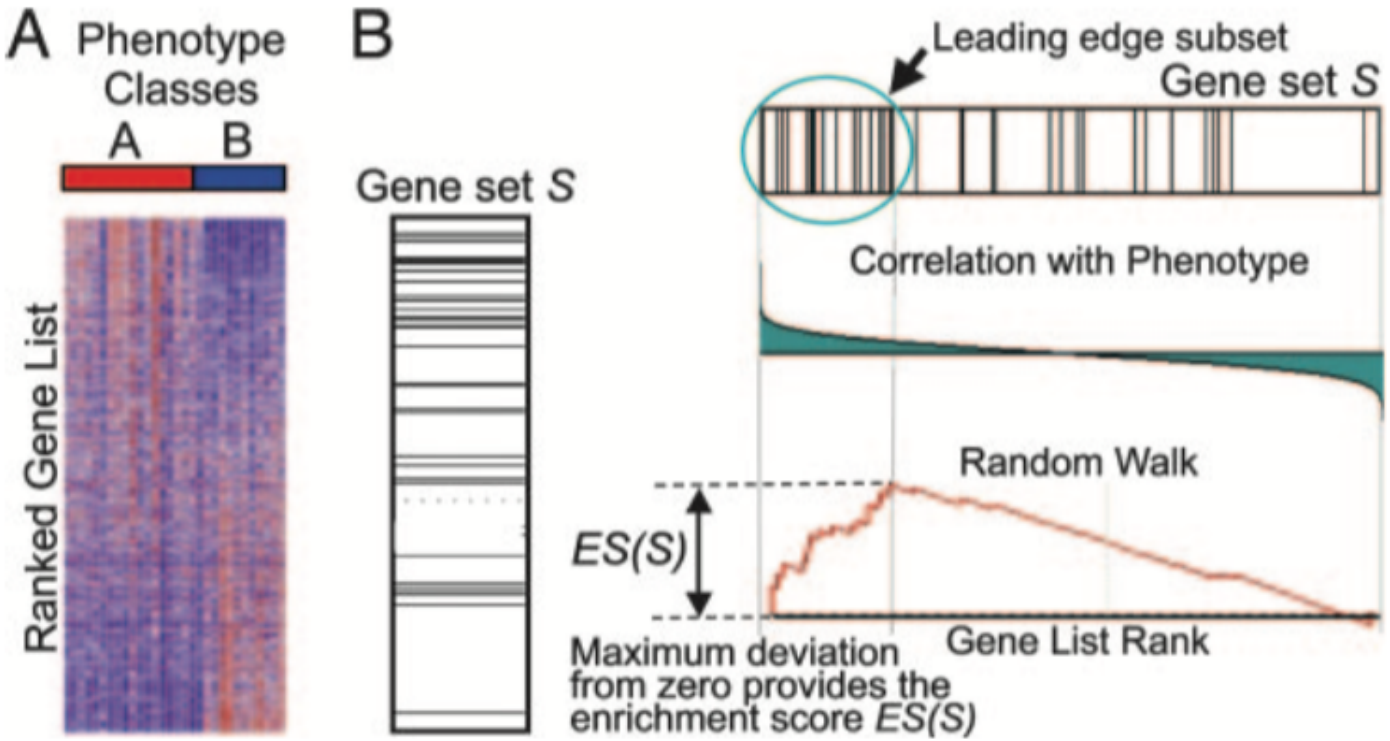


Figure 4b from Marcotte et al, *Cell*, 2016
GSEA of trans-essential genes for MYC targets (FDR < 0.0001).

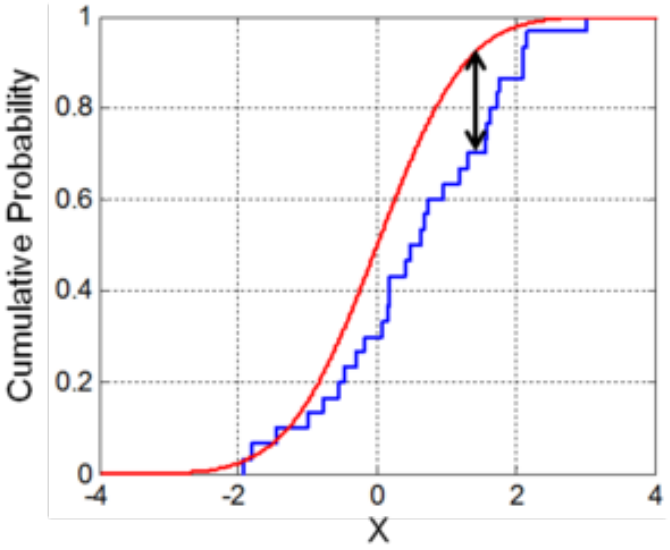
GSEA



1. All genes are ranked by using a signal-to-noise ratio;
2. For each gene set, the distribution of gene ranks from the gene set is compared against the distribution for the rest of the genes by using the enrichment score (ES) based on a one-sided Kolmogorov–Smirnov statistic;
3. Class labels are permuted to generate a null distribution of ES; and
4. Statistical significance of the observed score is assessed for the top-ranking gene set by comparison with the null distribution of maximum scores from each permutation.

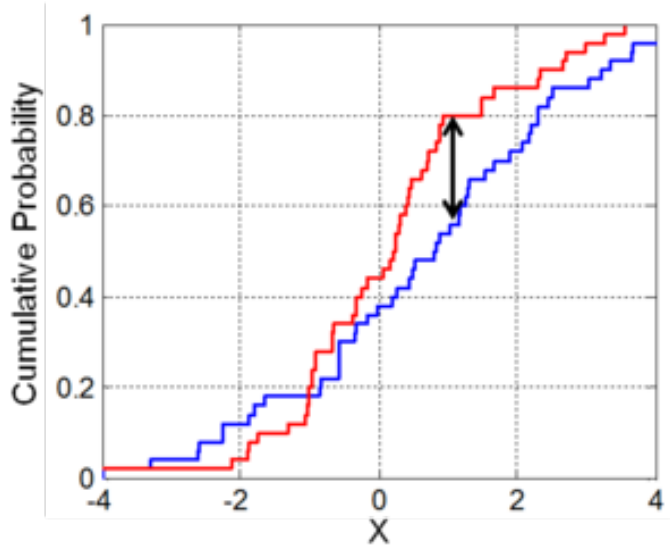
Kolmogorov-Smirnov test

- We have *n* observations $X_1,\ldots,X_n$. We would to test whether they came from a distribution P
- H_0: the samples come from P
- H_1: the samples do not come from P



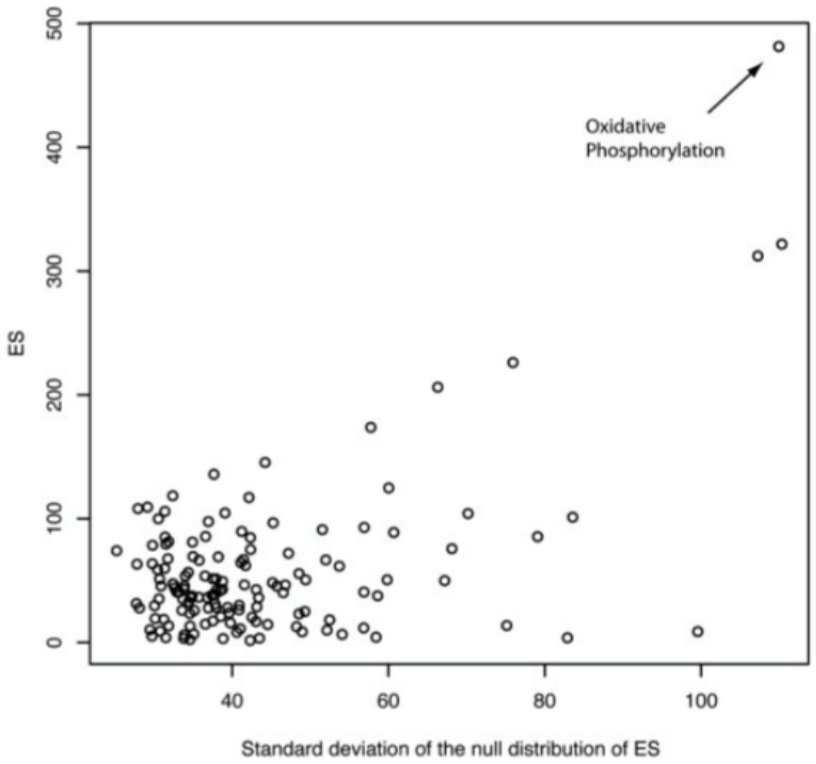
- Kolmogorov-Smirnov (K-S) statistic

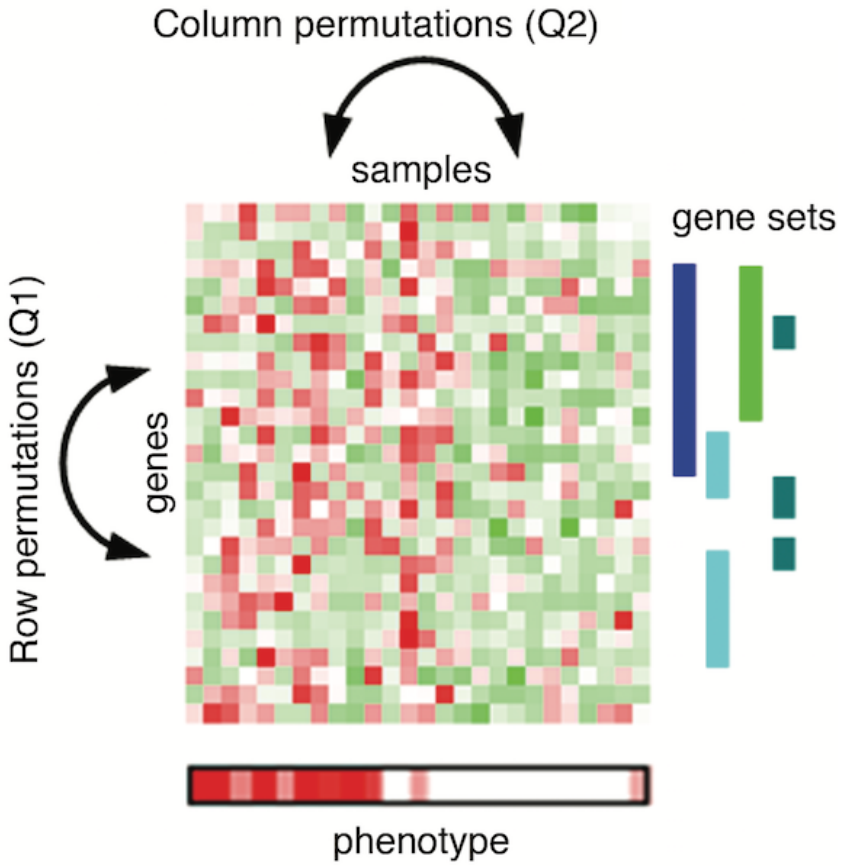
$$\max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|$$



GSEA - optimal?

- Enrichment score (ES) - does it depend on gene set size? On correlation structure?





Tian et al, *PNAS*, 2005

Discovering statistically significant pathways in expression profiling studies

Lu Tian[†], Steven A. Greenberg[‡§], Sek Won Kong[¶‖], Josiah Altschuler[¶], Isaac S. Kohane[§††], and Peter J. Park[§†††‡‡]

[†]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 680 North Lake Shore Drive, Chicago, IL 60611; [‡]Department of Neurology, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115; [§]Children's Hospital Informatics Program, 300 Longwood Avenue, Boston, MA 02115; [¶]Bauer Center for Genomics Research, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; [‖]Molecular Medicine, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA 02215; and [††]Harvard-Partners Center for Genetics and Genomics, 77 Avenue Louis Pasteur, Boston, MA 02115

Communicated by Louis M. Kunkel, Harvard Medical School, Boston, MA, August 2, 2005 (received for review February 7, 2005)

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [c]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [d]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [e]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [f]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [g]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [i]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [j]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Eric S. Lander, August 2, 2005

Improvements:

- Weigh the steps according to each gene's correlation with a phenotype, so that the sets clustered in the middle of the list do not score high
- Normalize for gene set size
- FWER -> FDR

The Annals of Applied Statistics
2007, Vol. 1, No. 1, 107–129
DOI: 10.1214/07-AOAS101

ON TESTING THE SIGNIFICANCE OF SETS OF GENES

BY BRADLEY EFRON[1] AND ROBERT TIBSHIRANI[2]

Stanford University

This paper discusses the problem of identifying differentially expressed groups of genes from a microarray experiment. The groups of genes are externally defined, for example, sets of gene pathways derived from biological databases. Our starting point is the interesting Gene Set Enrichment Analysis (GSEA) procedure of Subramanian et al. [*Proc. Natl. Acad. Sci. USA* **102** (2005) 15545–15550]. We study the problem in some generality and propose two potential improvements to GSEA: the *maxmean* statistic for summarizing gene-sets, and *restandardization* for more accurate inferences. We discuss a variety of examples and extensions, including the use of gene-set scores for class predictions. We also describe a new R language package *GSA* that implements our ideas.

The point is that any method for assessing gene-sets should compare a given gene-set score not only to scores from permutations of the sample labels, but also take into account scores from sets formed by random selections of genes.

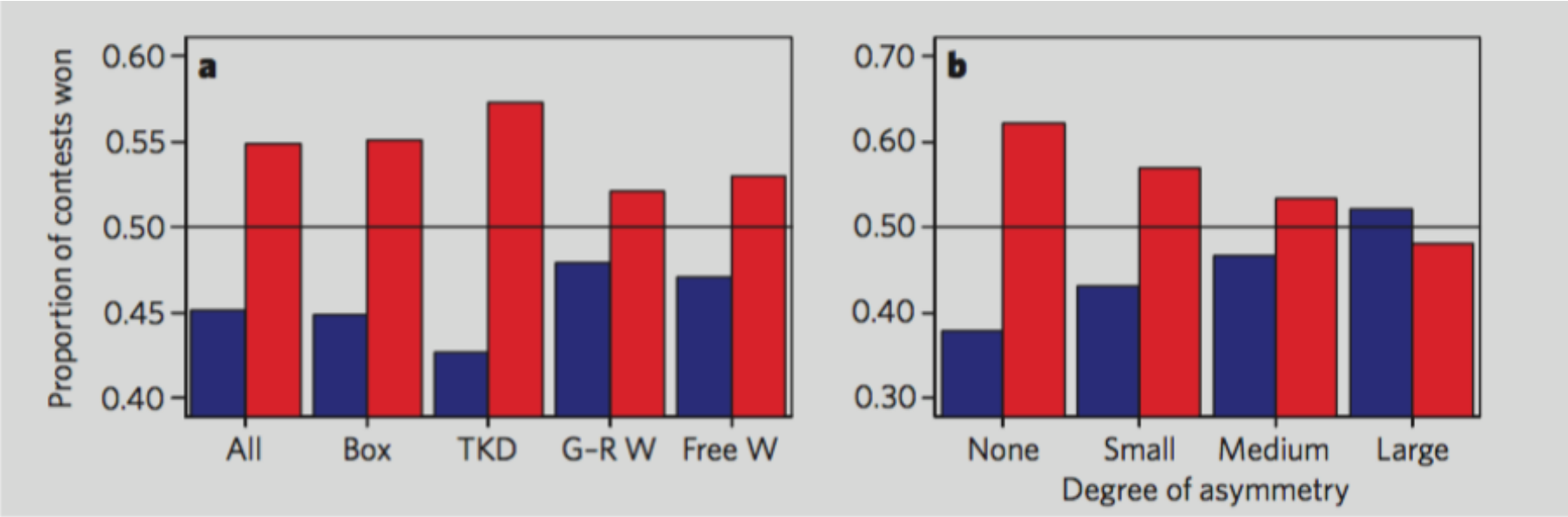
NATURE|Vol 435 | 19 May 2005 nature

BRIEF COMMUNICATIONS

Red enhances human performance in contests

Signals biologically attributed to red coloration in males may operate in the arena of combat sports.

Red coloration is a sexually selected, testosterone-dependent signal of male quality in a variety of animals[1–5], and in some non-human species a male's dominance can be experimentally increased by attaching artificial red stimuli[6]. Here we show that a similar effect can influence the outcome of physical contests in humans — across a range of sports, we find that wearing red is consistently associated with a higher probability of winning. These results indicate not only that sexual selection may have influenced the evolution of human response to colours, but also that the colour of sportswear needs to be taken into account to ensure a level playing field in sport.



- Boxing, Tae Kwon Do, Greco–Roman wrestling and freestyle wrestling in 2004 Olympics
- “Randomly assigned red or blue outfits”
- Fig 1: Chi-sq = 4.19, d.f. 1, P=0.041
- Nature News: “A red face is commonly associated with anger and aggression, so a bright red shirt or headgear may intimidate an opponent, suggests Hill, who unveils his results in this week's *Nature*. Alternatively, red clothes could actually boost the wearer's testosterone levels, he says: "Maybe you get a surge when you pull on that red shirt.” ”

Red enhances performance?

- “remarkably consistent across rounds in each competition, with 16 of 21 rounds having more red than blue winners, and only four rounds having more blue winners (sign test, P=0.012).”
- In team sports too? A preliminary analysis Euro 2004 (in which teams wore shirts of different colours in different matches)
- Five teams that wore “predominantly red”; four played the other matches in white, one in blue. All five had better results when playing in red (paired t-test, t= -3.15, d.f.=4, P=0.034), largely as a result of scoring more goals (t= -2.98, d.f.=4, P=0.041)

	A	B	C	D	E	F	G	H
1	Weight Class	Red Boxer ID	Blue Boxer ID	Round of Competition	Winner	Method of Win	Points Scored by Red	Points Scored by Blue
2	48kg	4804	4805	Last 32	Red	Points	20	8
3	48kg	4806	4807	Last 32	Red	Points	48	25
4	48kg	4808	4809	Last 32	Blue	Referee Stopped Contest - Outscored	.	.
5	48kg	4810	4811	Last 32	Red	Points	22	7
6	48kg	4812	4813	Last 32	Blue	Points	8	23
7	48kg	4814	4815	Last 32	Blue	Points	9	22
8	48kg	4816	4817	Last 32	Red	Points	26	21
9	48kg	4818	4819	Last 32	Red	Points	21	7
10	48kg	4820	4821	Last 32	Blue	Points	20	27
11	48kg	4822	4823	Last 32	Red	Referee Stopped Contest - Outscored	.	.
12	48kg	4824	4825	Last 32	Blue	Points	12	17
13	48kg	4826	4827	Last 32	Red	Referee Stopped Contest - Outscored	.	.
14	48kg	4828	4829	Last 32	Red	Points	26	14
15	48kg	4801	4802	Last 16	Blue	Points	20	29