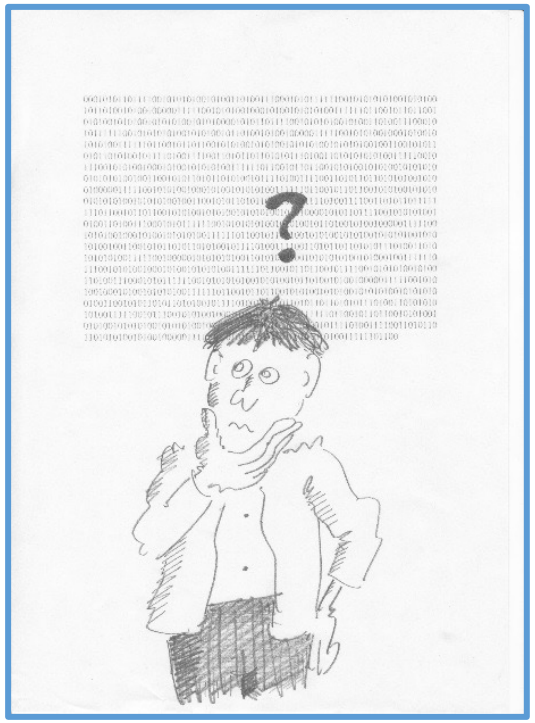
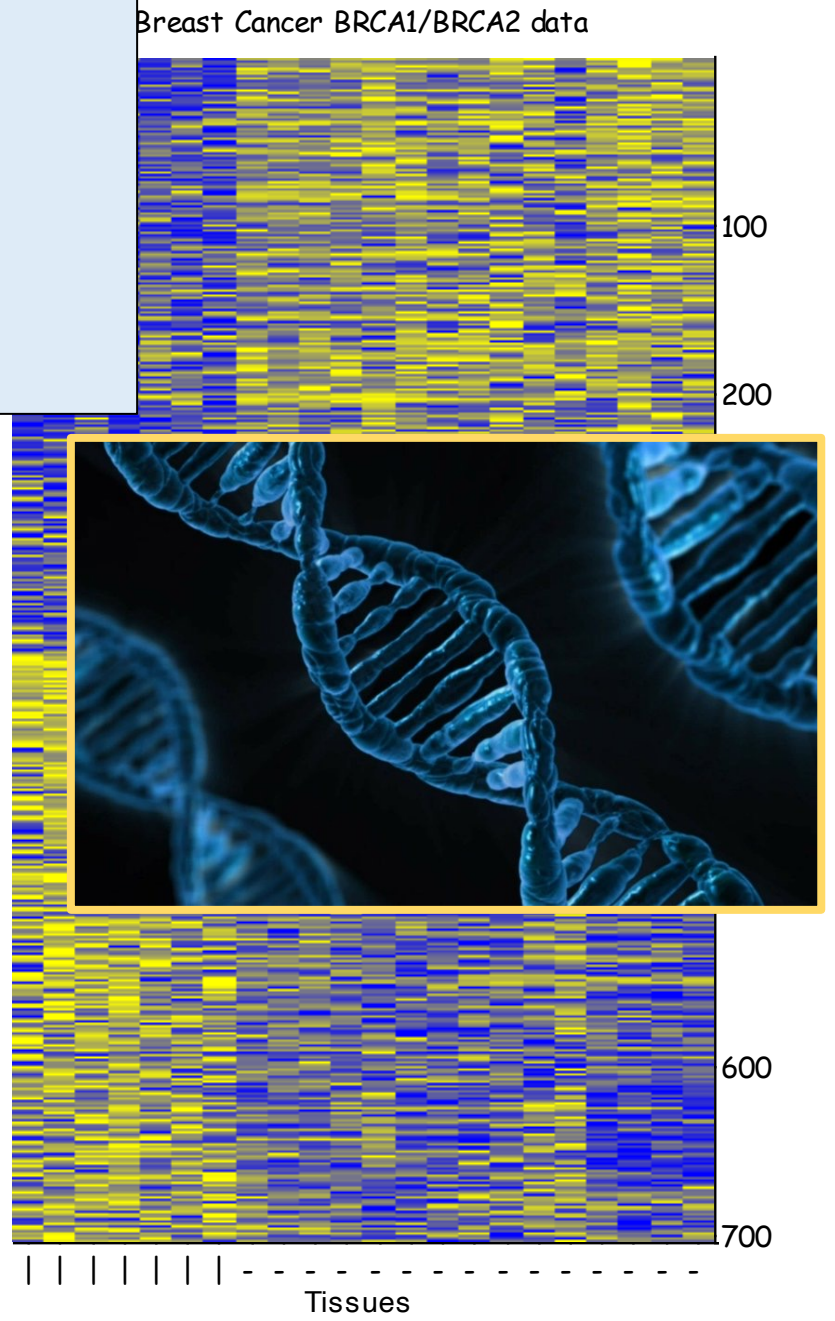
Introduction to differential gene expression, multiple testing and FDR

Zohar Yakhini, Leon Anavy,
Ben Galili – Reichman.
Partially based on slides from Zohar Yakhini,
Doron Lipson, Itai Sharon and others,
at the Technion

IDC HERZLIYA





The concept of p-value

Under a given NULL MODEL, what is the probability of observing a value for a MEASURED QUANTITY which is as or more extreme than the one actually measured in data?

We tossed a coin 100 times and observed 27 Hs.

Under a FAIR INDEPENDENT COIN TOSSING MODEL, what is the probability of observing a NUMBER OF Hs which is less than 27 out of 100.

IDC HERZLIYA

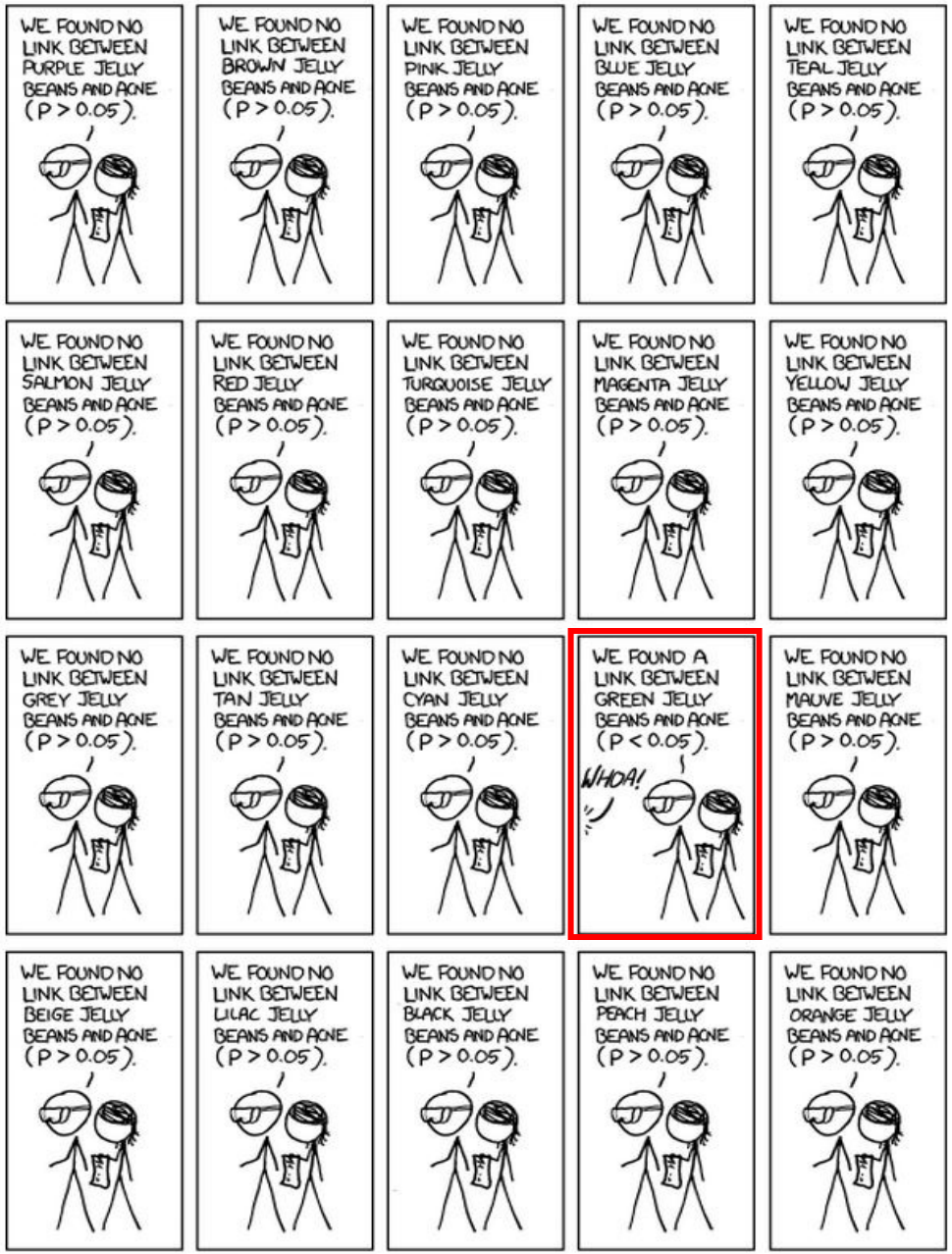
Tall people

- What is the probability that the person sitting next to you on the bus going back home is >1.90m tall?
- What is the probability of SOME person in the bus being >1.90m tall?
- What is the probability that someone on the IDC campus is >1.90m tall?

$$P(X > 1.9) = 0.01$$

IDC HERZLIYA

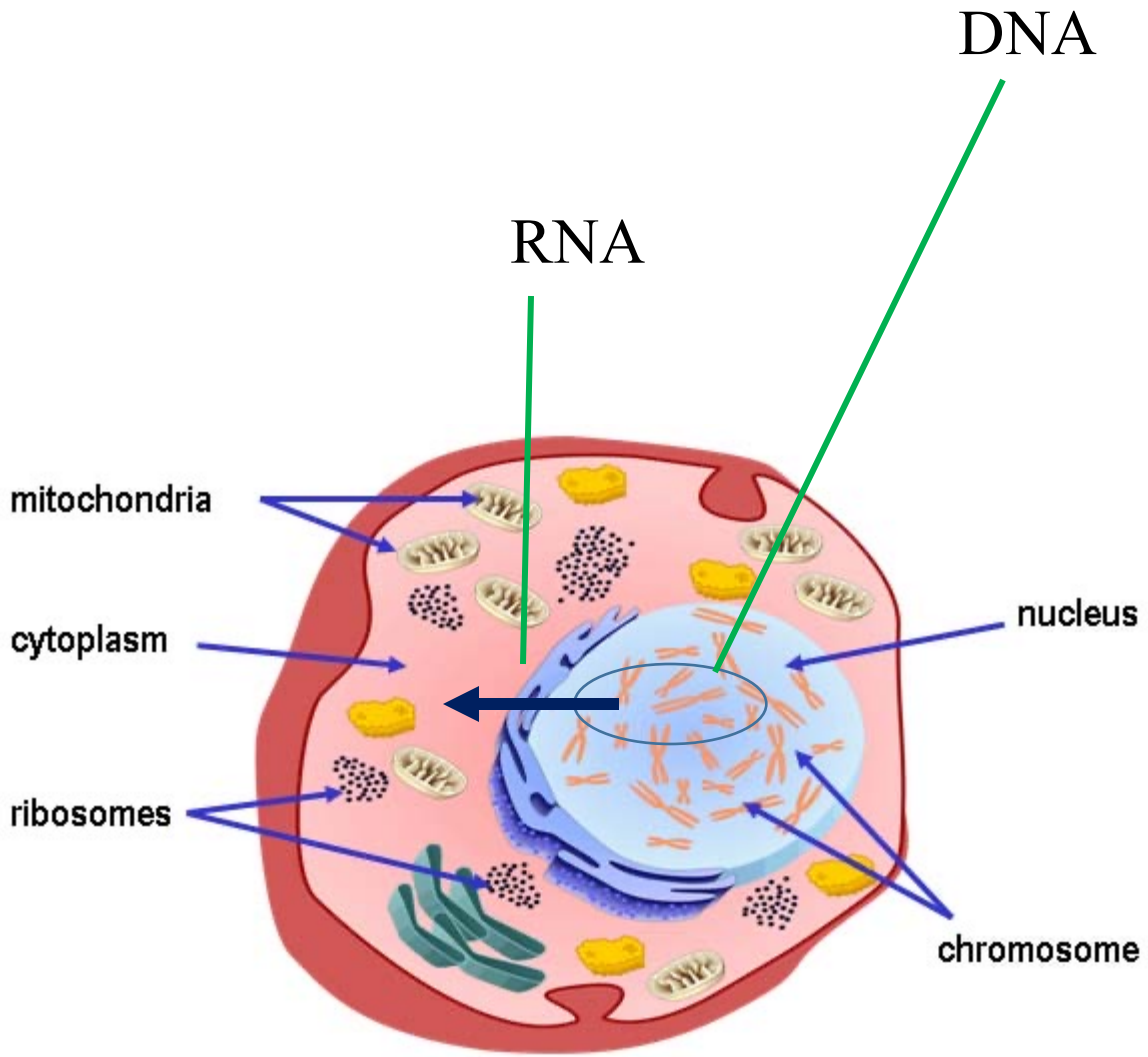
Multiple testing



IDC HERZLIYA

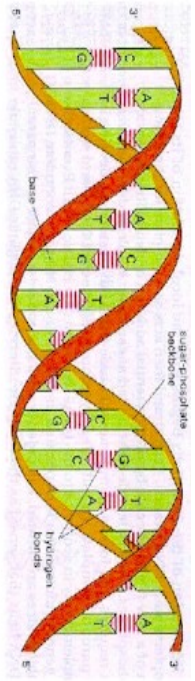
Zohar Yakhini

A living cell



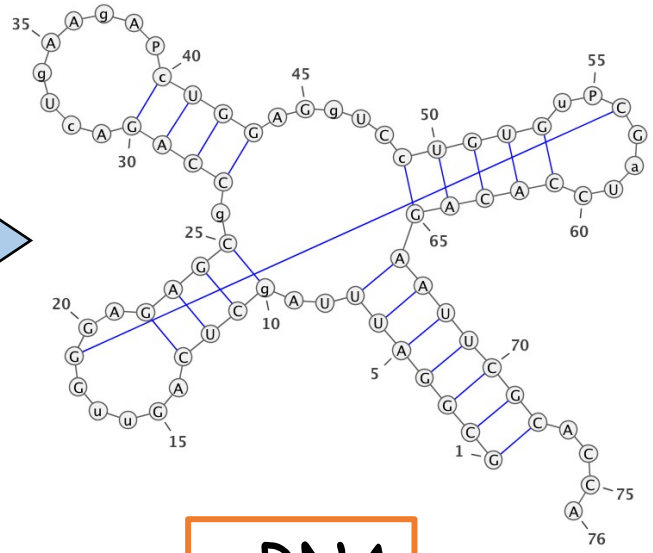
IDC HERZLIYA

Central Dogma, information flow in living cells



Gene
(DNA, nuclear)

Transcription



mRNA

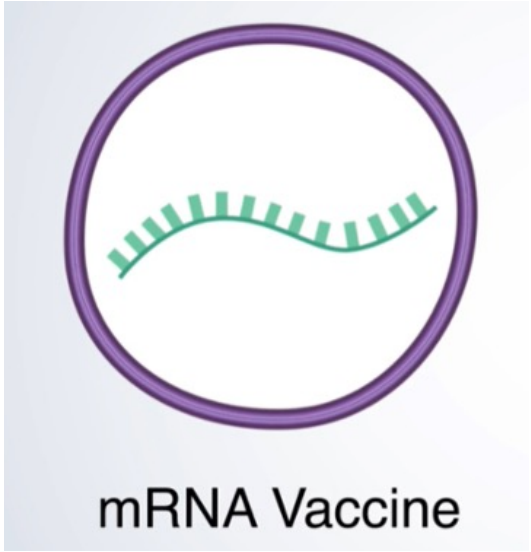
Translation

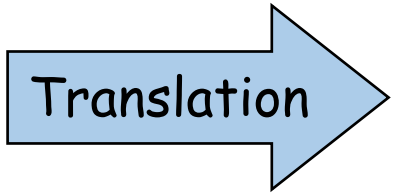
Protein

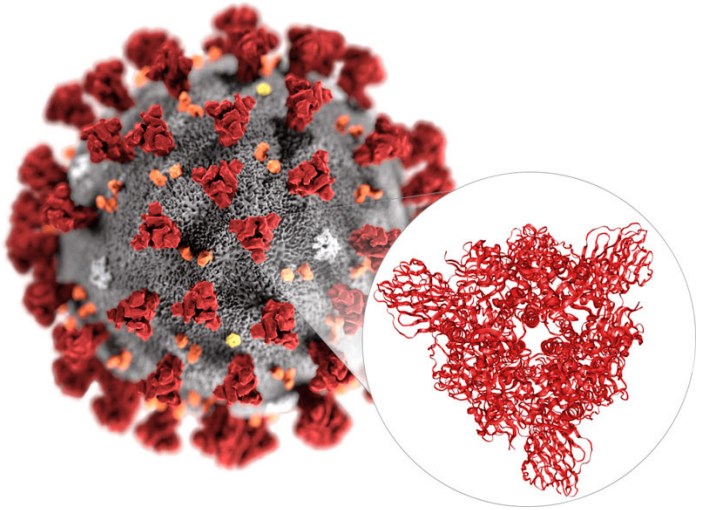
Cells express **different** subsets of the (organismal) genes in different tissues and under different conditions

IDC HERZLIYA

Pfizer COVID-19 mRNA vaccine







IDC HERZLIYA

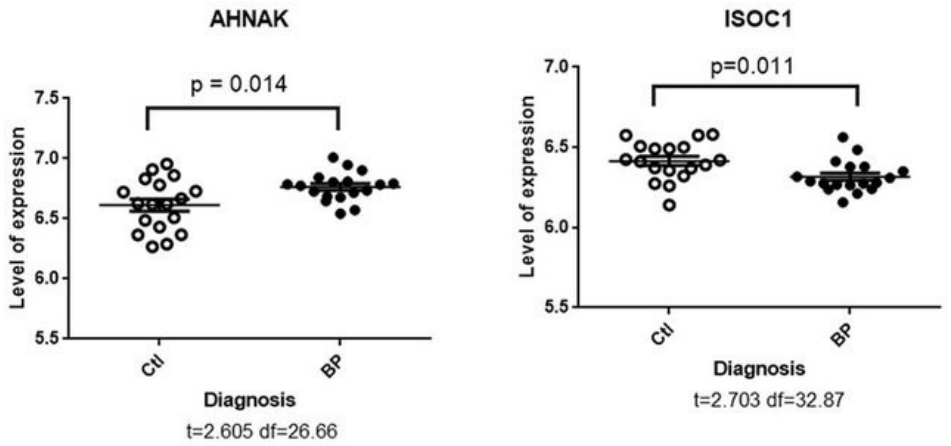
BioNTech CoVid19 vaccine is a 4,284nt mRNA which mimics coding of the pathogen's spike protein.
The entire virus is coded by ~30Knt RNA genome

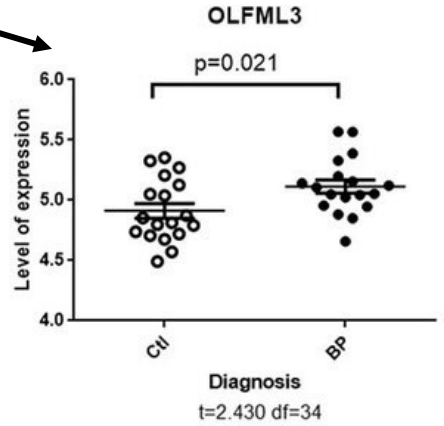
Zohar Yakhini

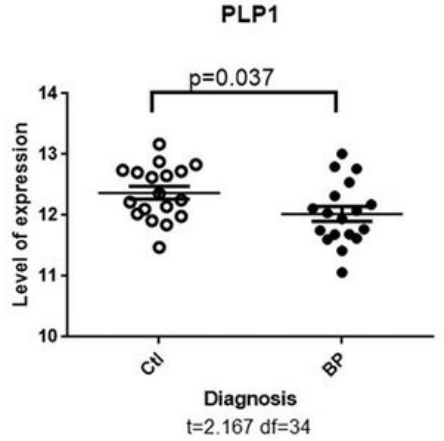
Differentially expressed genes

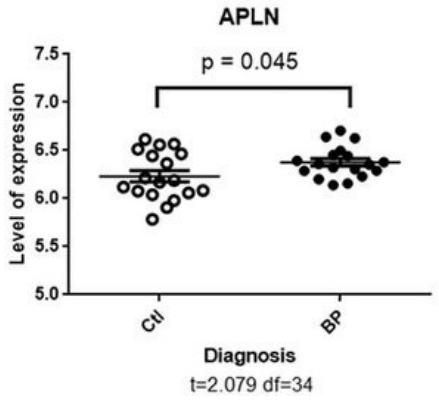
Postmortem measurement of gene expression in brain cells, Bi-Polar (BP) vs Control (Ctl)

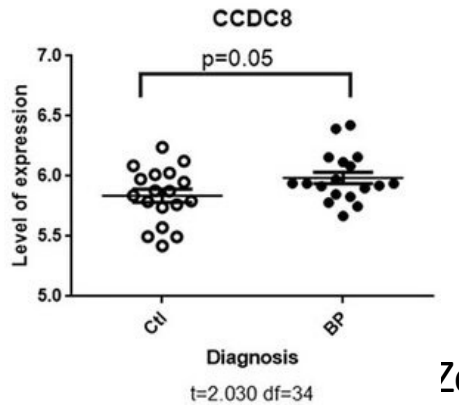
PCR measurement









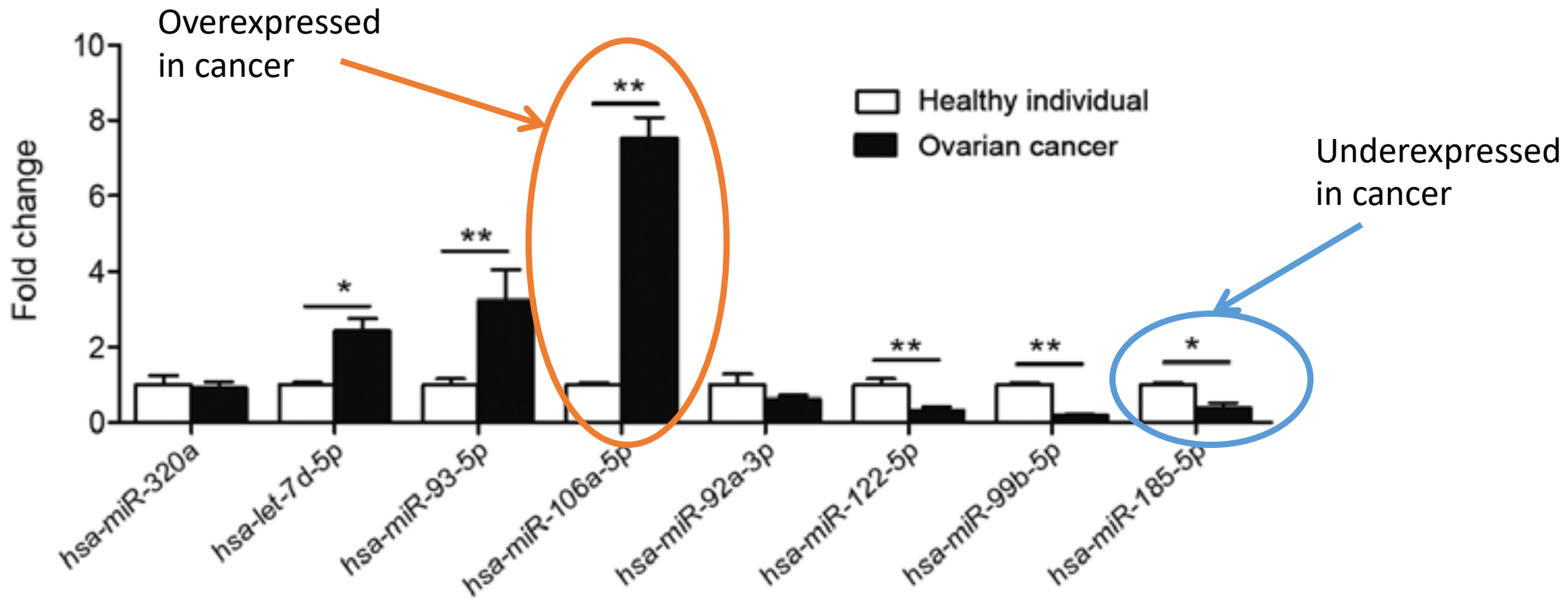


From:
Kidnapillai et al,
The use of a gene expression signature and connectivity map to repurpose drugs for bipolar disorder,
Ken Walder's Lab,
Biological Psychiatry 2018

IDC HERZLIYA

Zohar Yakhini

Differentially expressed genes



Expression levels of plasma exosomal miRNA measured in patients with ovarian cancer (n=30) and healthy women (n=30) using RT-qPCR.
*P<0.05 and **P<0.01.

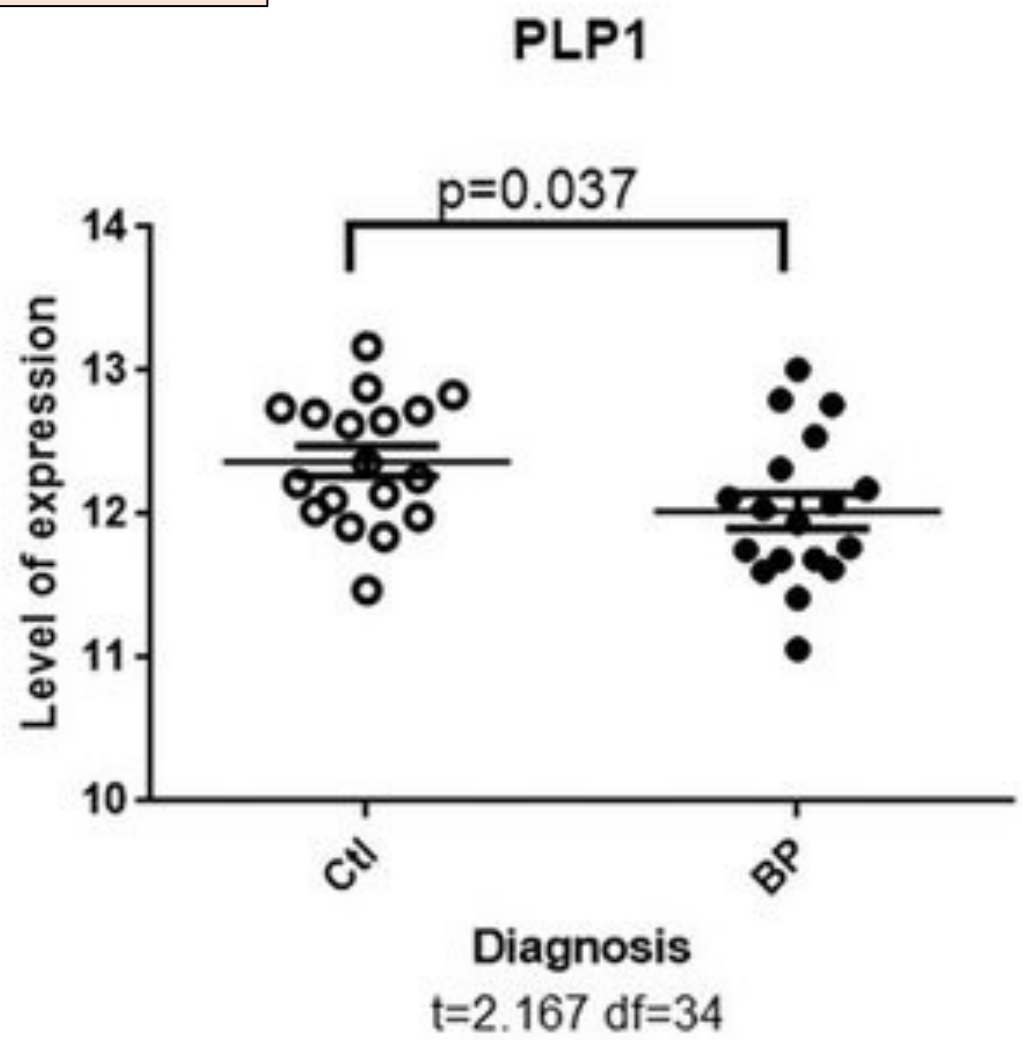
Zhang et al,
Oncology Letters 2019

IDC HERZLIYA

Zohar Yakhini

Differentially expressed genes

- t-Test
- WRS
- Other methods…

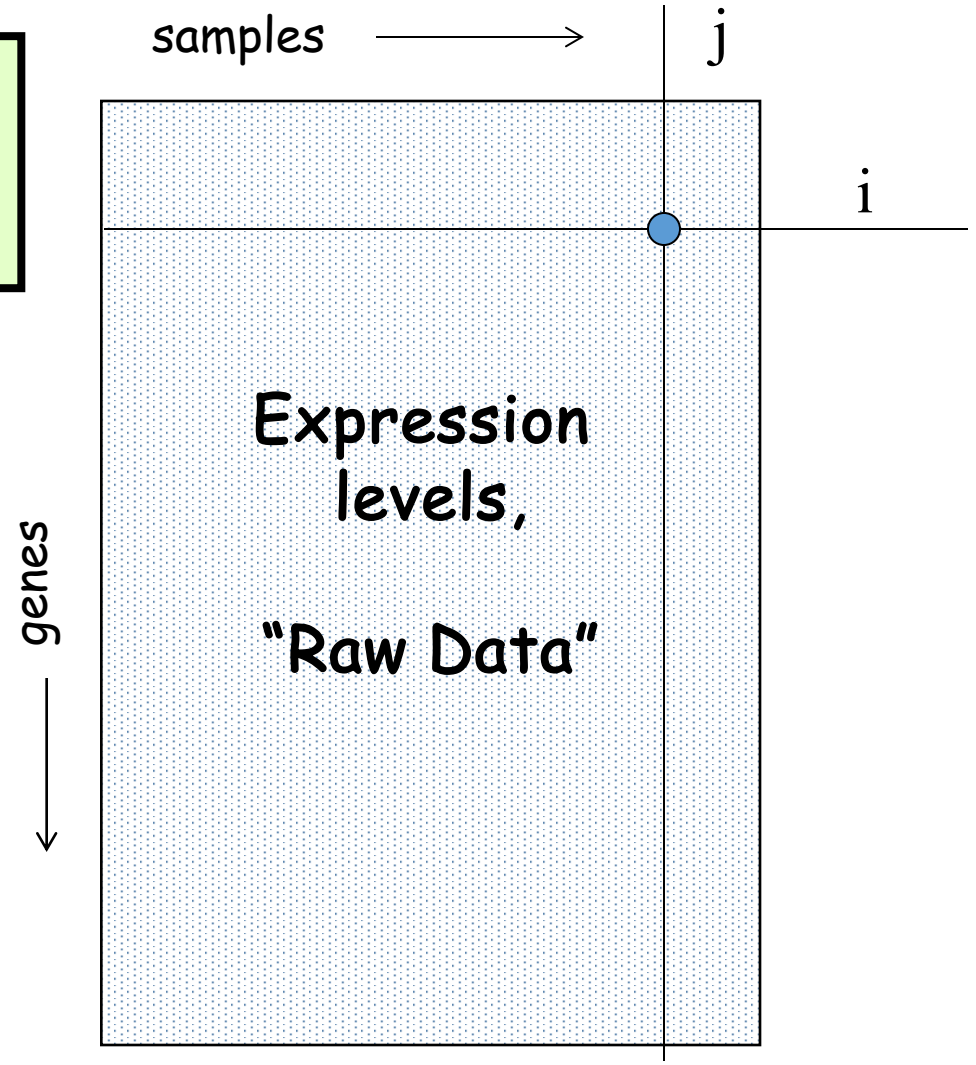


IDC HERZLIYA

Zohar Yakhini

Gene Expression Data, Matrix Representation

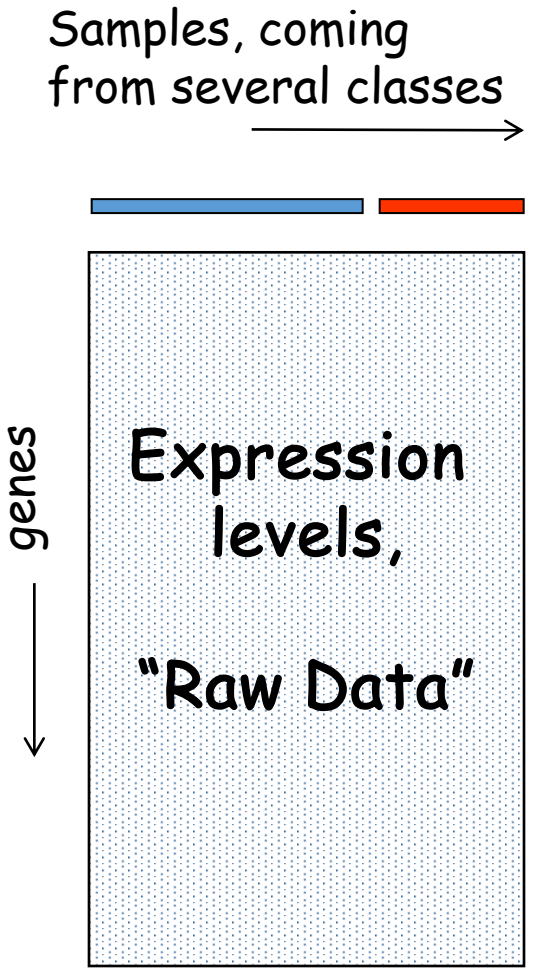
$E(i,j)$ represents the expression level of the gene indexed by i
in the experiment/sample indexed by j.



IDC HERZLIYA

Classified Gene Expression Data, Informative Genes

- Examples:
  - Tumor vs Normal
  - Subtypes of a pathology
  - Prognosis (responders vs non-responders)
  - etc
- Informative genes: genes that sharply separate the two classes. They are *DIFFERENTIALLY EXPRESSED*
- How do we assign DE scores to genes?



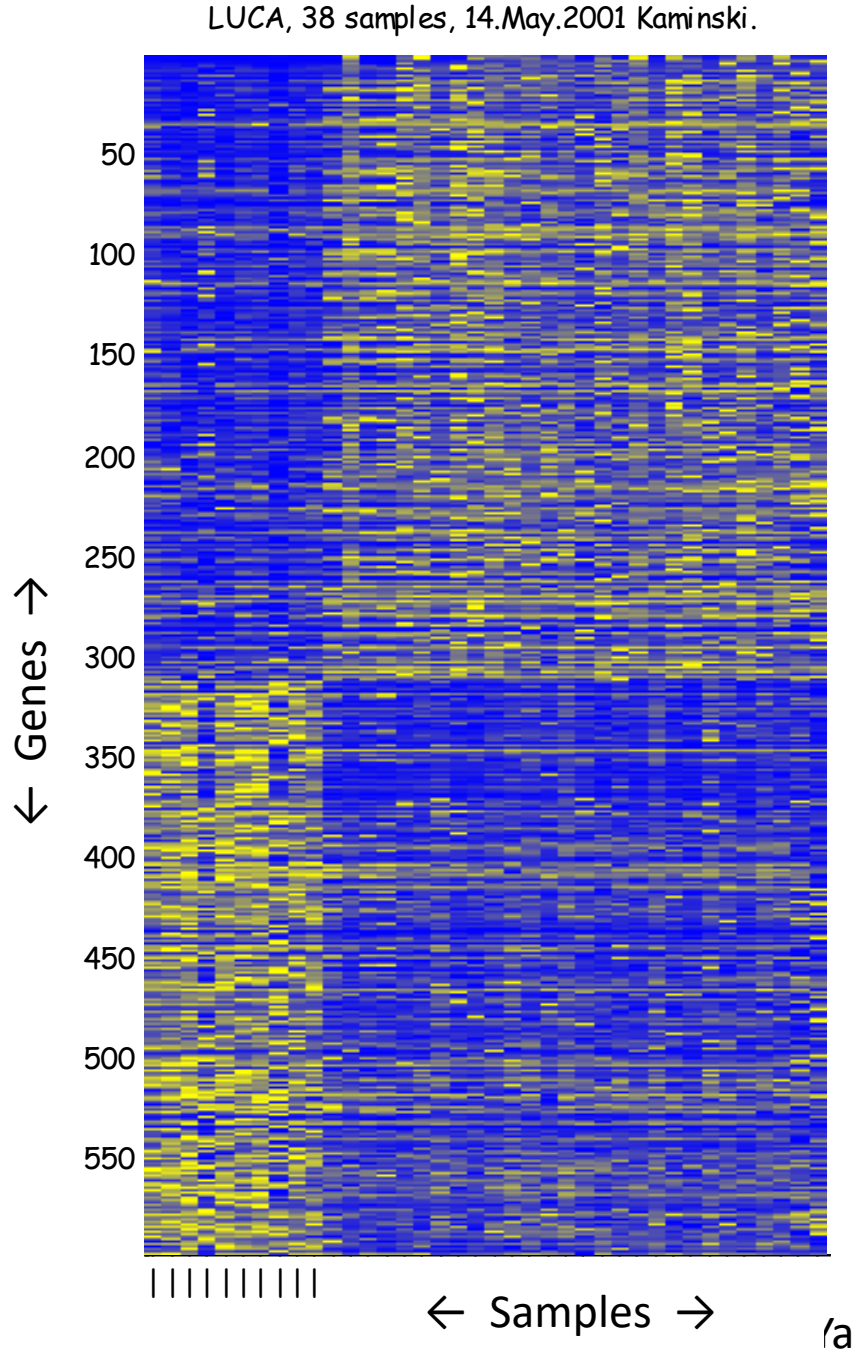
IDC HERZLIYA

Lung Cancer Informative Genes

Dehan et al, (Kaminski Lab at Sheba), Lung Cancer 2007

- 24 tumors (various types and origins)
- 10 normals (normal edges and normal lung pools)

IDC HERZLIYA



Yakhini

Wilcoxon Rank Sum test

- Compute the sum of the ranks of the +s:

+	+	–	–	+	+	+	–	–	–	–	–	+	–	–
a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15

- In this case:

  RS(+) = 1+2+5+6+7+13 = 34

- The null model: all +/- configurations are equiprobable
- The mean value under the null model is E(RS(+)) = ?
- The p-value of the deviation can be computed using normal approximations

Wilcoxon example

+ + + + + − − − − − − − − − −

RS(+) = ?, p-val = ?

+ + + − − − − − + − − − + − −

RS(+) = ? , p-val = ?

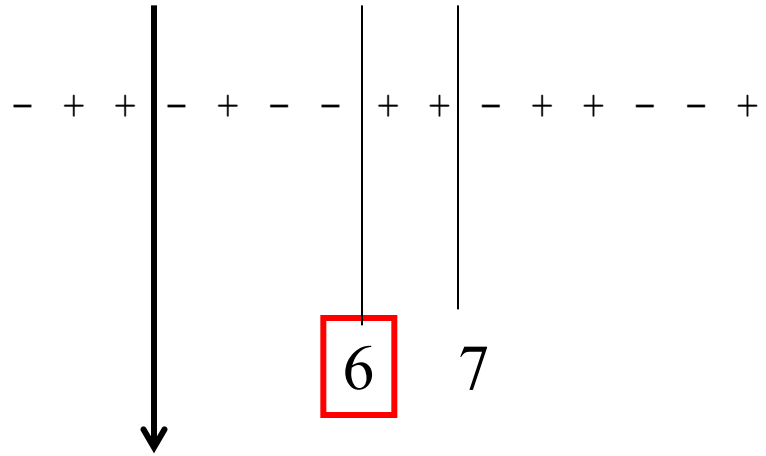
E(RS(+)) = ?

IDC HERZLIYA

Threshold Error Rate (TNoM) Score

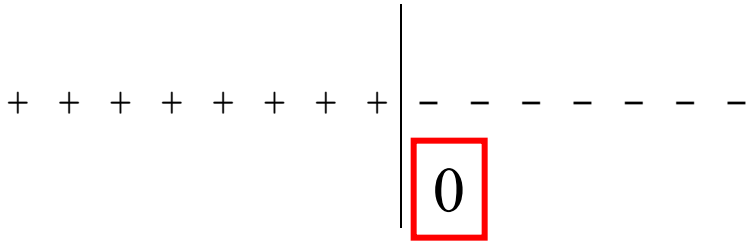
Find the threshold that best separates tumors from normals, count the number of errors committed there.

Ex 1:



# of errors = min(7,8) = 7.

Ex 2: A perfect single gene classifier gets a score of 0.



IDC HERZLIYA

Zohar Yakhini

Threshold Error Rate (TNoM) Score

Consider the ordered vectors of labels below:

v1 = 0 0 0 0 1 0 0 0 1 1 1 0

v2 = 0 0 1 0 0 1 0 0 0 1 1 0

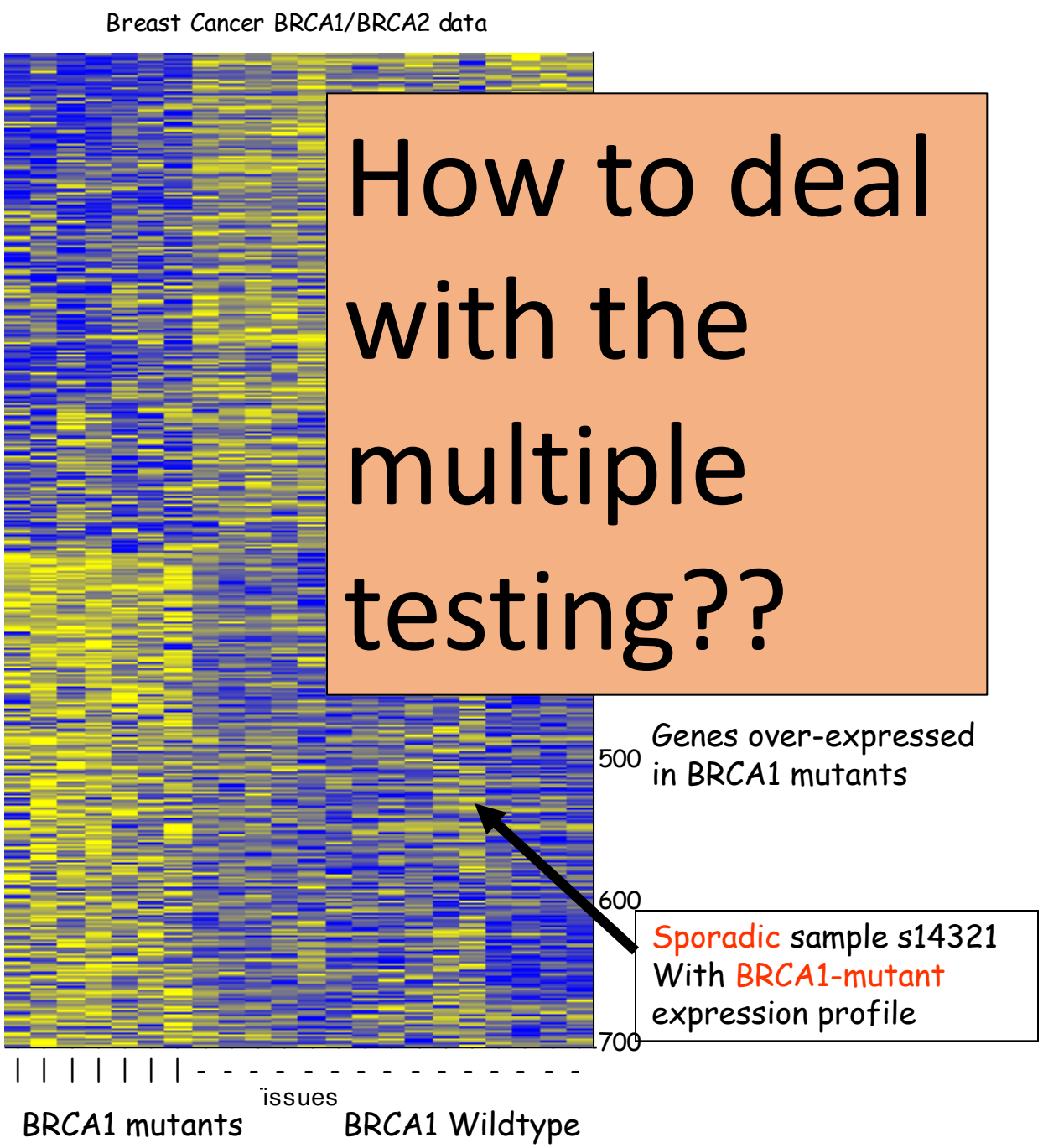
Compute the TNoM score for these.

IDC HERZLIYA

Zohar Yakhini

BRCA1 Differential Expression

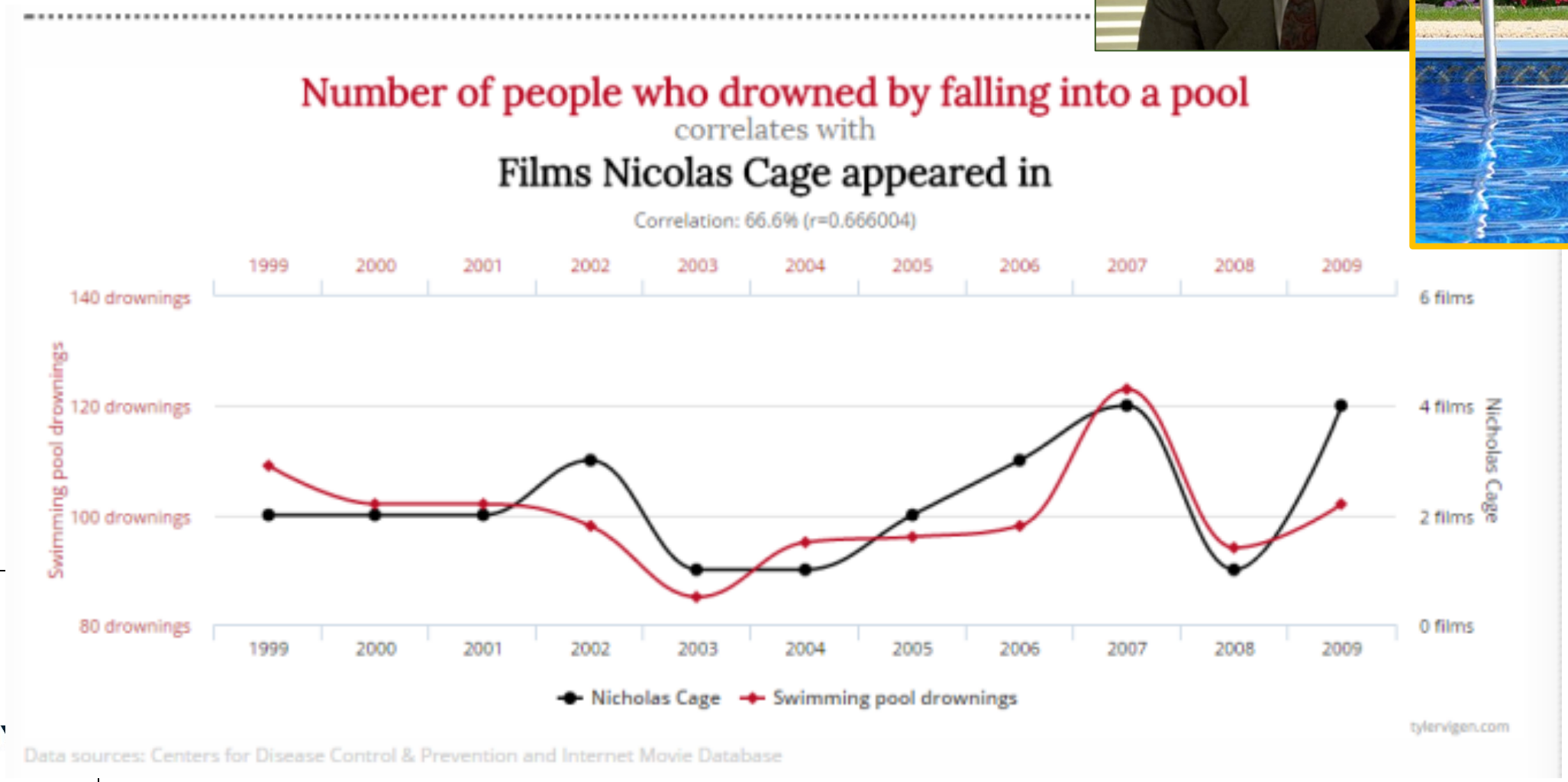
Collab with NIH; Hedenfalk et al, NEJM 2001



IDC HERZLIYA

Zohar Yakhini

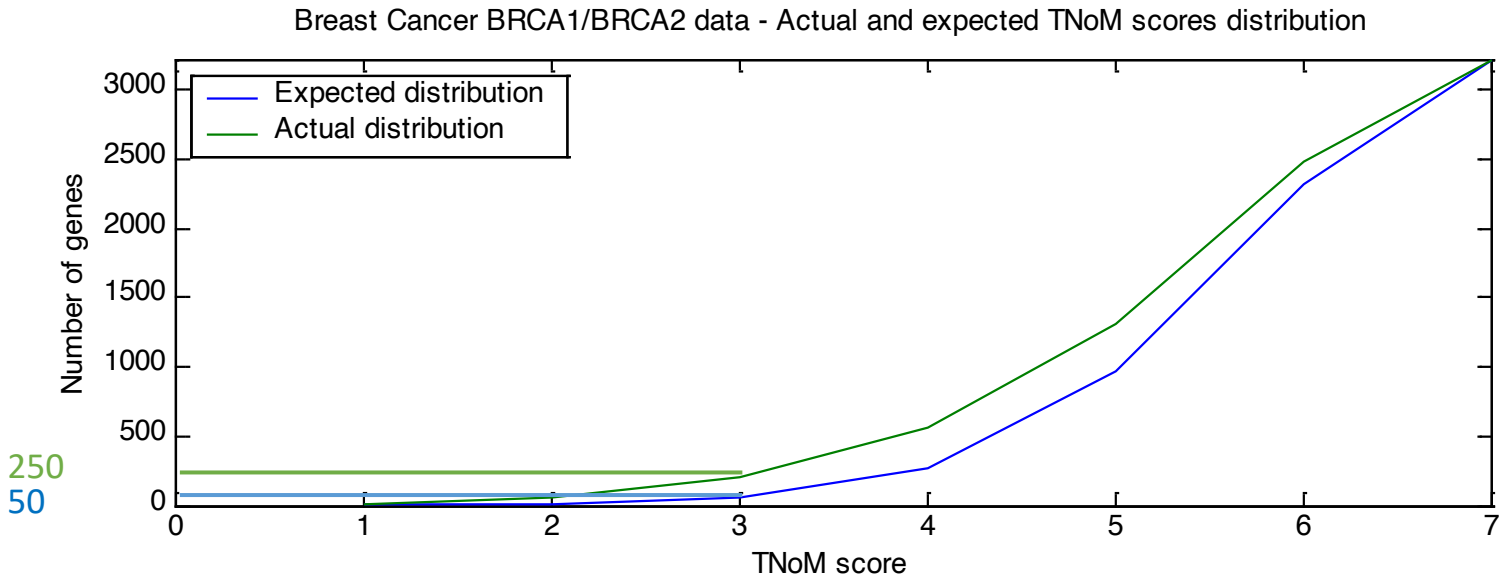
Spurious Correlations



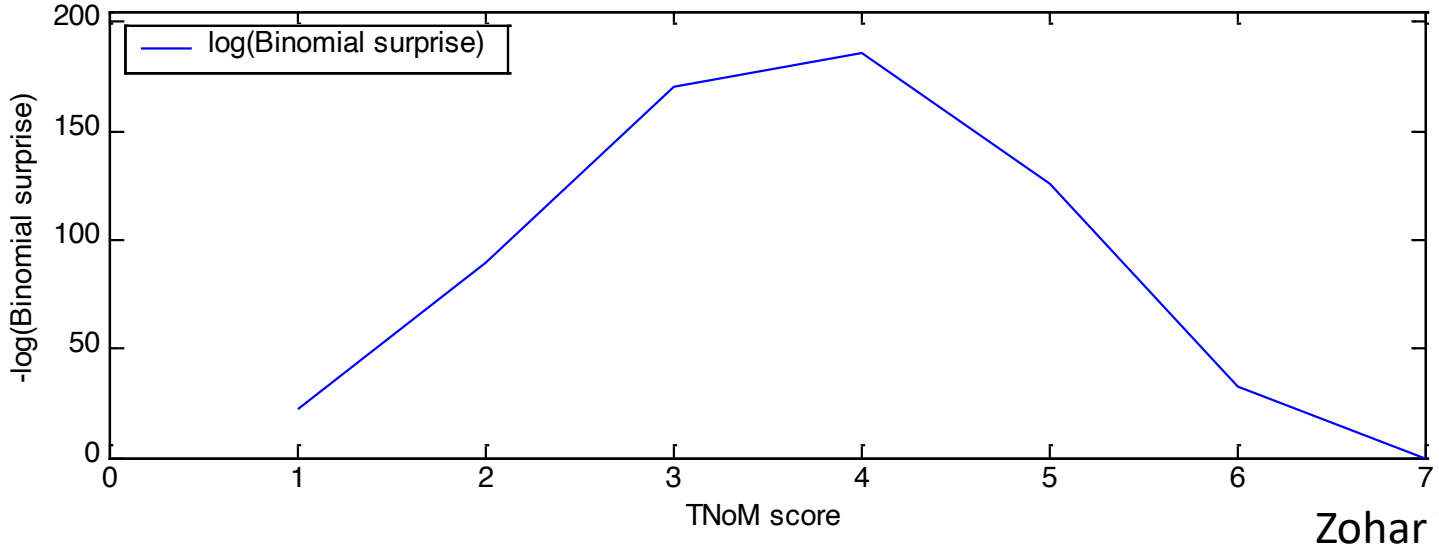
IDC HERZLI

Zohar Yakhini

BRCA1 Differential Expression: Overabundance Analysis



$$TNoM(3) \sim Binomial(pval(3), 3000)$$
$$P(TNoM(3) \geq 250) = ?$$



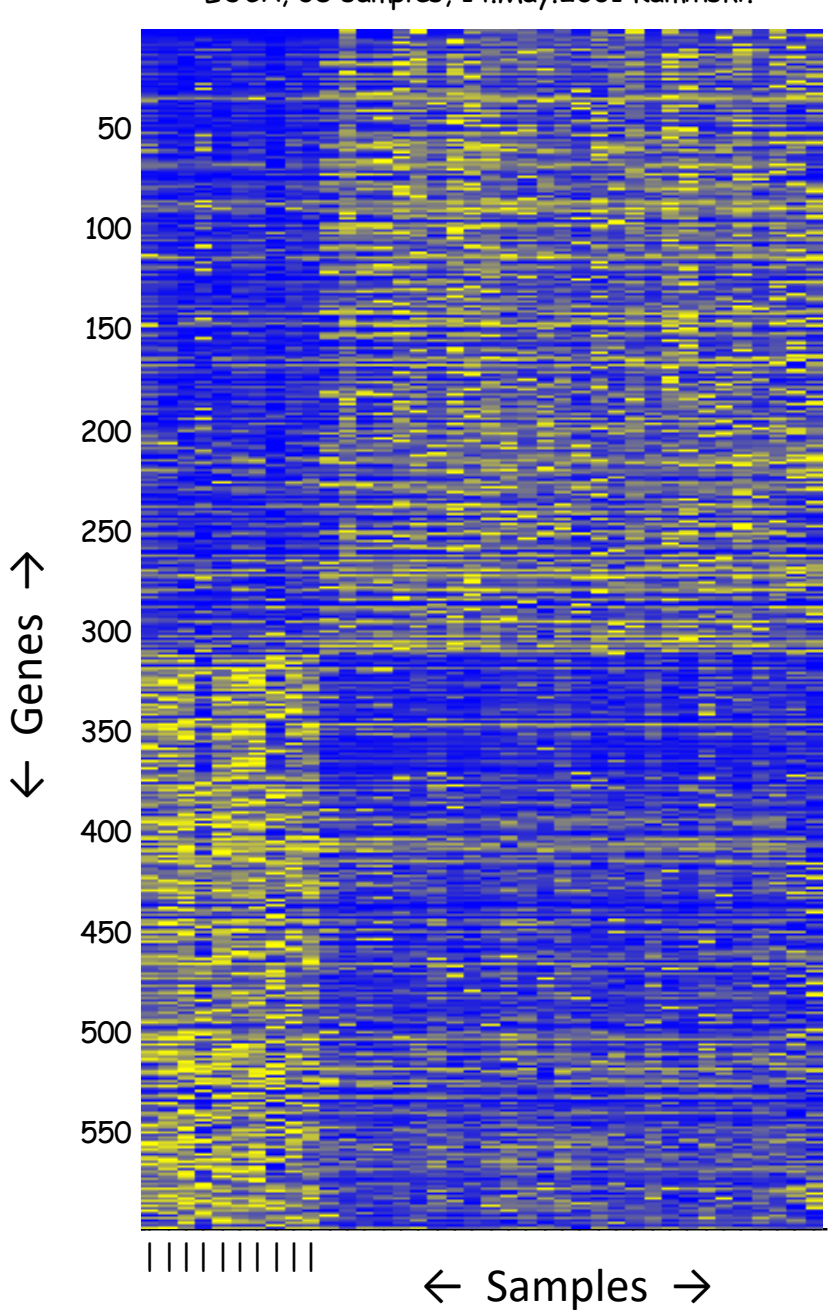
IDC HERZLIYA

Zohar Yakhini

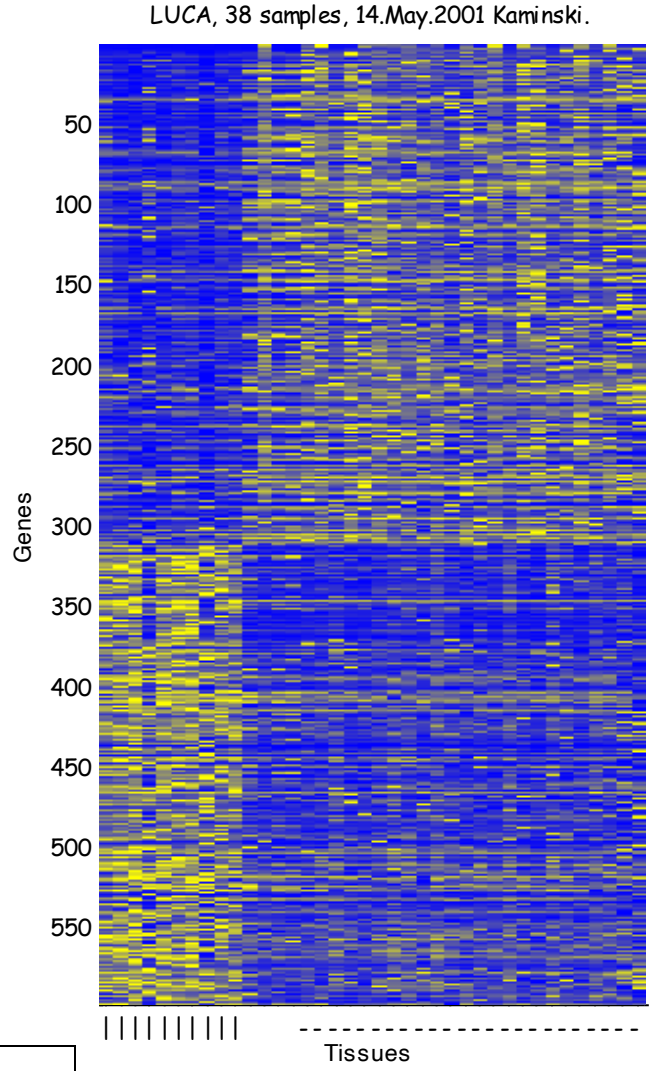
Lung Cancer Informative Genes

Dehan et al,
(Kaminski Lab at Sheba),
Lung Cancer 2007

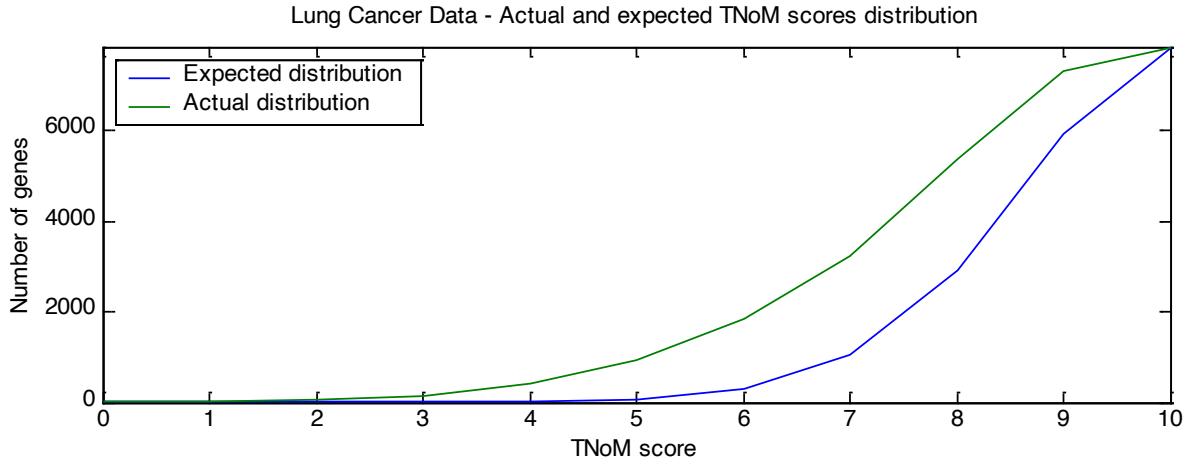
- 24 tumors (various types and origins)
- 10 normals (normal edges and normal lung pools)

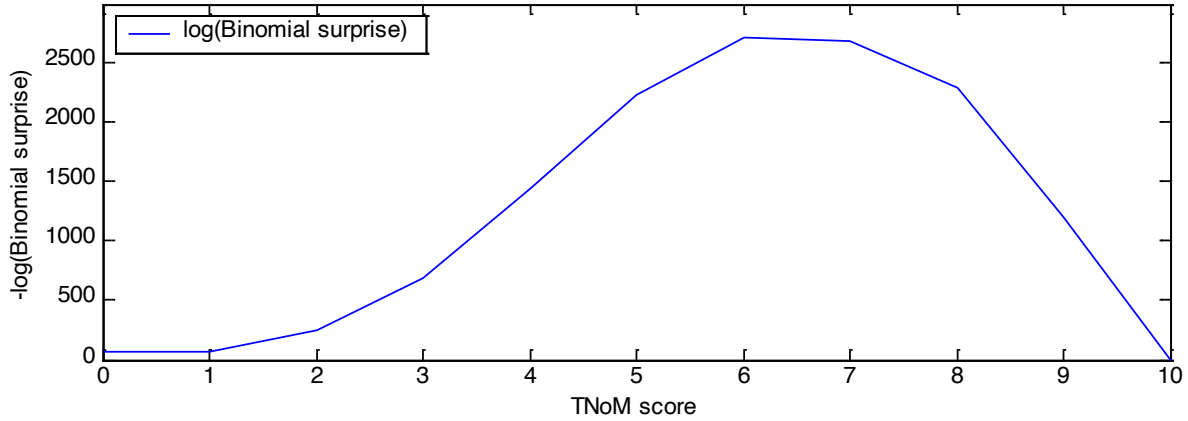


Zohar Yakhini



Lung Cancer Overabundance Analysis





IDC HERZLIYA

Zohar Yakhini

Tall people, Bonferroni

- What is the probability that the person sitting next to you on the bus this evening is >1.90m tall?
- What is the probability of SOME person in the bus being >1.90m tall?
- What is the probability that someone on the IDC campus is >1.90m tall?

- What is a possible naïve (and not very tight) correction? p/N

IDC HERZLIYA

Bonferroni - cont

What is the probability of rejecting at least one TRUE hypothesis (each test is one hypothesis testing)? (FWER – Family-Wise Error Rate)

The probability if rejecting a specific TRUE hypothesis is at most p/N.

$$\mathrm{FWER} = P\left\{\bigcup_{i=1}^{N_0}\left(p_i \le \frac{\alpha}{N}\right)\right\} \le \sum_{i=1}^{N_0}\left\{P\left(p_i \le \frac{\alpha}{N}\right)\right\} = N_0\frac{\alpha}{N} \le \alpha$$

IDC HERZLIYA

Bonferroni - cont

Assuming that we profiled 20K genes.

Is a DE p-value of $5 \cdot 10^{-6}$ significant (at 0.05)?

How many genes with a p-value at most $5 \cdot 10^{-6}$ are we expecting to see?

IDC HERZLIYA

How are p-values distributed under the null model???

- Single Experiment:

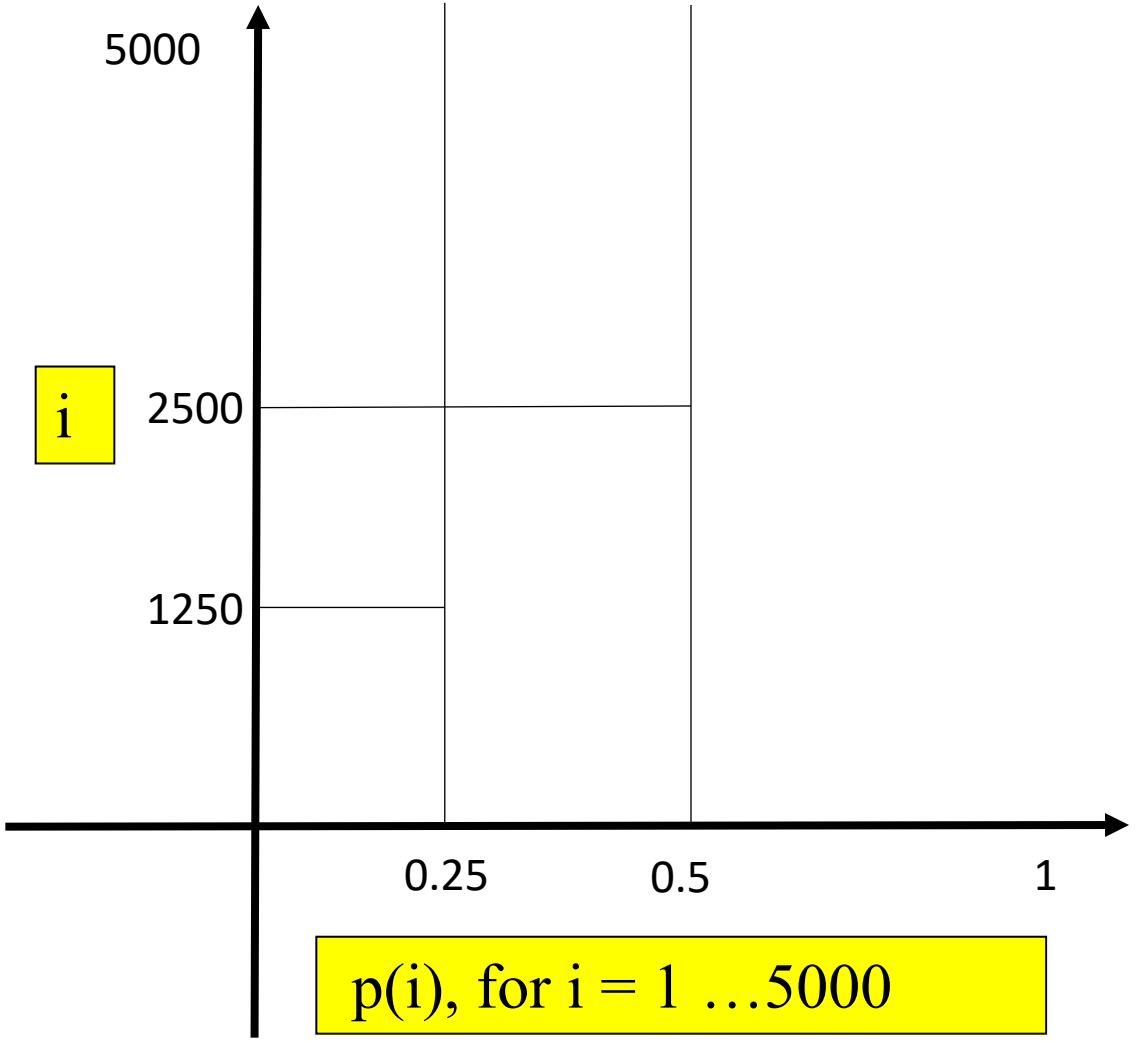
1. Toss a fair coin 100 times
2. Compute $\hat{\theta} = \frac{\#1}{100}$
3. Compute the (one sided **left**) p-value of $\hat{\theta}$ under the fair coin null model $\mathrm{P}(X \leq \#1)$

$$X \sim Binom(0.5,100)$$

- Repeat 5000 times:

$$\left(\hat{\theta}(1), p(1)\right), \left(\hat{\theta}(2), p(2)\right), \ldots, \left(\hat{\theta}(5000), p(5000)\right)$$

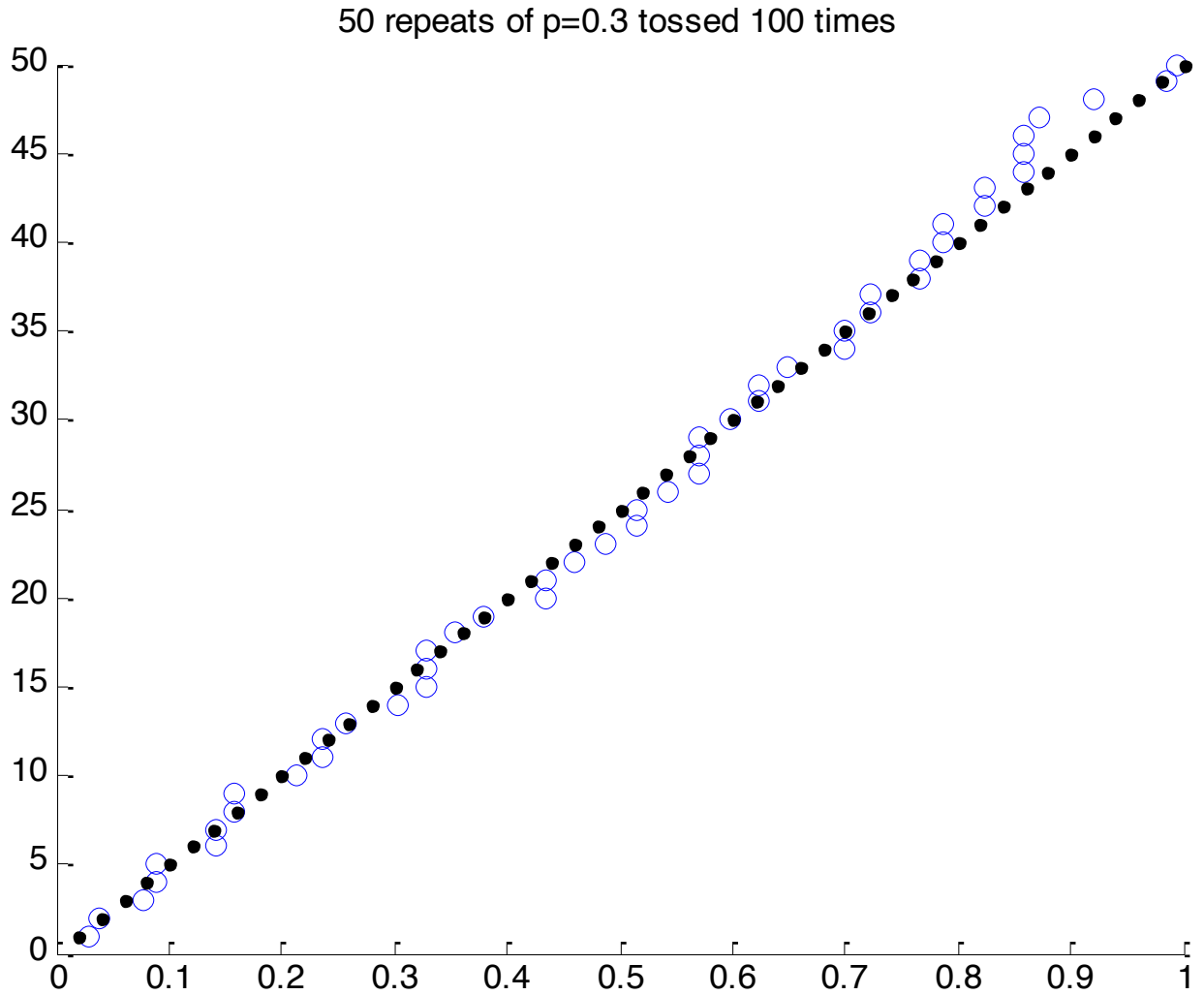
$$p(1) \leq p(2) \leq \cdots p(5000)$$



IDC HERZLIYA

Distribution of p-values under the null model

- Generating data using a coin with $\boldsymbol{p = 0.3}$
- under a null model of $\boldsymbol{p = 0.3}$



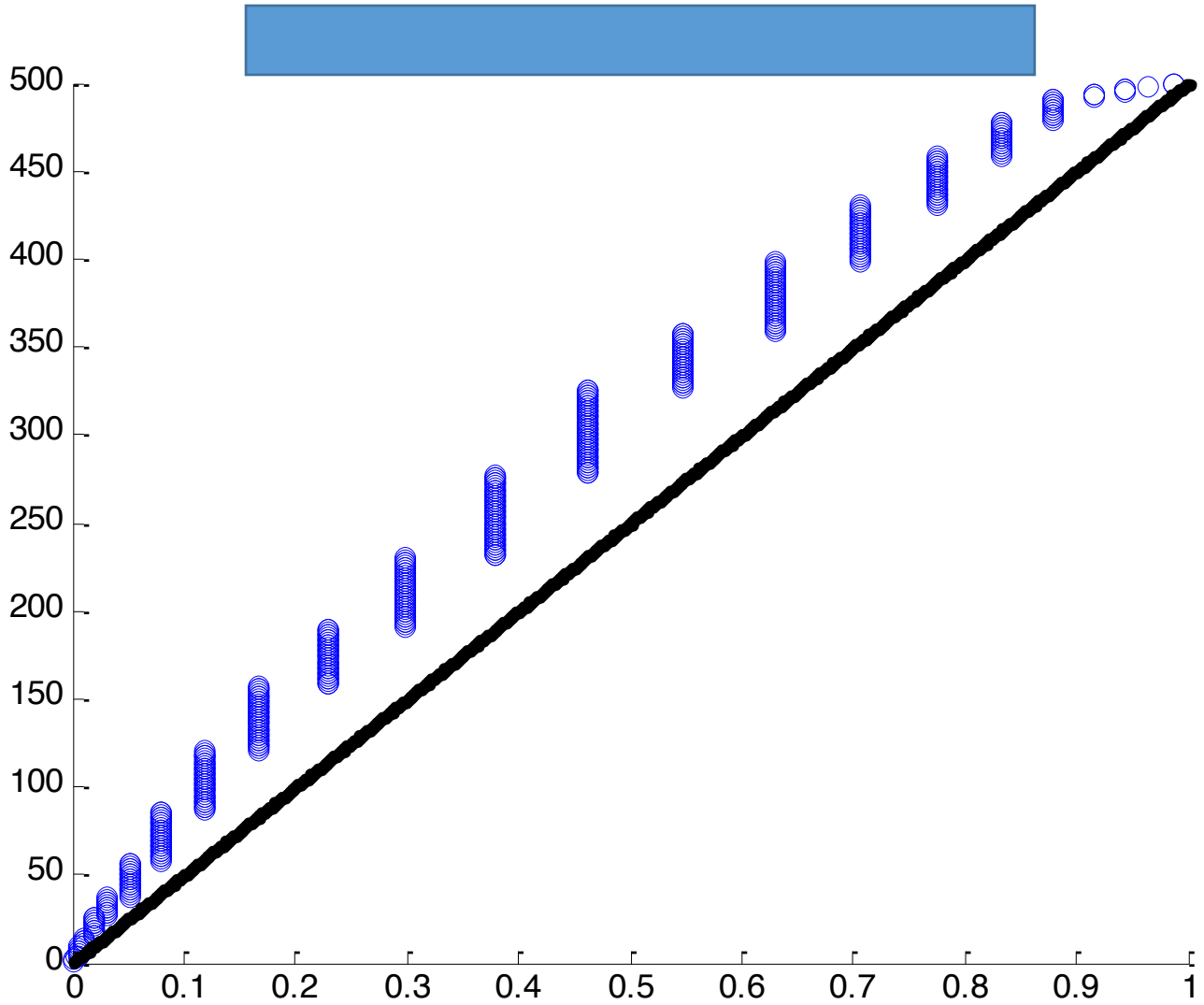
IDC HERZLIYA

Zohar Yakhini

Distribution of p-values under the null model

$$\mathrm{P}(X \leq \#1)$$

- Generating data using a coin with $\boldsymbol{p = 0.3}$
- under a null model of $\boldsymbol{p = 0.32}$



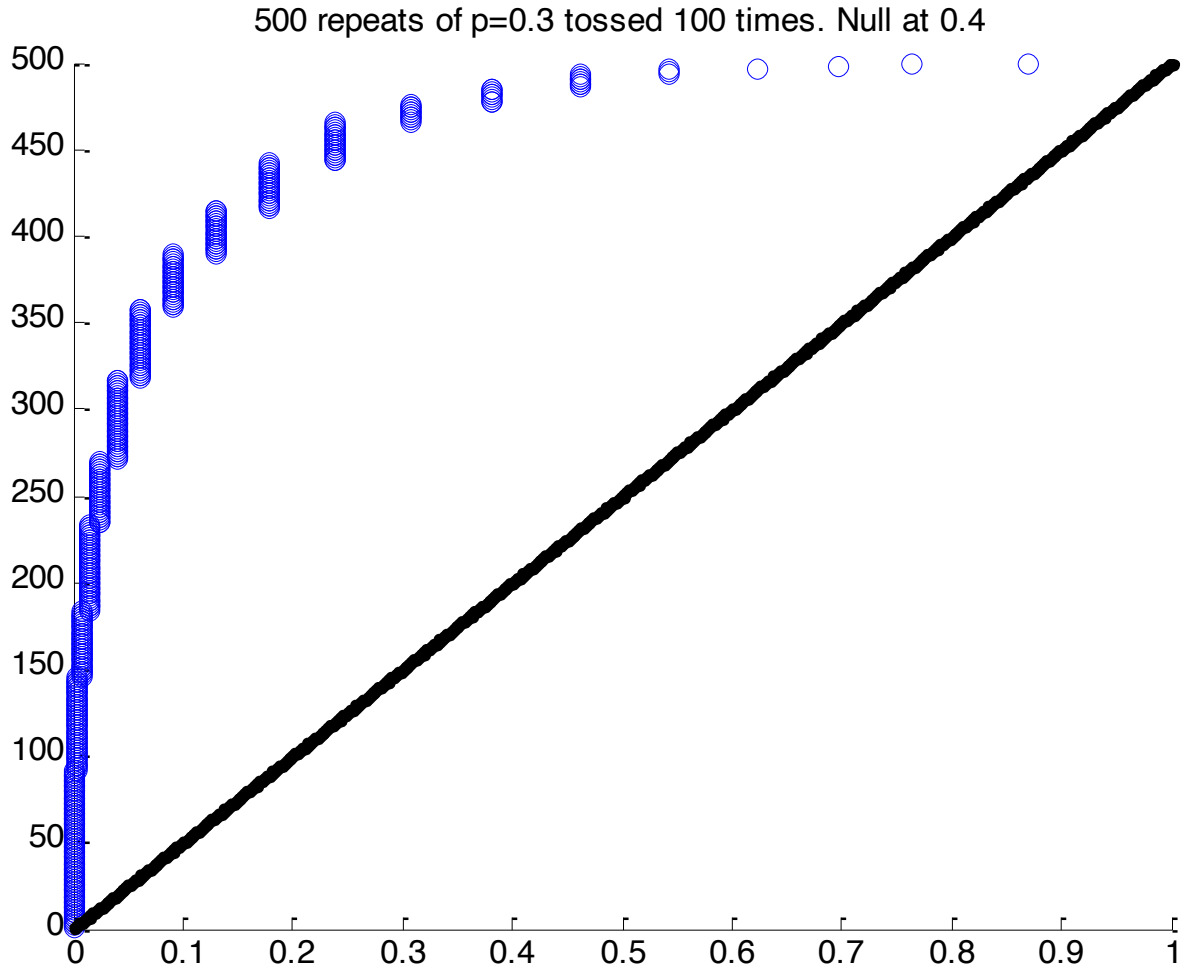
IDC HERZLIYA

Zohar Yakhini

Distribution of p-values under the null model

$$\mathrm{P}(X \leq \#1)$$

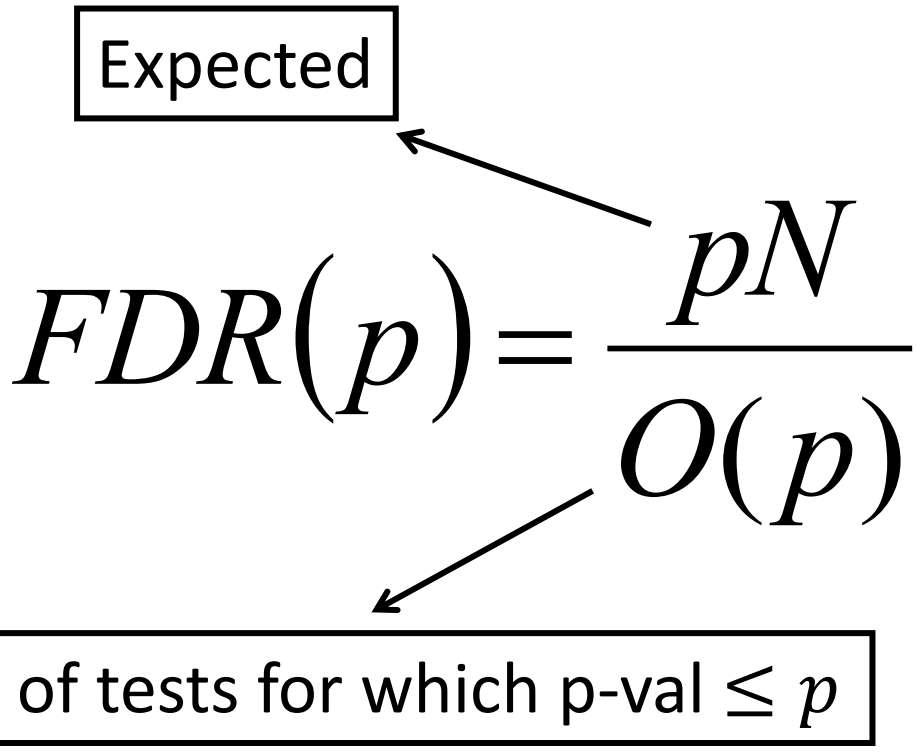
- Generating data using a coin with $\boldsymbol{p = 0.3}$
- under a null model of $\boldsymbol{p = 0.4}$



IDC HERZLIYA

False Discovery Rate (FDR)

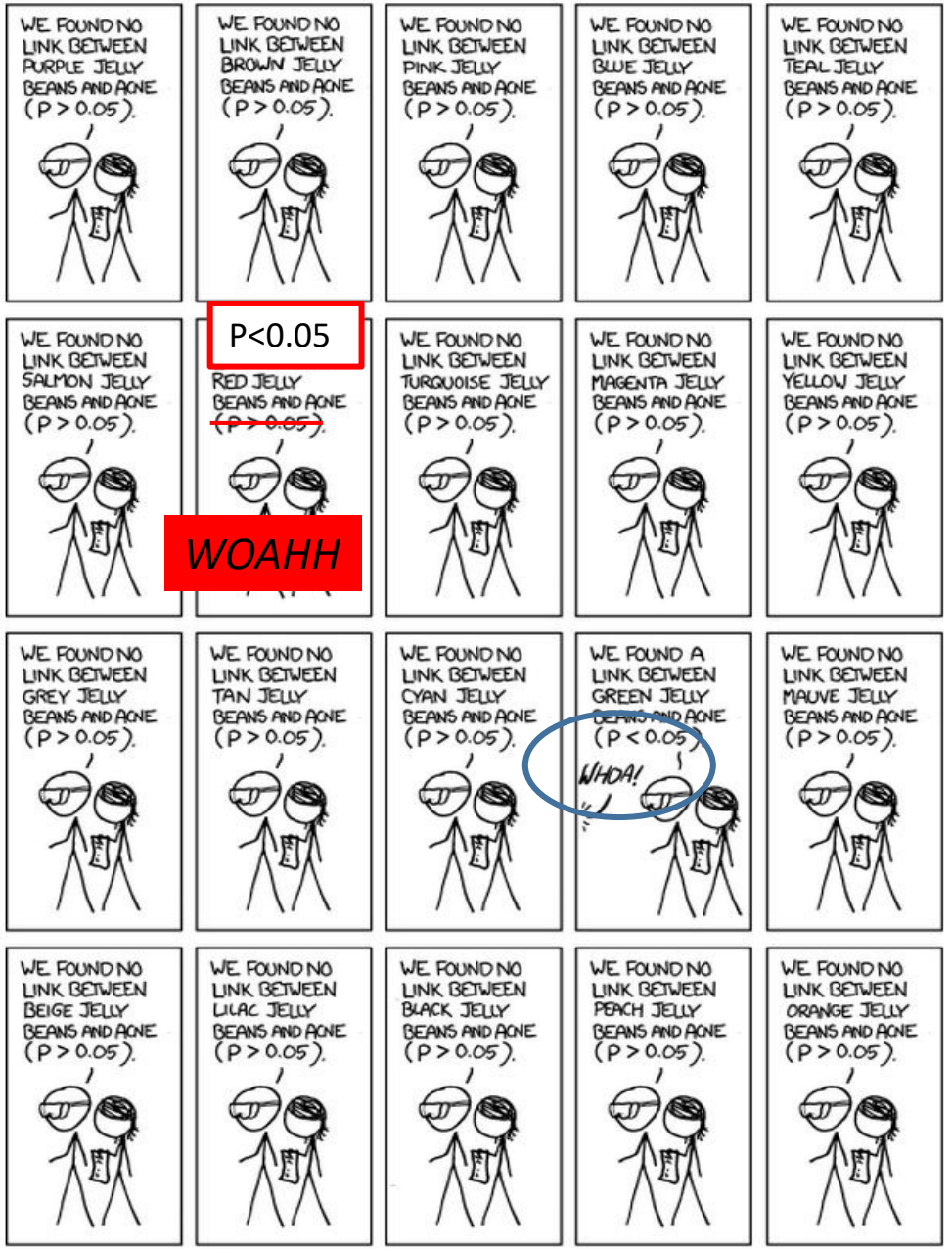
What fraction of the observed DE is expected at random (under a null-model)?



IDC HERZLIYA

More events

What if two colors were linked at 0.05? Three?

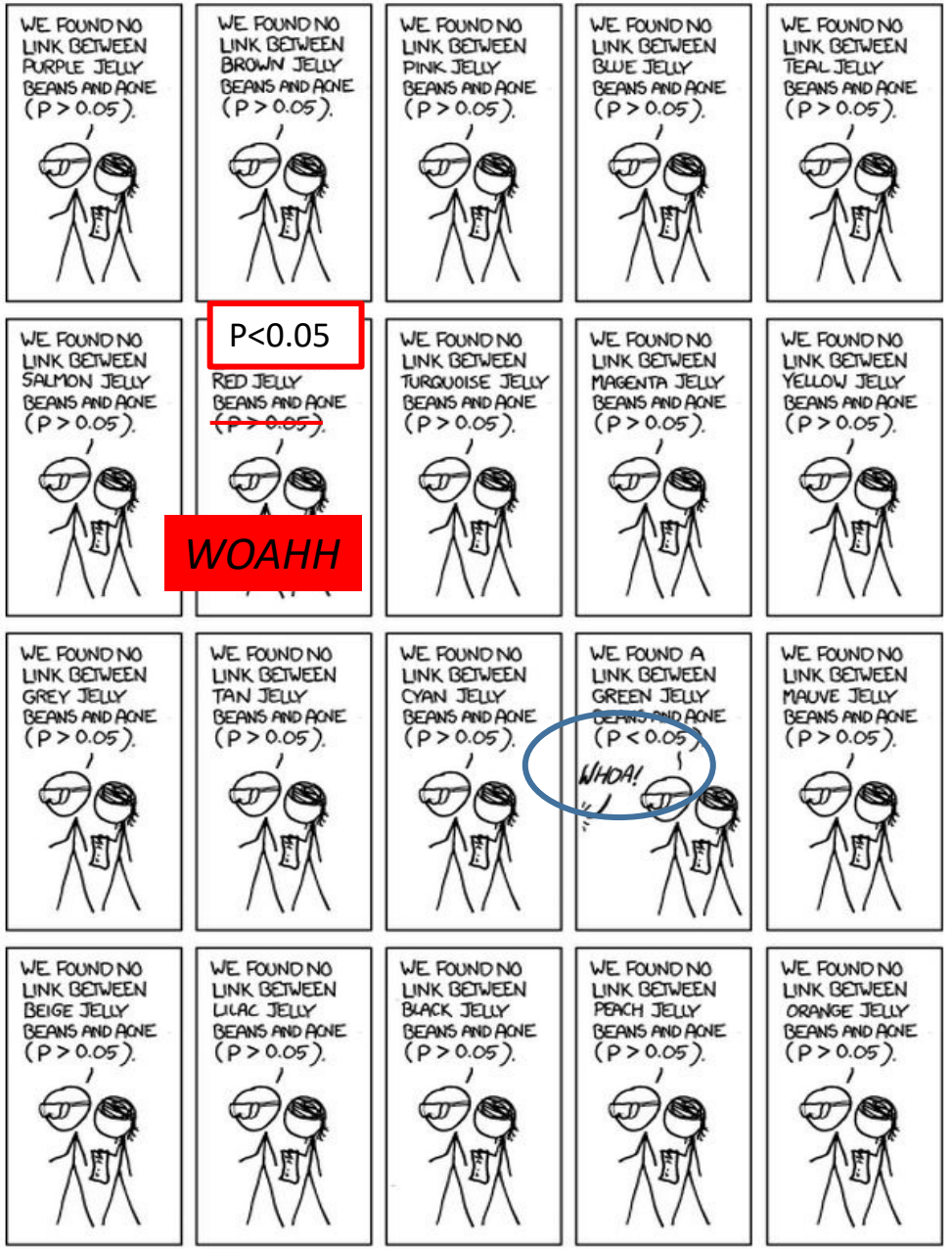


IDC HERZLIYA

Zohar Yakhini

More events

$$FDR(0.05) = ?$$



IDC HERZLIYA

Zohar Yakhini

FDR at any i

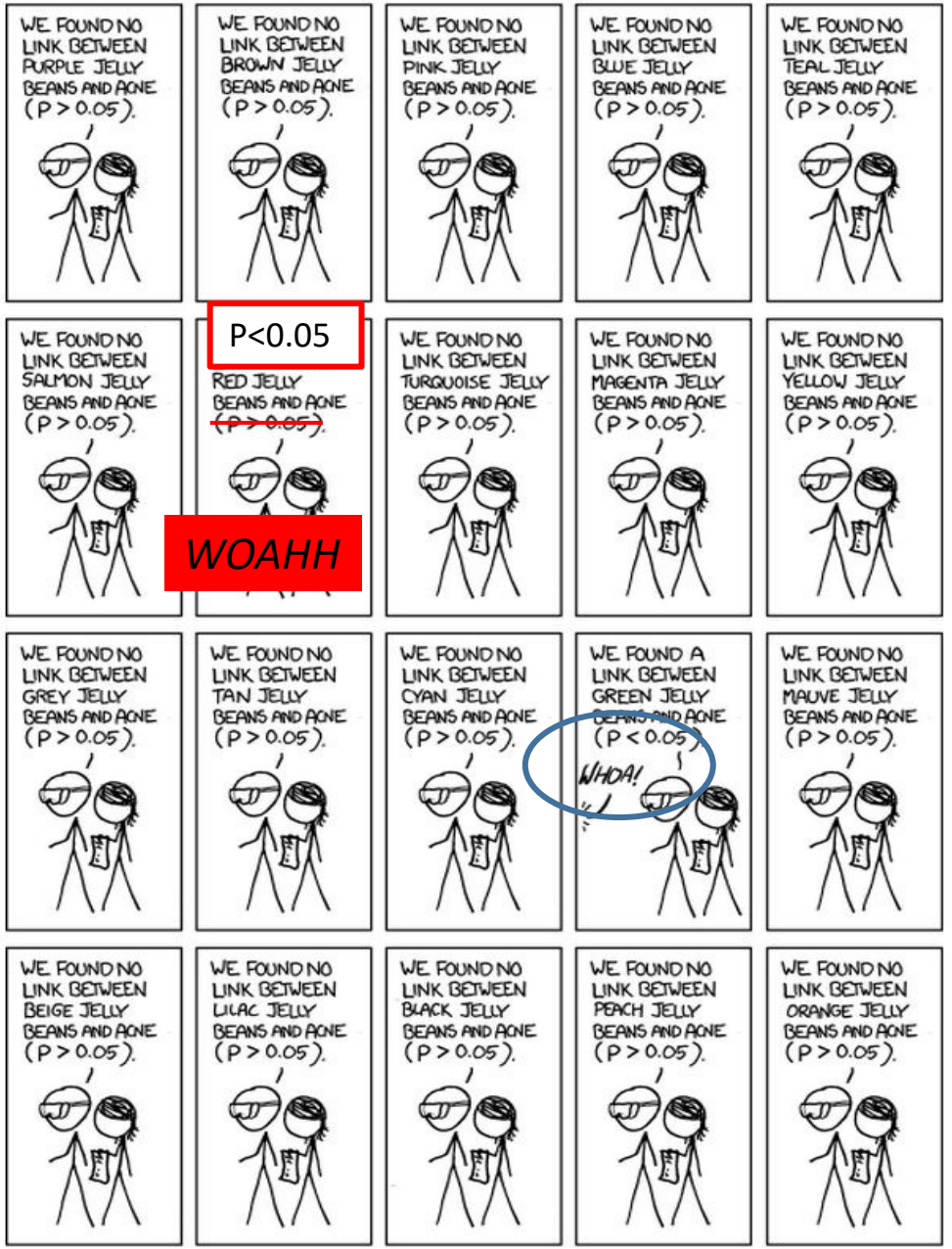
$$FDR(p) = \frac{pN}{O(p)}$$

$$FDR(i) = \frac{p(i)N}{i}$$

IDC HERZLIYA

More events

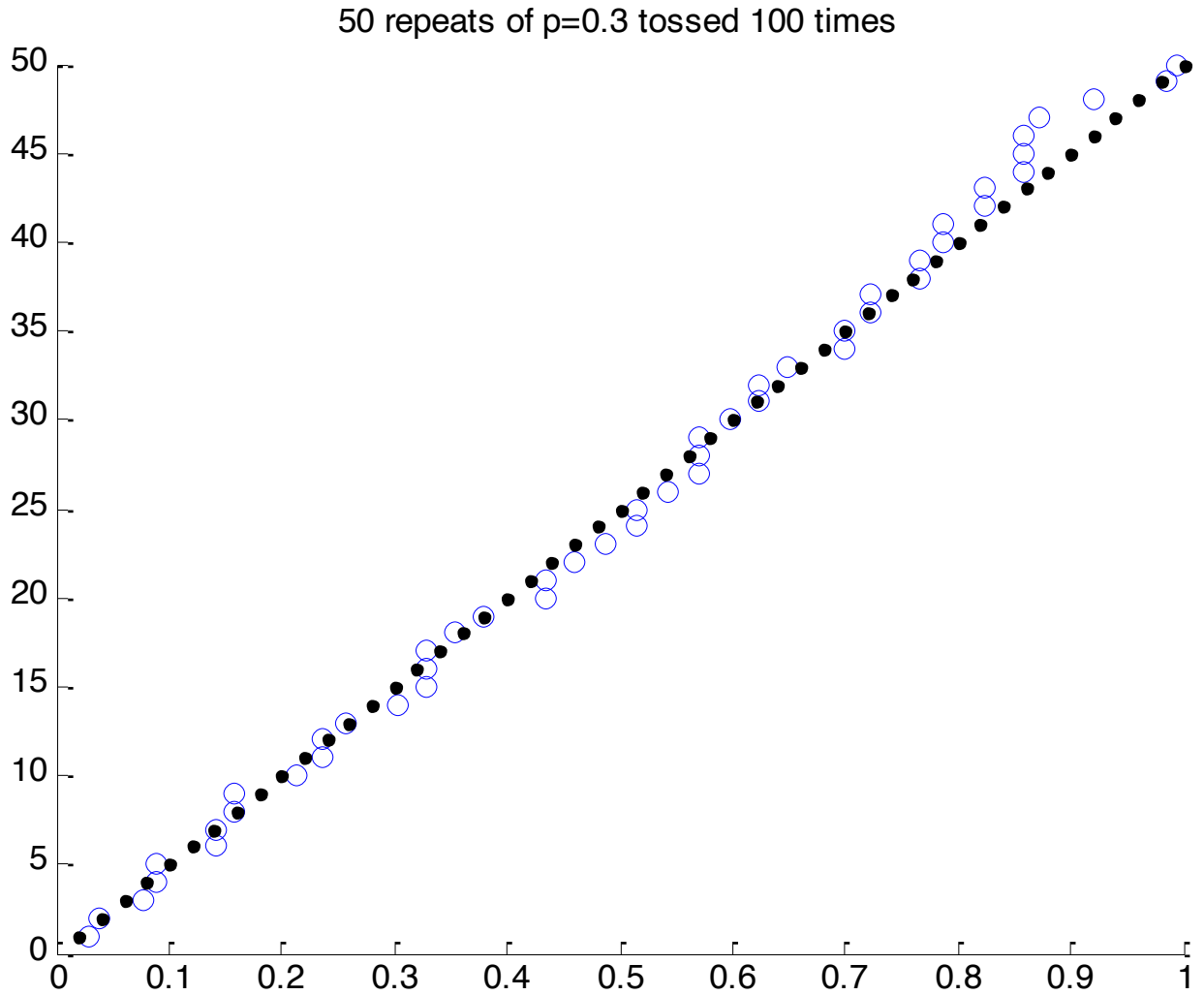
$$FDR(2) = ?$$



IDC HERZLIYA

Distribution of p-values under the null model

- Generating data using a coin with $\boldsymbol{p = 0.3}$
- under a null model of $\boldsymbol{p = 0.3}$

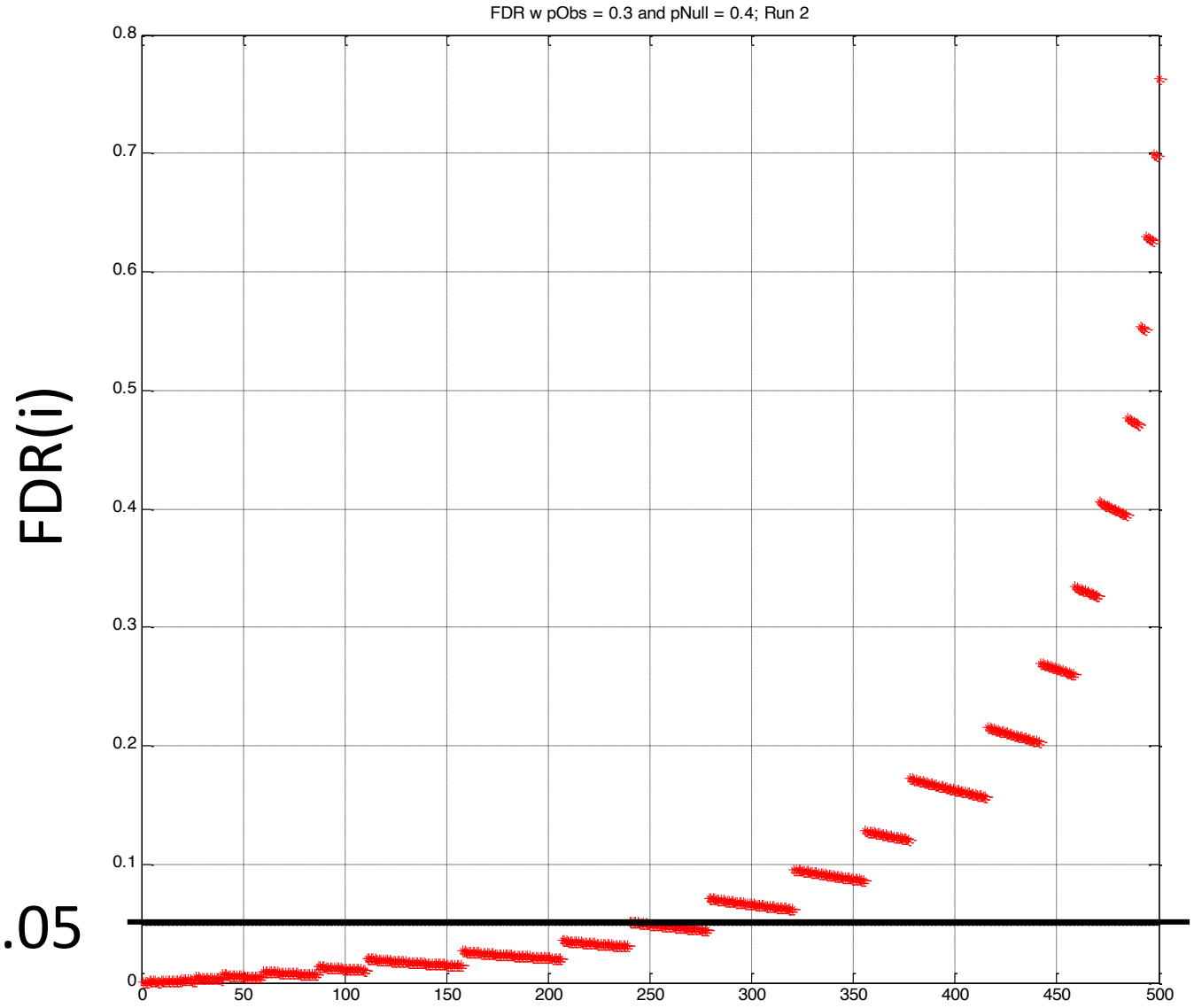


IDC HERZLIYA

Zohar Yakhini

FDR

- Generating data using a coin with $\boldsymbol{p = 0.3}$
- under a null model of $\boldsymbol{p = 0.4}$



$$FDR(i) = p(i) \cdot N / i$$

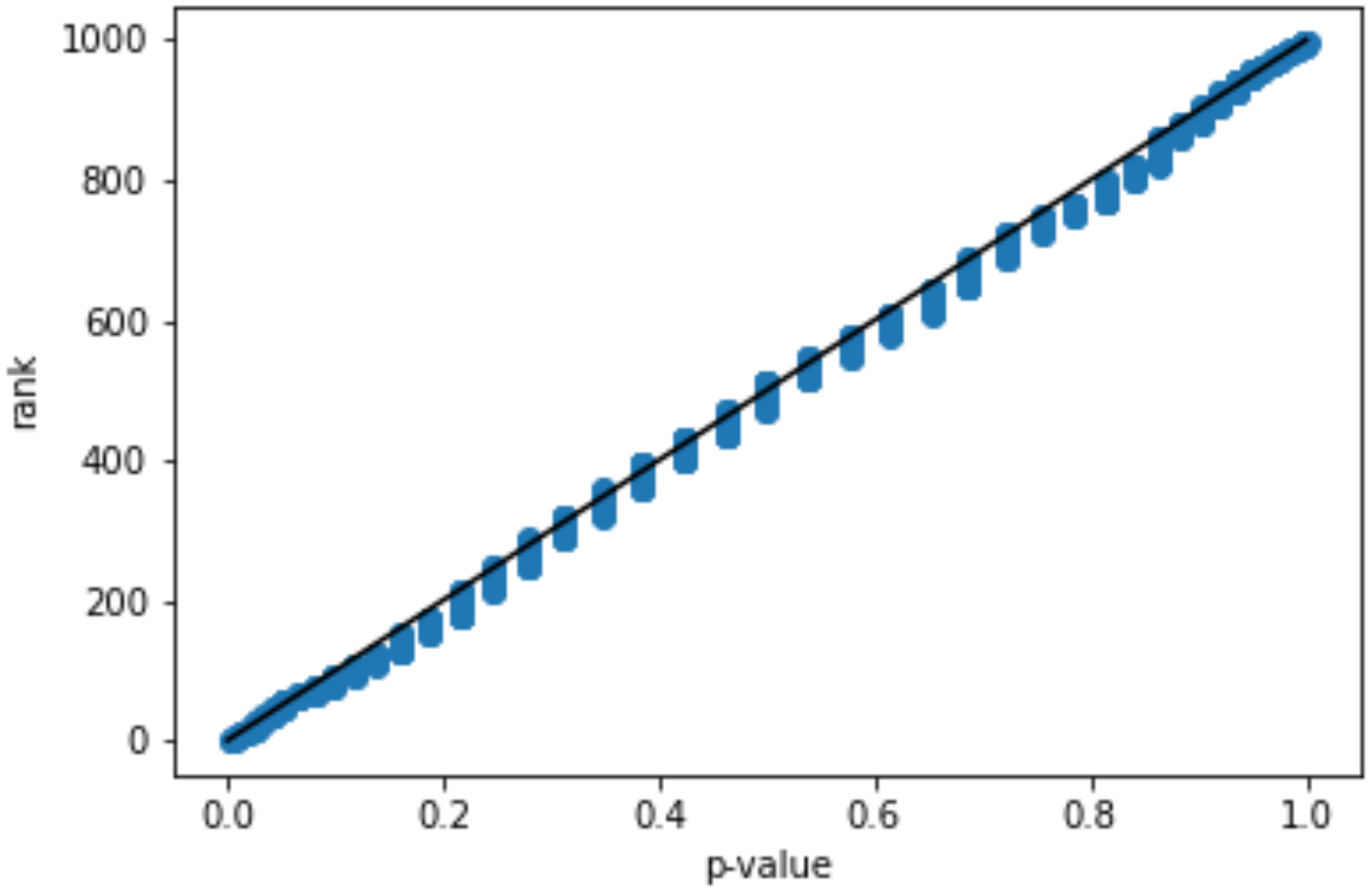
IDC HERZLIYA

A more subtle correction

$$\mathrm{RobFDR}(i) = \min_{j \geq i} \left(\frac{p(j) \cdot N}{j} \right)$$

IDC HERZLIYA

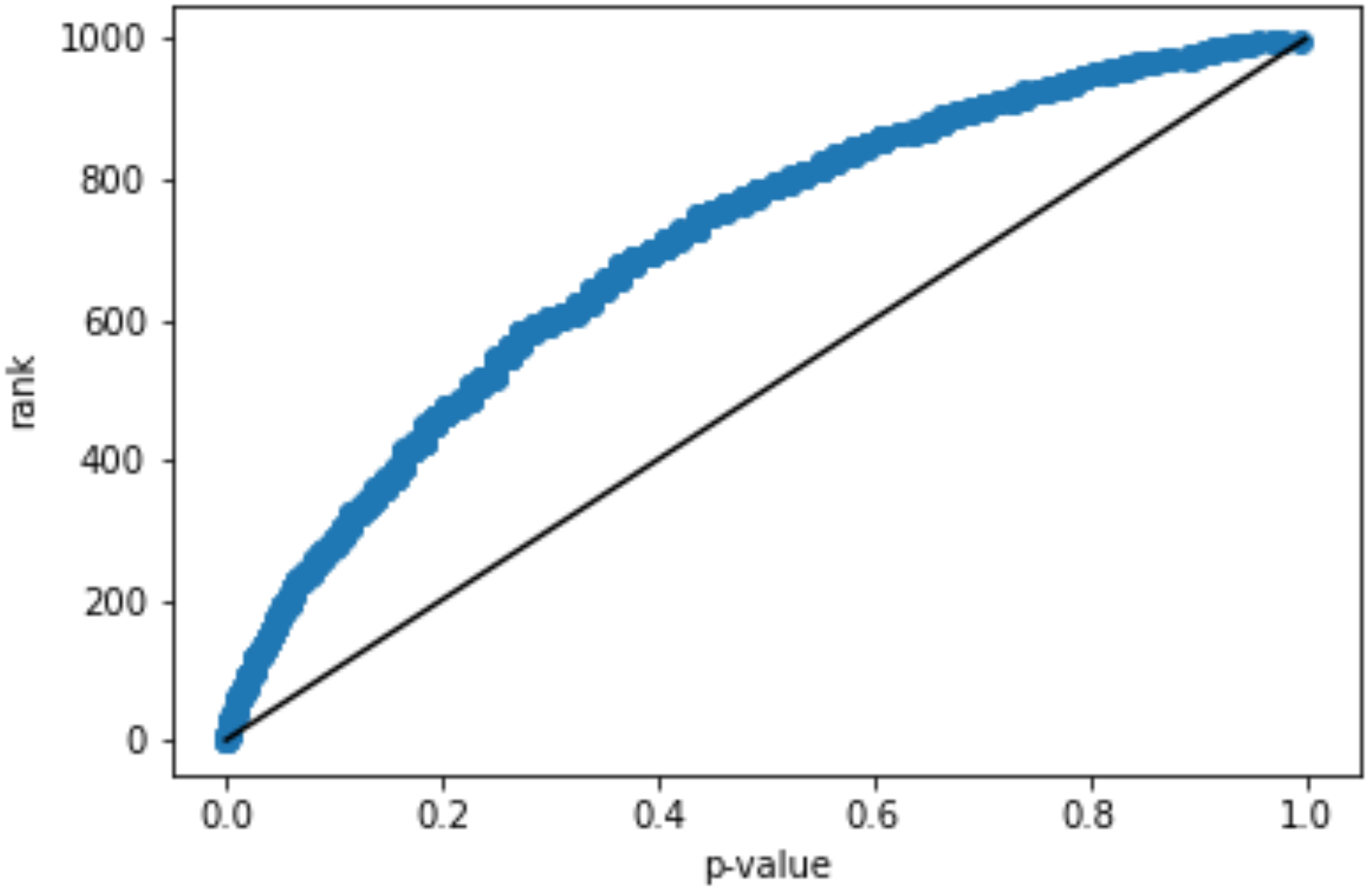
WRS FDR: A ~N(0,1) and B ~ N(0,1), n= 20,20



IDC HERZLIYA

Zohar Yakhini

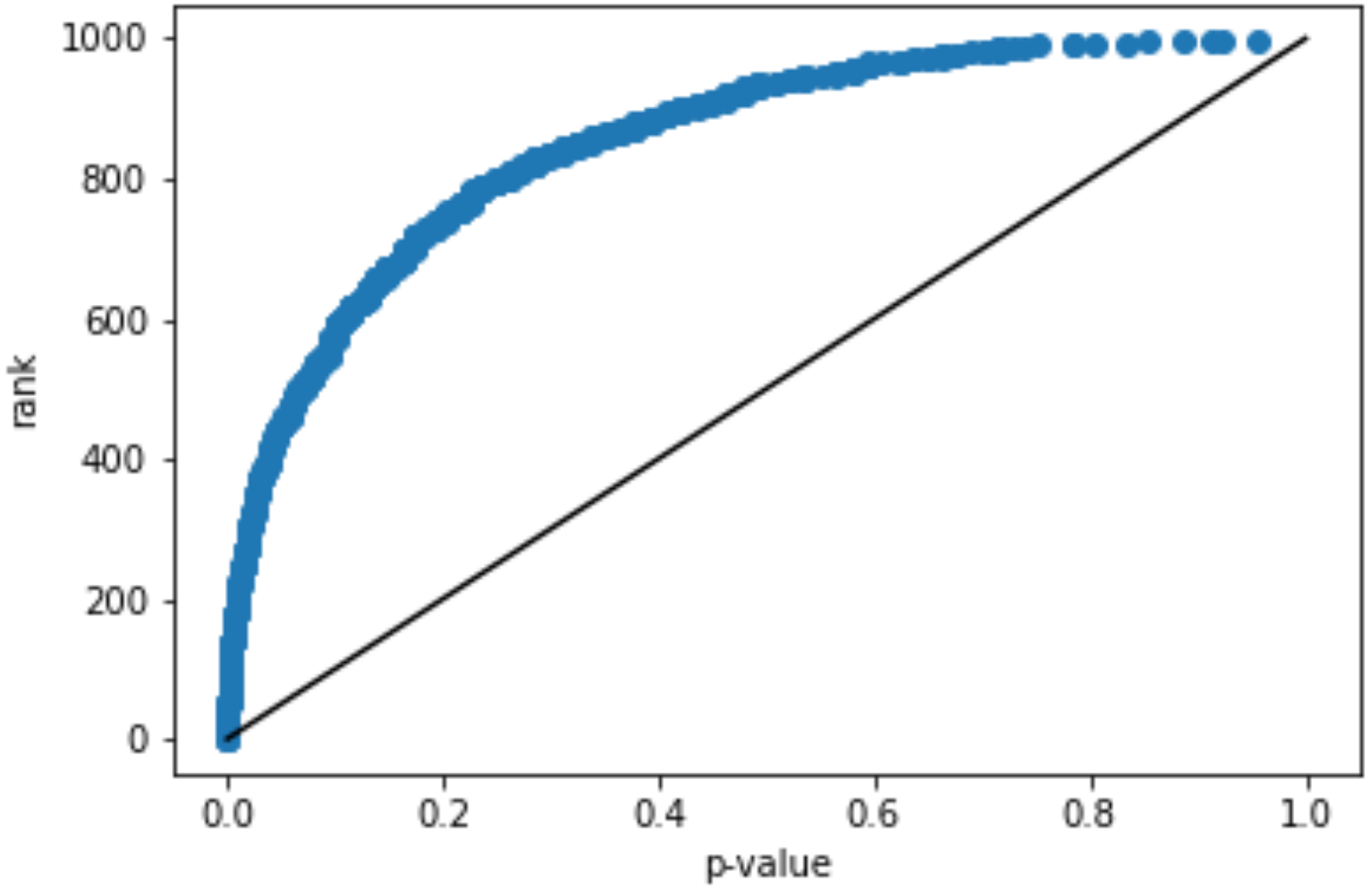
WRS FDR: A ~N(0,1) and B ~ N(0.25,1), n= 20,20



IDC HERZLIYA

Zohar Yakhini

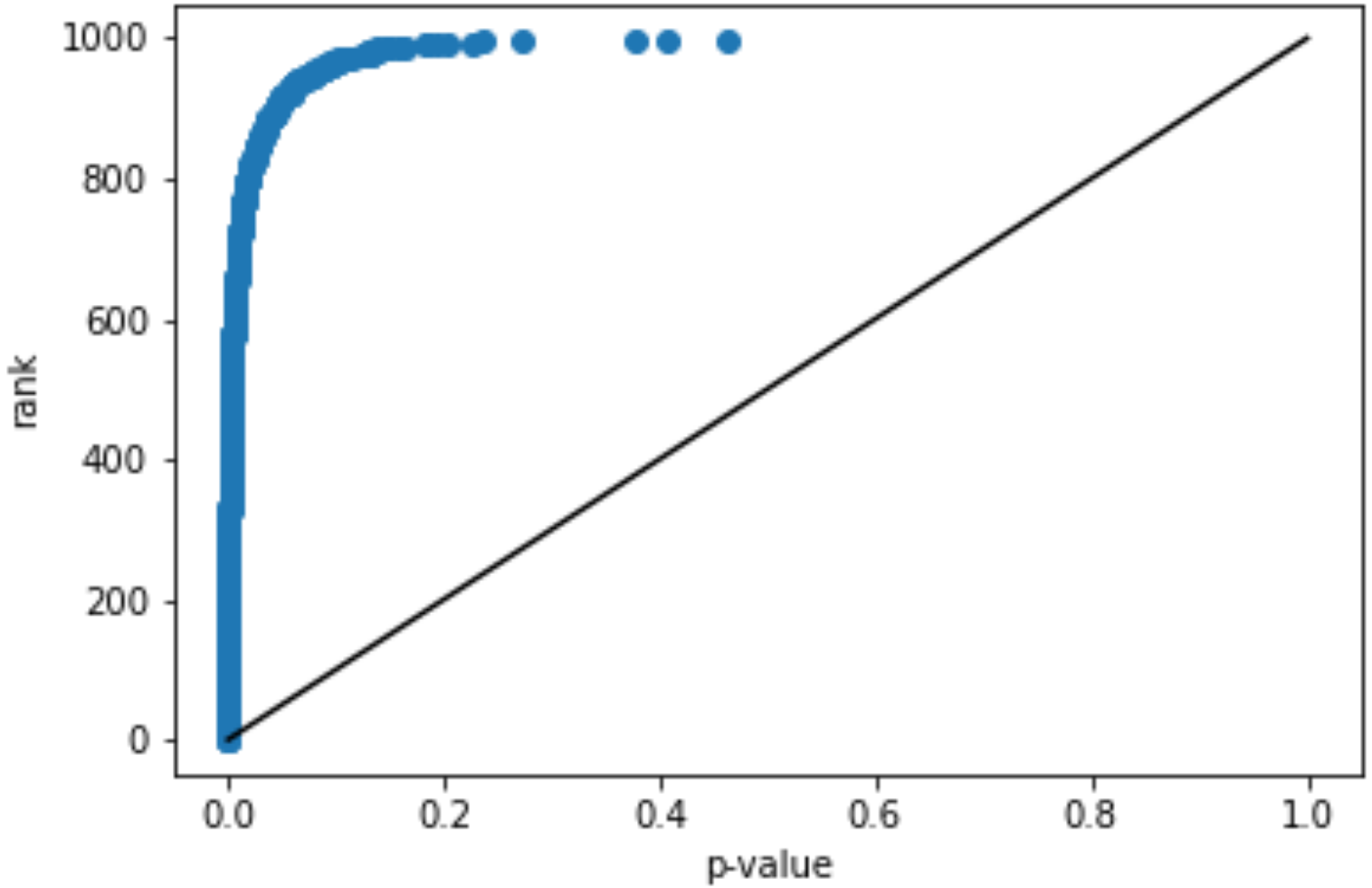
WRS FDR: A ~N(0,1) and B ~ N(?,1), n= 20,20



IDC HERZLIYA

Zohar Yakhini

WRS FDR: A ~N(0,1) and B ~ N(1,1), n= 20,20



IDC HERZLIYA

Zohar Yakhini

FDR – the procedure

Yoav Benjamini and Yosef Hochberg 1995

Replace by the robust version

- Assume that we performed N measurements (comparisons, observations etc)
- Rank the computed significance of the findings:
  $p(1) \le p(2) \le p(3) \le p(4) \le \ldots. \le p(N-1) \le p(N)$
- Under the null model, the expected number of observations with p-value better than p(i) is $p(i) \cdot N$
- The false discovery rate at i is therefore:

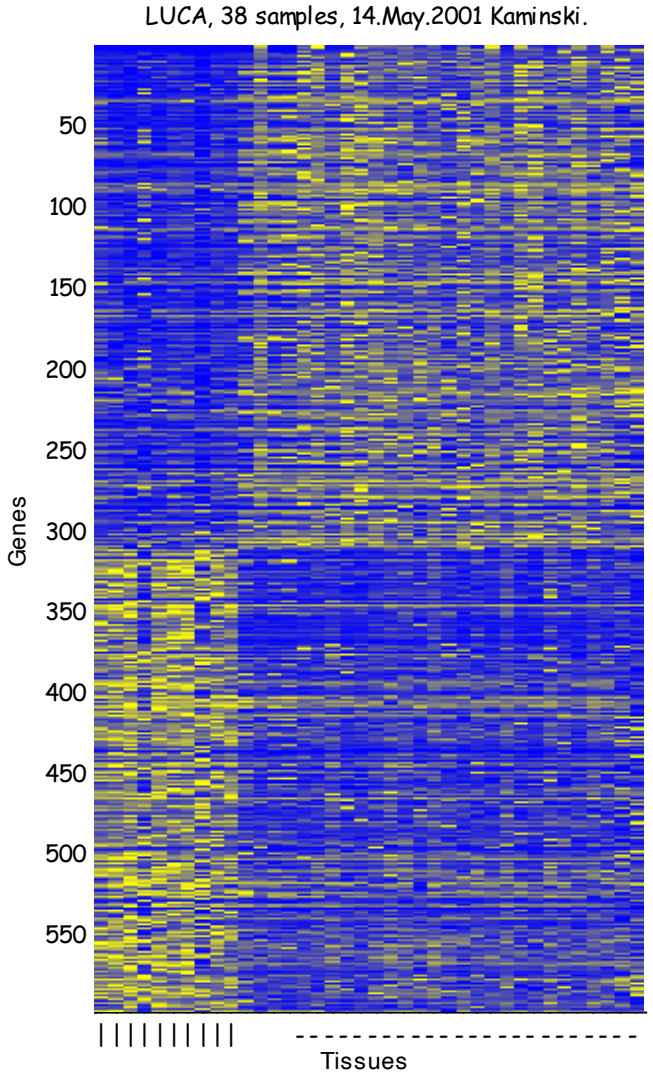
$$\mathrm{RobFDR}(i) = \min_{j \ge i} \left(\frac{p(j) \cdot N}{j} \right)$$

- A corrected hypothesis testing in this case would be to find the max i that satisfies $\mathrm{FDR}(i) \le \tau$, where τ (e.g 0.05) is the required false discovery rate.

IDC HERZLIYA

Zohar Yakhini

In the context of DGE …



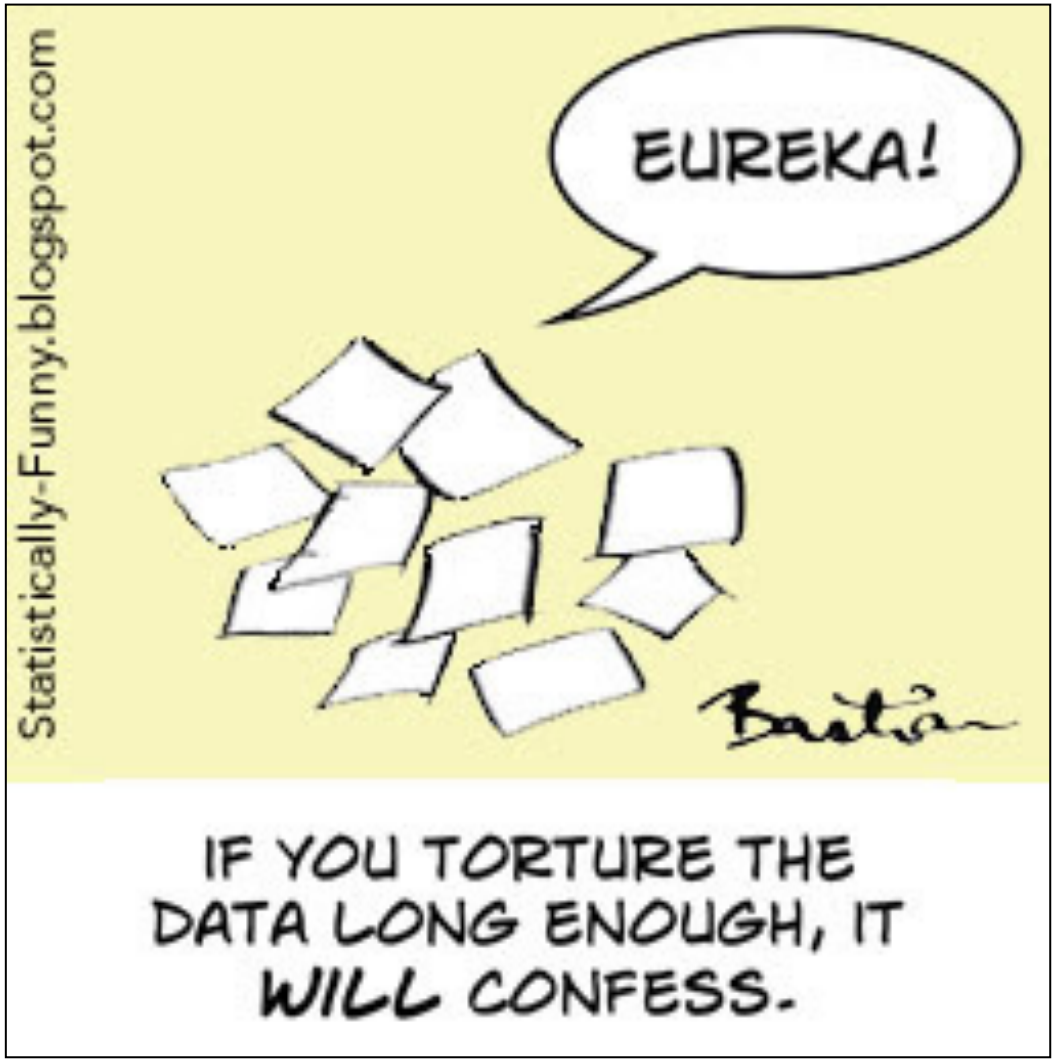
We observed (say) 200 genes at FDR=0.05, using a WRS test

IDC HERZLIYA

Zohar Yakhini

Multiple testing and FDR

- Greater data volumes require more careful inferential statistics approaches.
- Approaches to addressing multiple testing:
  - Bonferroni correction
  - Report FDR results
  - Simulations under a null



IDC HERZLIYA