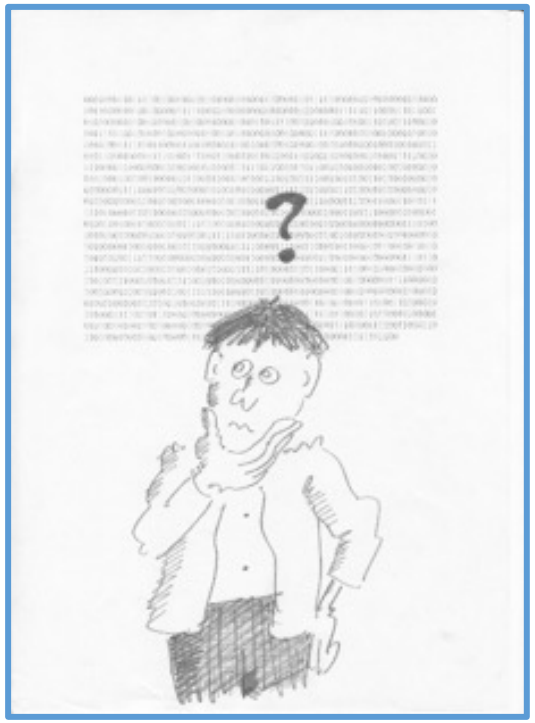
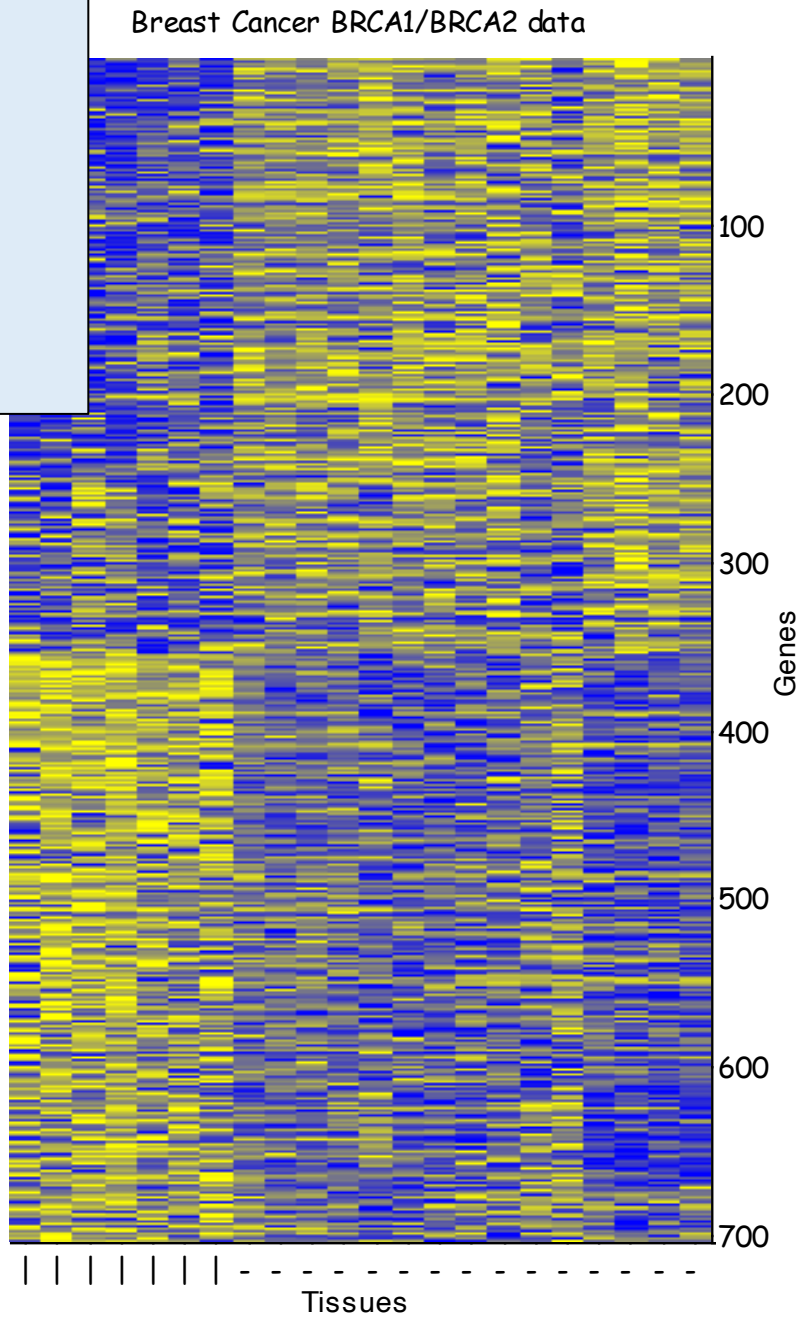
A brief guide to multiple testing and FDR

Zohar Yakhini,
Leon Anavy,
Ben Galili
– RUNI Herzliya

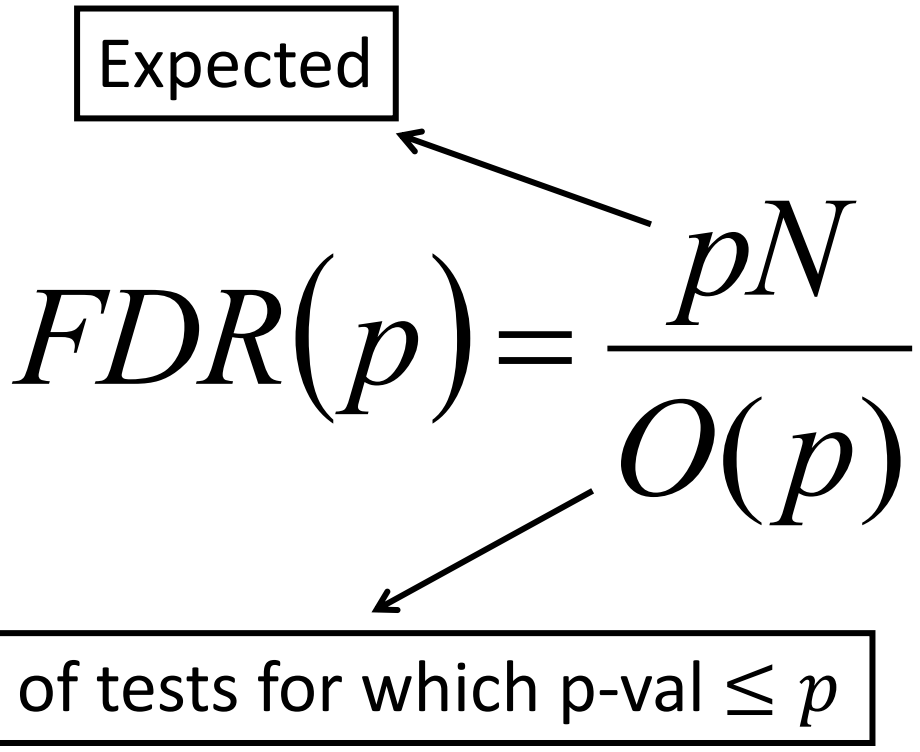
IDC HERZLIYA





False Discovery Rate (FDR)

What fraction of the observed discoveries (e.g DE) is explained by what is expected at random (under a null-model)?

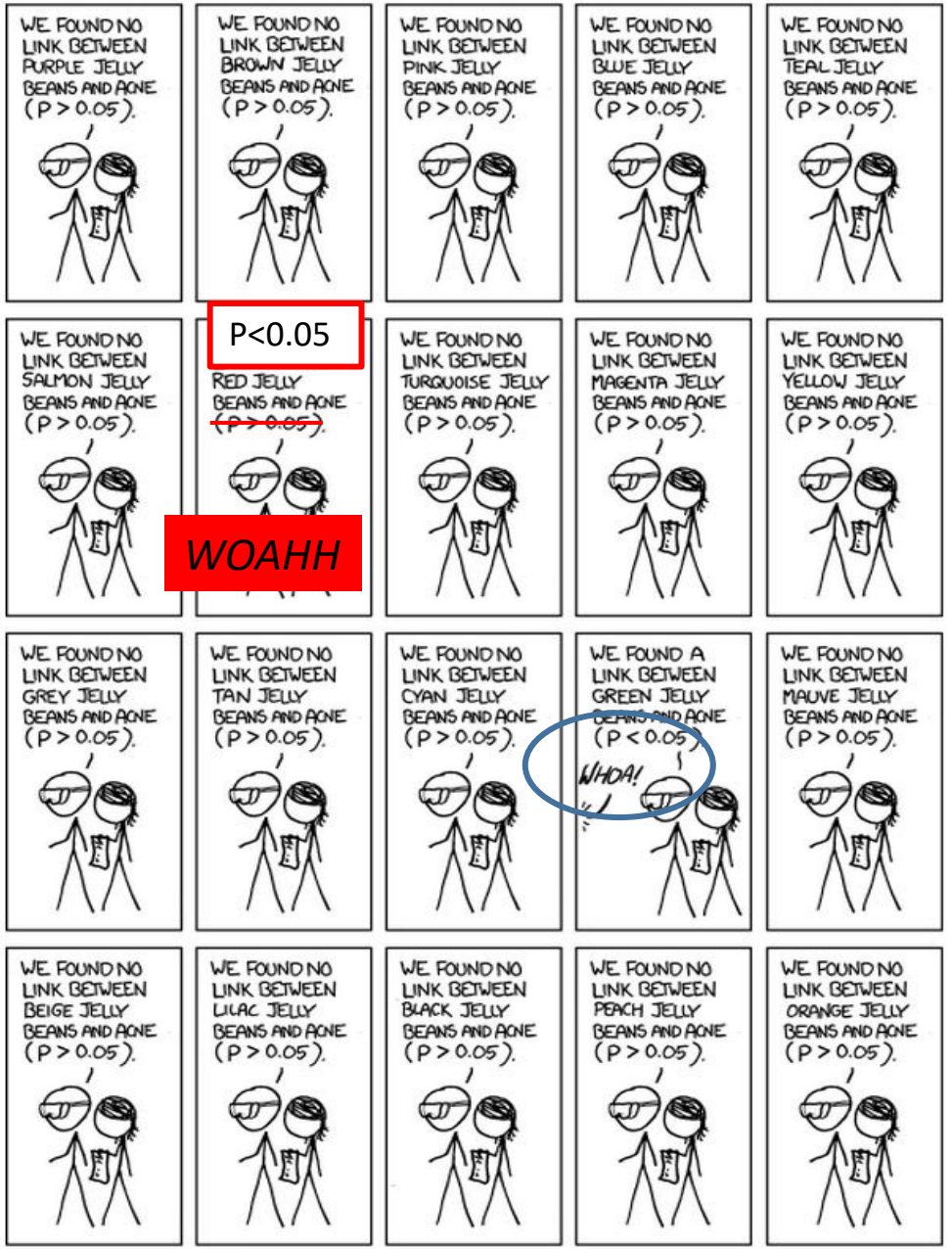


IDC HERZLIYA

Zohar Yakhini

More events

What if two colors were linked at 0.05? Three?



IDC HERZLIYA

Zohar Yakhini

FDR at any i

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

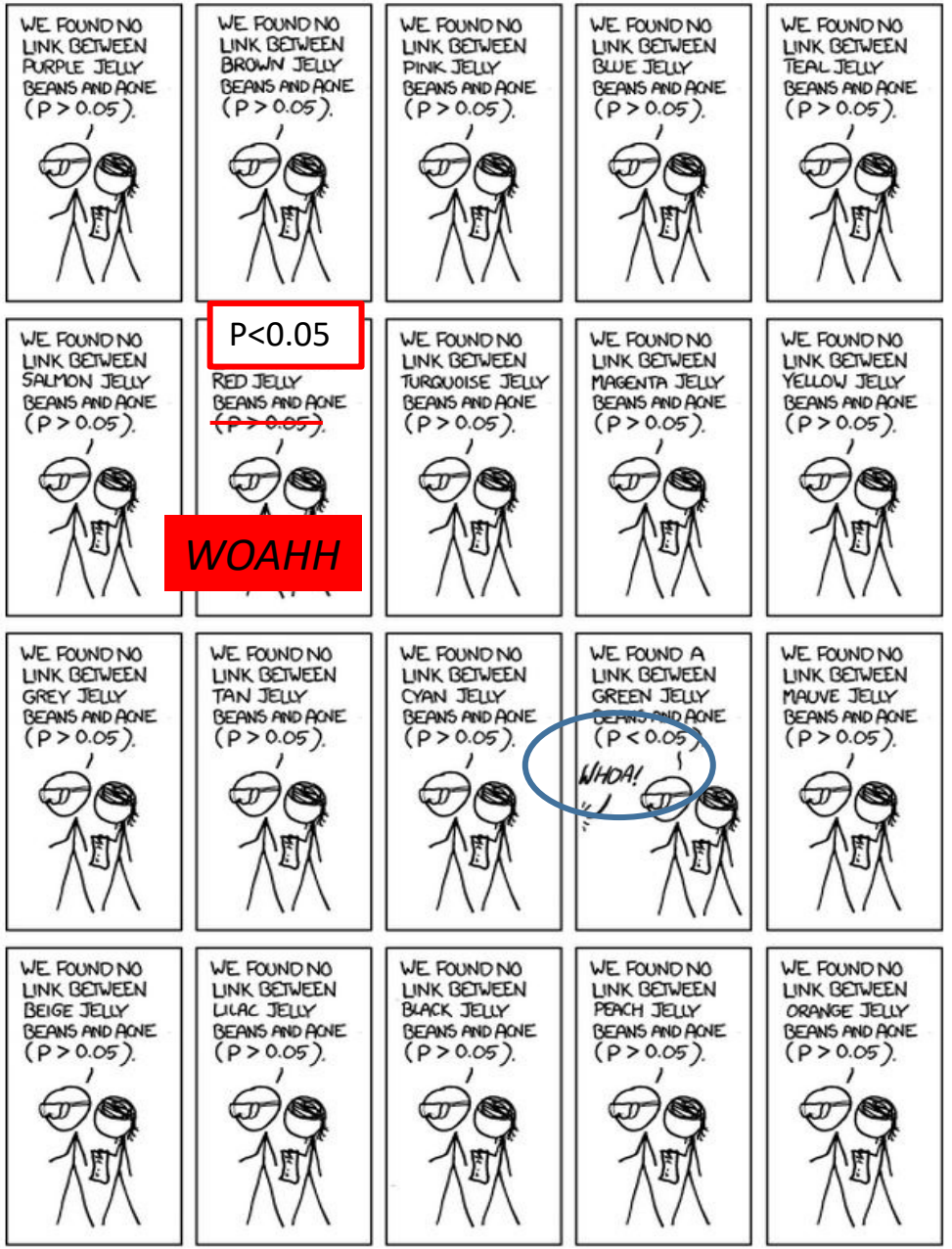
$$FDR(i) = \frac{p(i)N}{i}$$

IDC HERZLIYA

Zohar Yakhini

More events

$$FDR(2) = ?$$

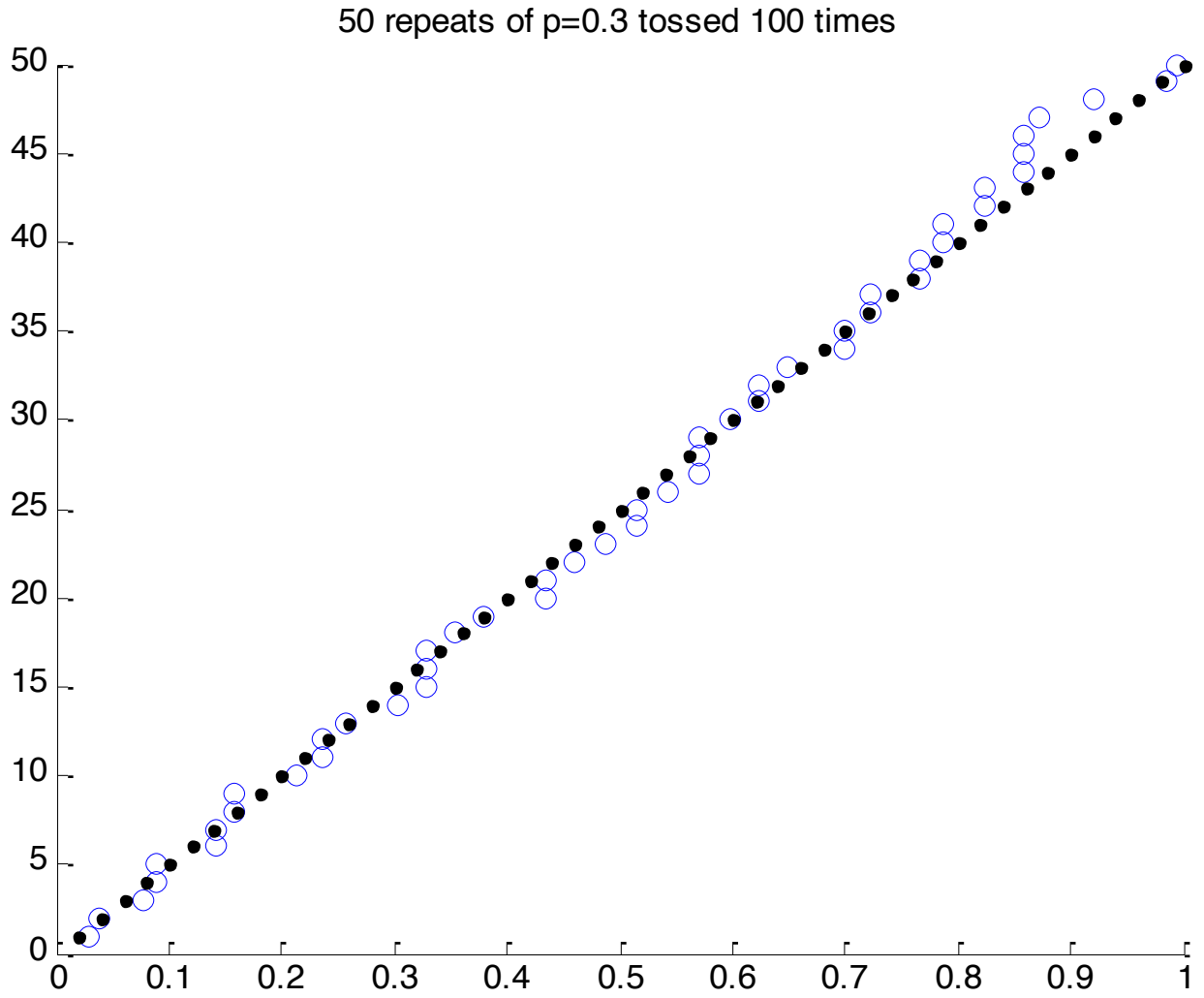


IDC HERZLIYA

Zohar Yakhini

Distribution of p-values under the null model

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

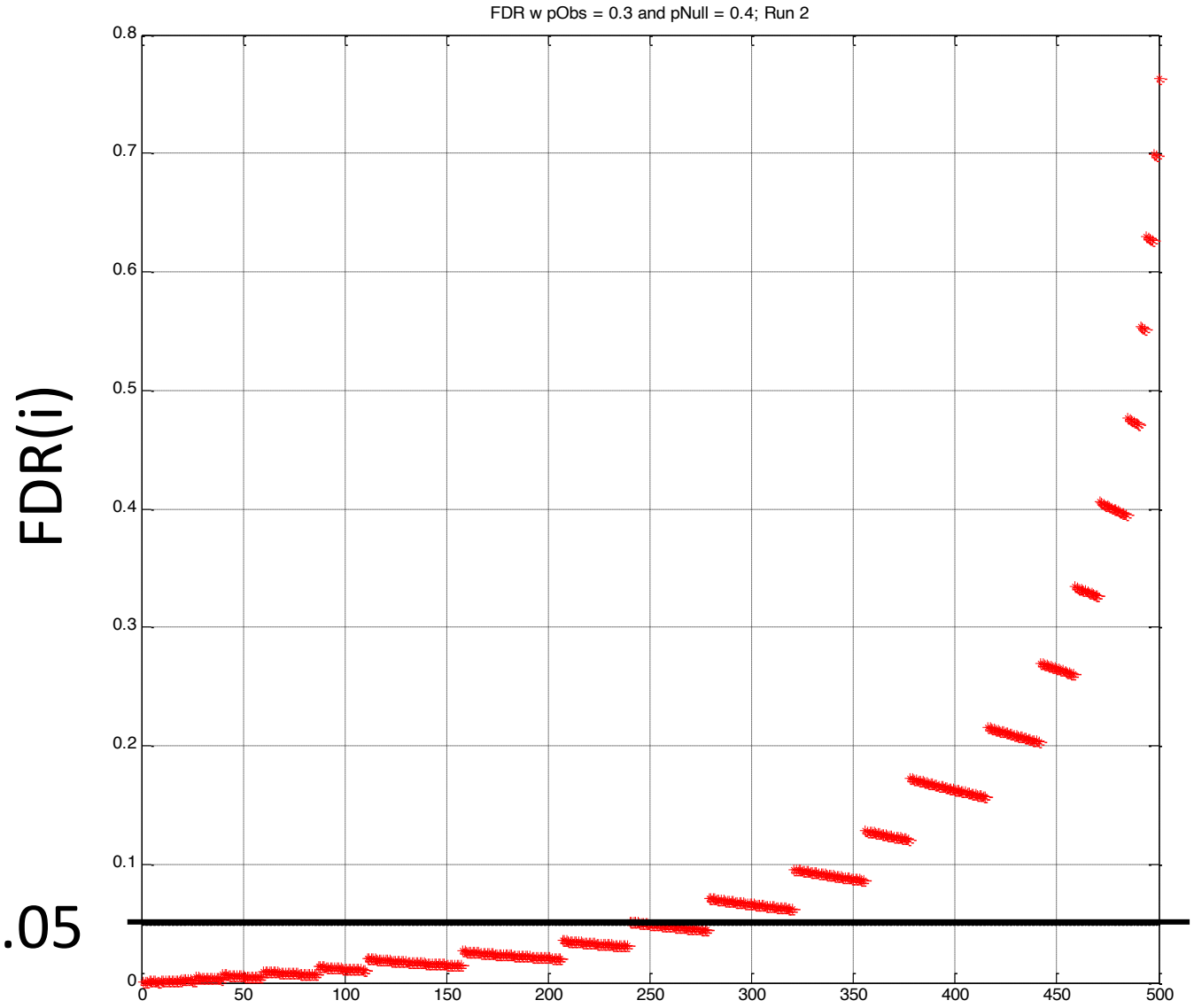


IDC HERZLIYA

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

Zohar Yakhini

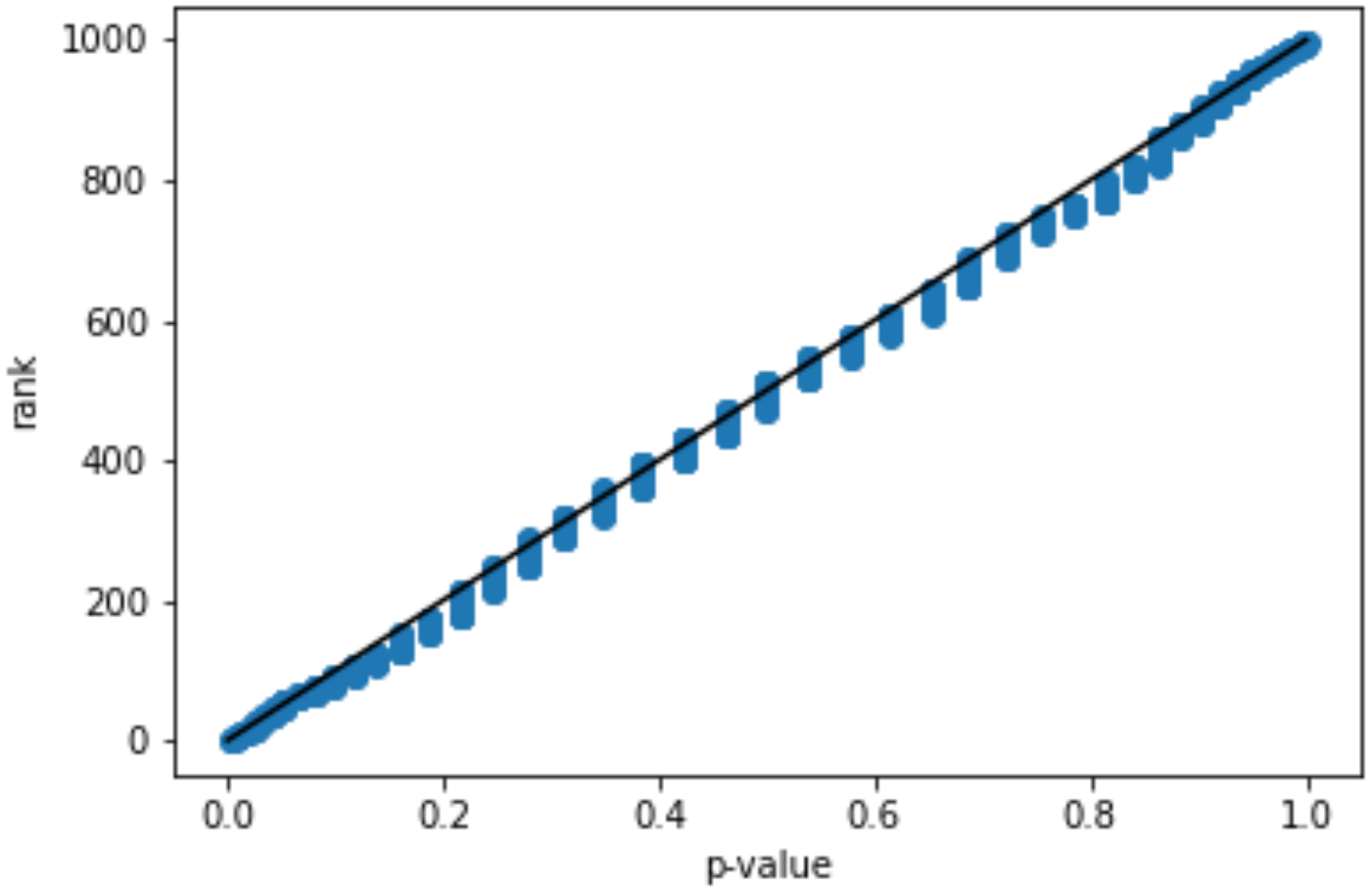
A more adequate correction

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

IDC HERZLIYA

Zohar Yakhini

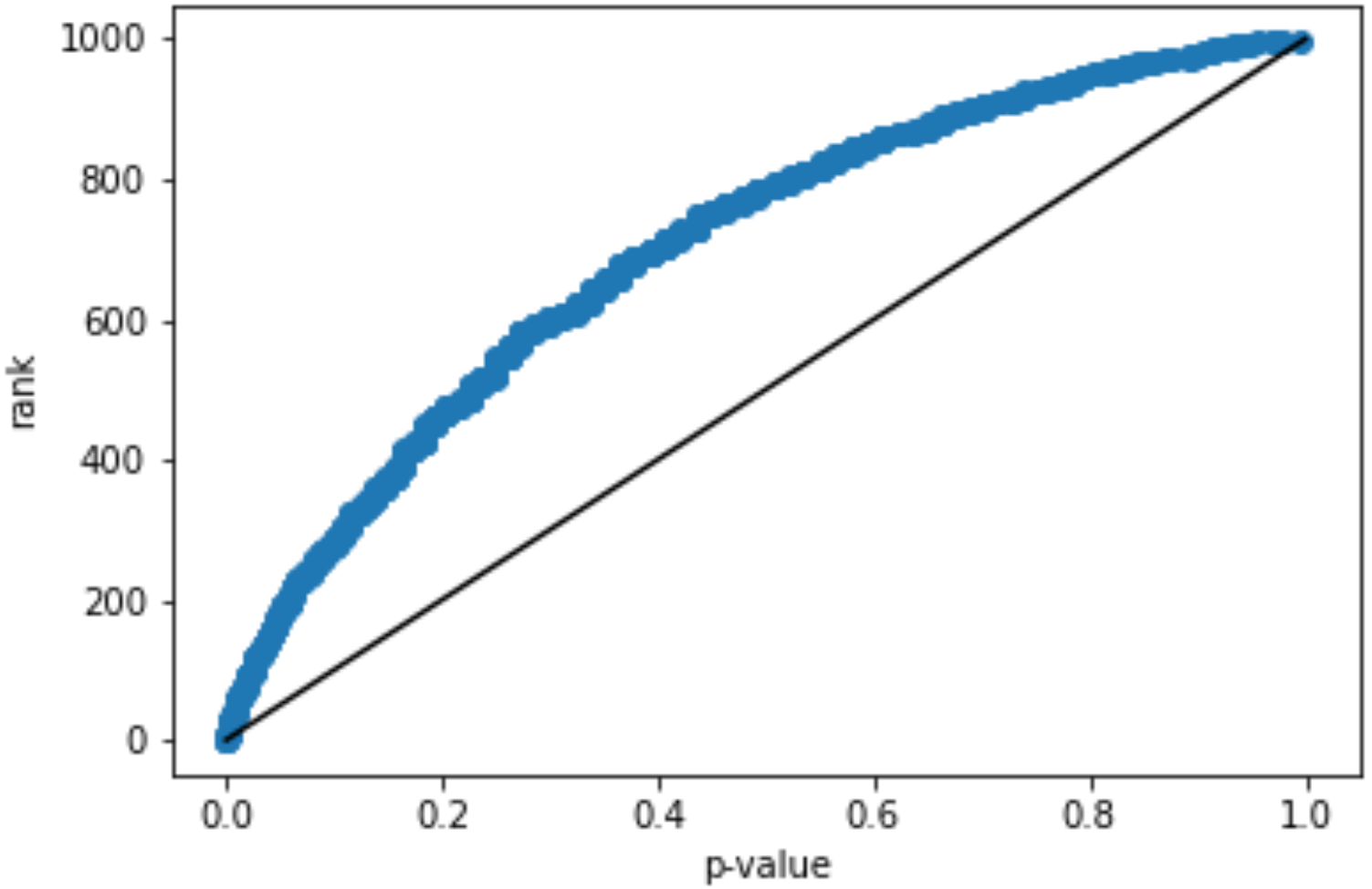
WRS FDR: A ~N(0,1) and B ~ N(0,1), n= 20,20



IDC HERZLIYA

Zohar Yakhini

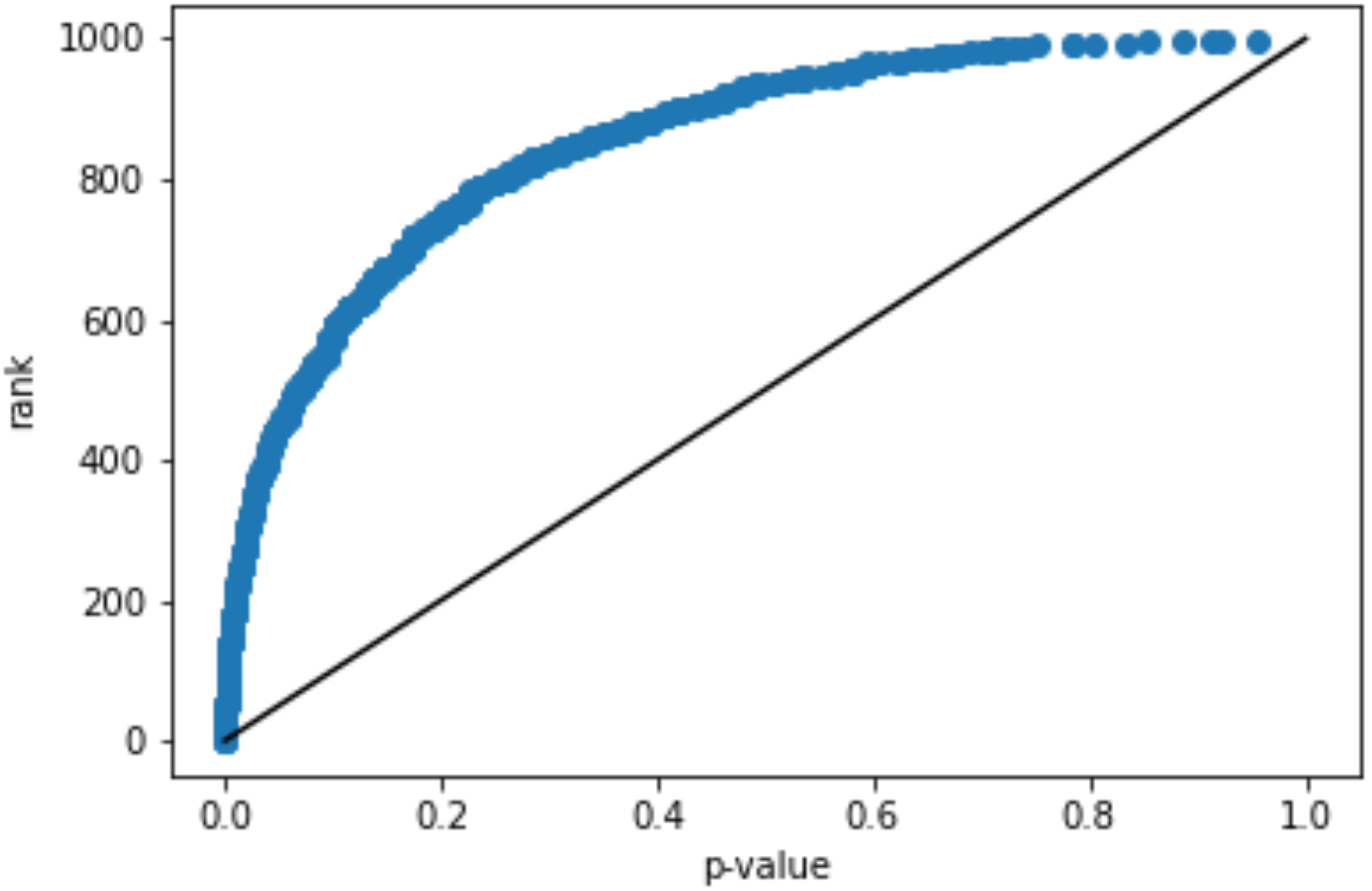
WRS FDR: A ~N(0,1) and B ~ N(0.25,1), n= 20,20



IDC HERZLIYA

Zohar Yakhini

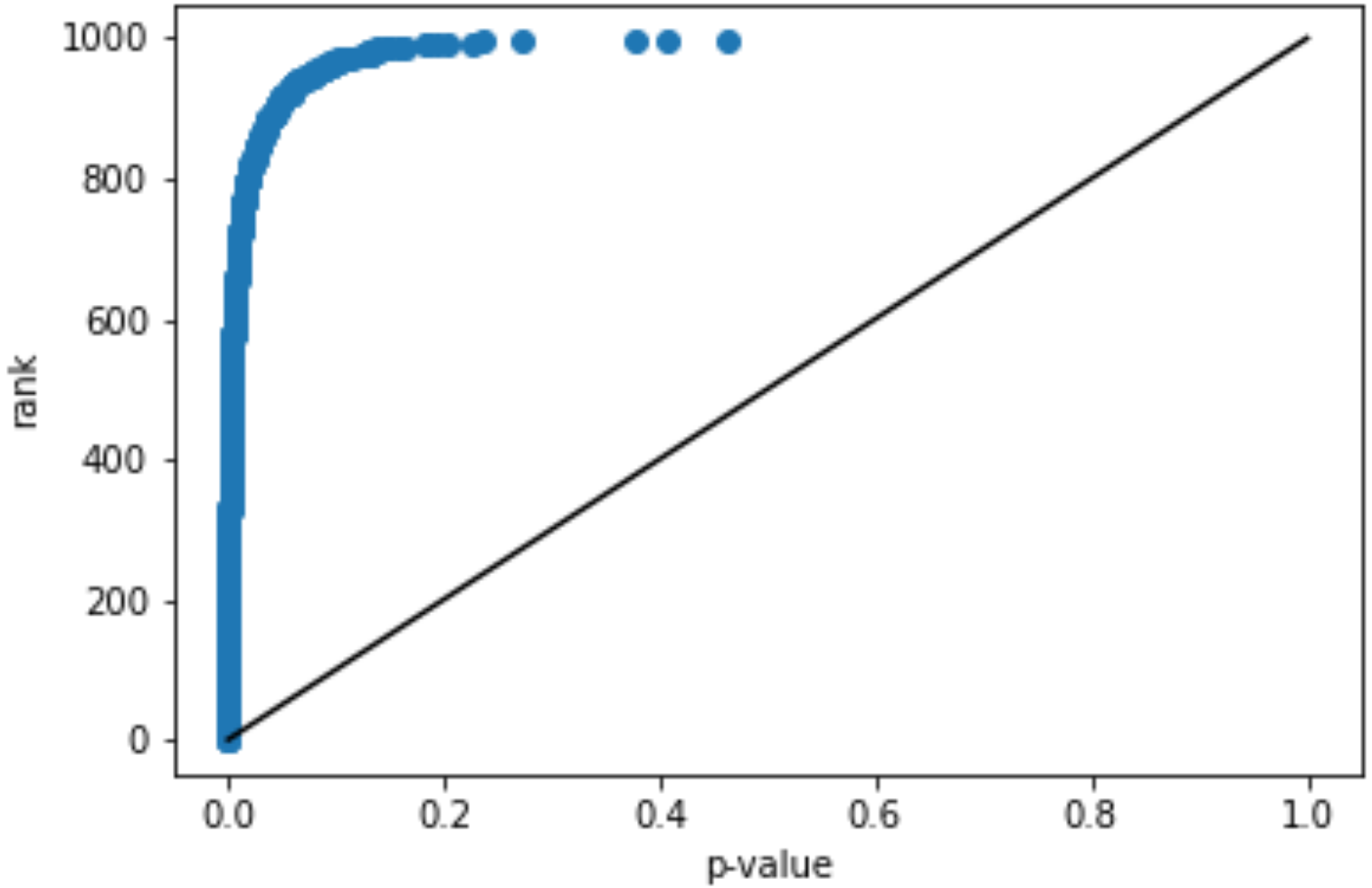
WRS FDR: A ~N(0,1) and B ~ N(?,1), n= 20,20



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

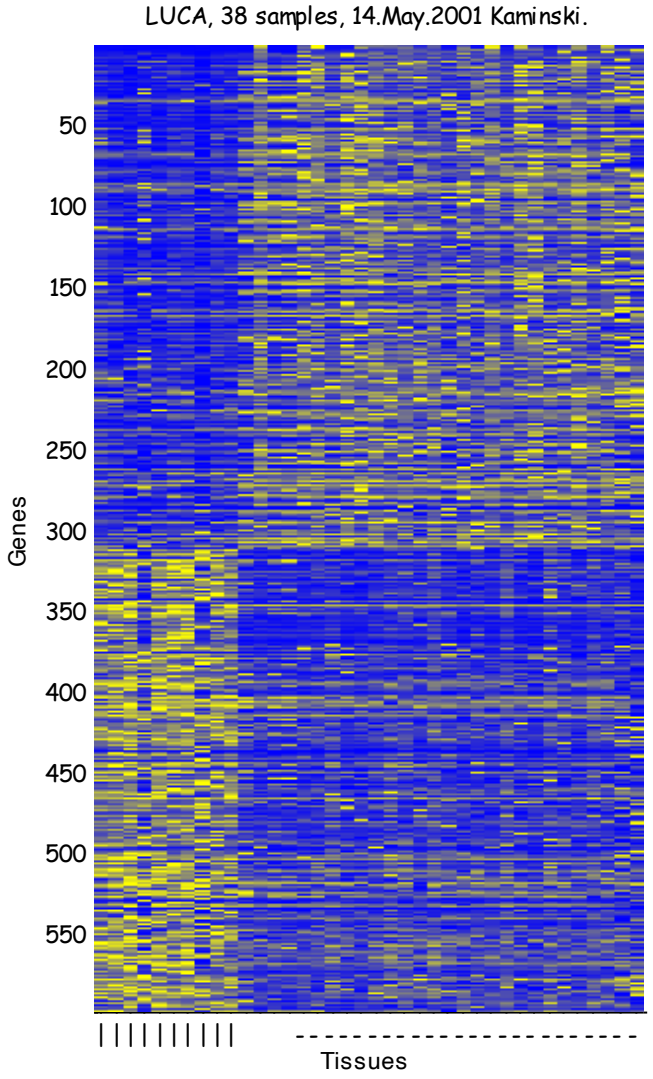
- Assume that we performed N measurements (comparisons, observations etc)
- Rank the computed significance of the findings: $p(1) \leq p(2) \leq p(3) \leq p(4) \leq \ldots. \leq p(N\text{-}1) \leq p(N)$
- Under the null model, the expected number of observations with p-value better than p(i) is $p(i)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

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

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

IDC HERZLIYA

In the context of DGE …



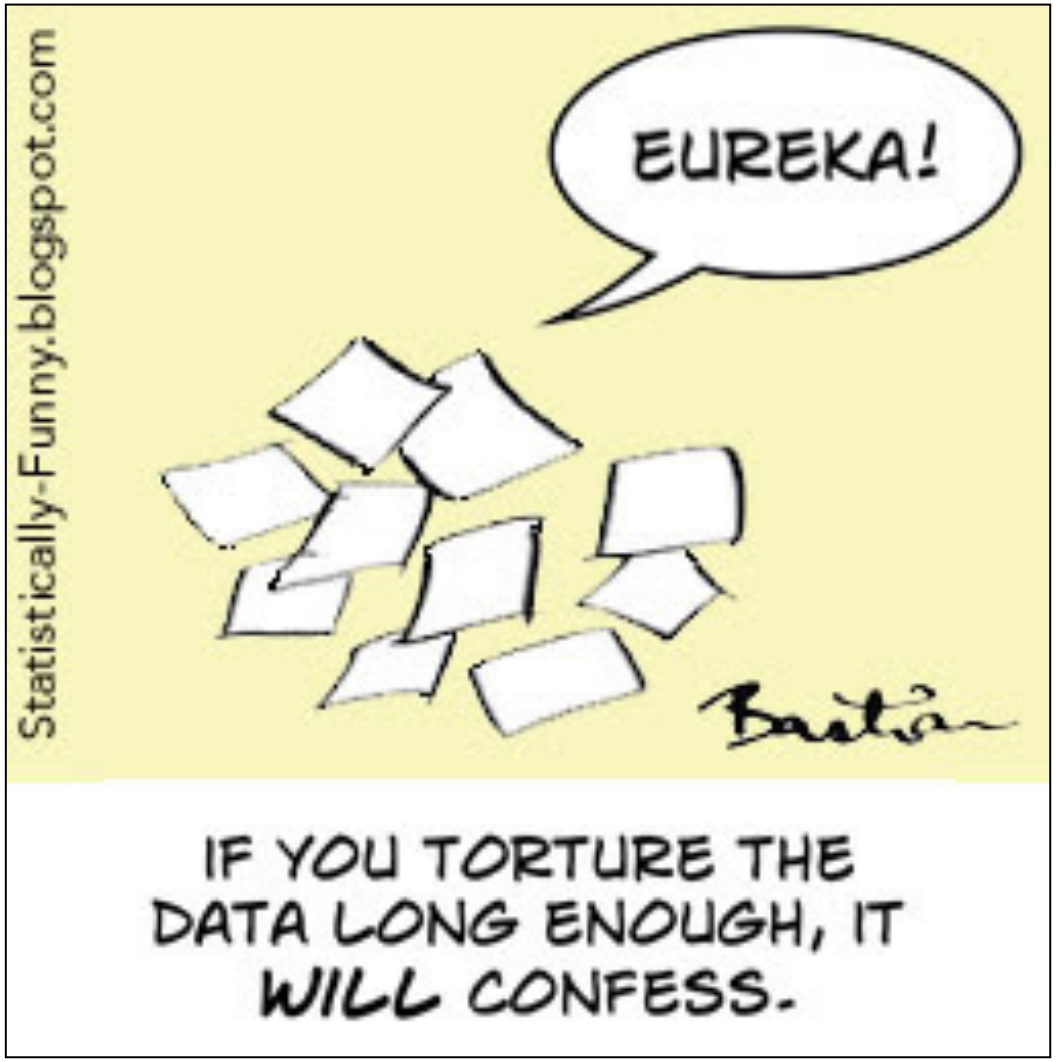
We observed (say) 200 genes at FDR=0.05, using a WRS test

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