Informative Genes

- Differentially expressed in the two classes.
- Identifying (statistically significant) informative genes
  - Provides biological insight
  - Indicate promising research directions
  - Reduce data dimensionality
  - Diagnostic assay

IDC HERZLIYA

Statistical significance of the observed DE

IDC HERZLIYA

Zohar Yakhini, Doron Lipson

Student t-test

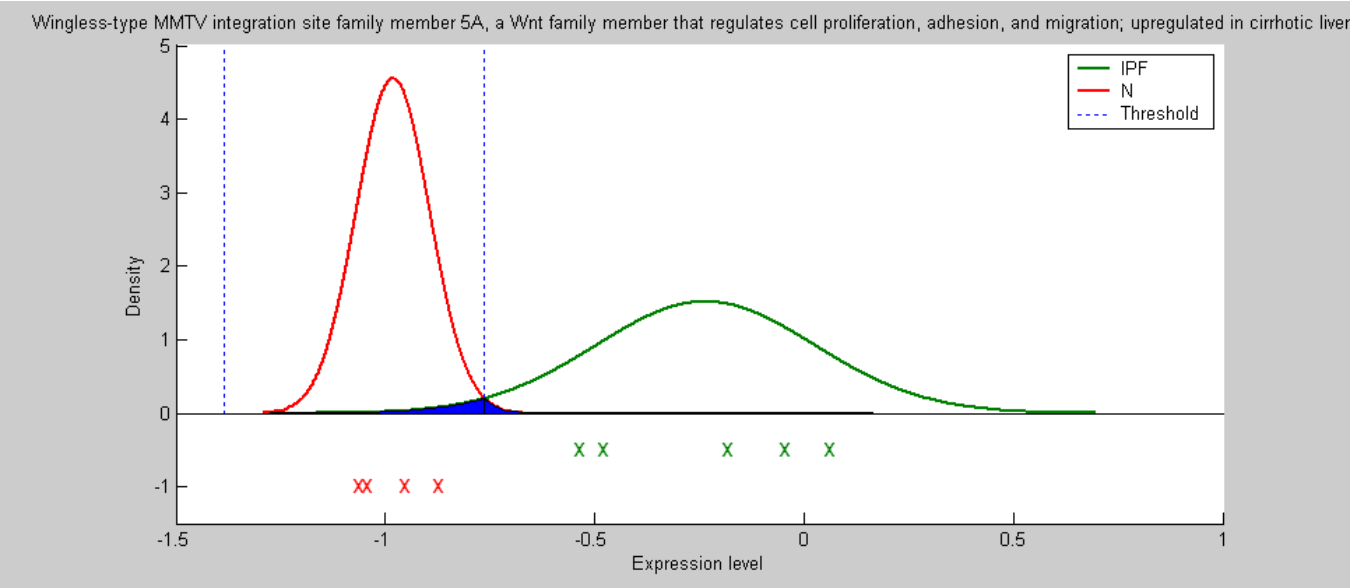
Perform a 2 sample t-test on the two vectors of values.

Zohar Yakhini, Doron Lipson

IDC HERZLIYA

Separation Score

- Compute a Gaussian fit for each class
  $\rightarrow (\mu_1 , \sigma_1) , (\mu_2 , \sigma_2)$.
- The Separation Score is
  $(\mu_1 - \mu_2)/(\sigma_1 + \sigma_2)$



IDC HERZLIYA

Non parametric models

Expression pattern and pathological diagnosis information (annotation), for a single gene

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

Permute the annotation by sorting the expression pattern (ascending, say).

Informative genes

+ + + + + + + + − − − − − − − −
− − − − − − − + + + + + + + +
− − − − + − − + + − + + + + +

etc

Non-informative genes

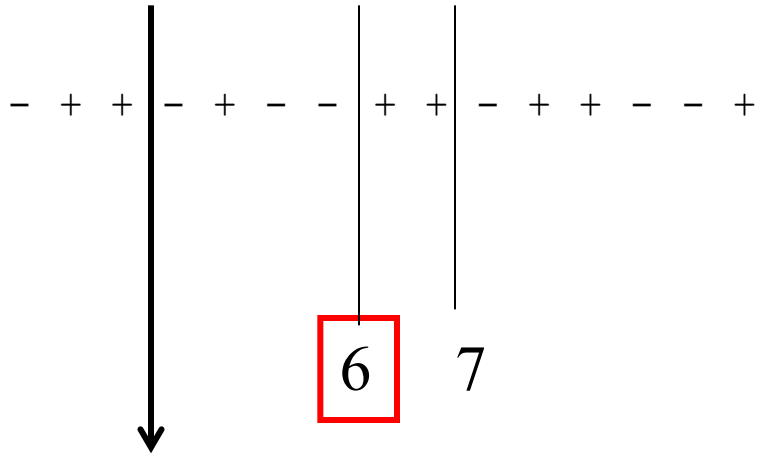
+ − + − + + + + − − + + − − −
− + + − + − − + + − + + − − +
+ − − − + + − + + − + + − + −

etc

Threshold Error Rate (TNoM) Score

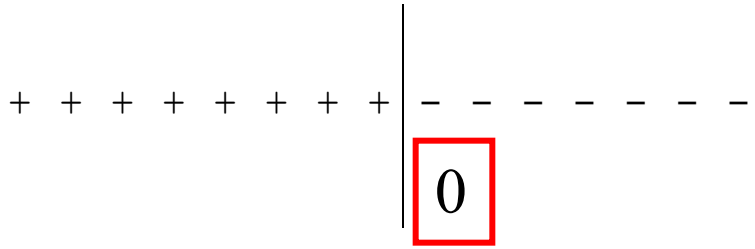
Find the threshold that best separates tumors from normals, count the number of errors committed there.

Ex 1:



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

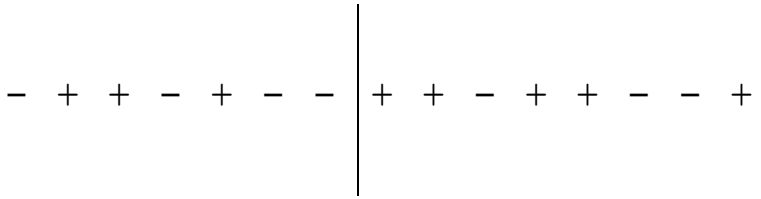
Ex 2: A perfect single gene classifier gets a score of 0.



Threshold Mutual Information (*Info*) Score

The minimal conditional entropy of the annotation, given a threshold partition (minimum taken over thresholds)

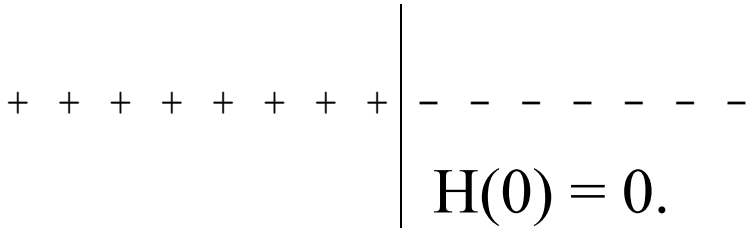
Ex 1:



7/15 H(3/7) + 8/15 H(3/8).

H(p) = p log p + (1-p) log (1-p).

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



H(0) = 0.

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:**
  SR(+) = 1+2+5+6+7+13 = 34

- **The null model: all +/- configurations are equiprobable**

- **The mean value under the null model E(RS(+)) = 6*8 = 48**

- **The p-value of the deviation can be computed**

Wilcoxon - cont

+ + + + + + - - - - - - - - - -

RS(+) = 21

IDC HERZLIYA

p-Values

- DE scores are more useful when we can compute their significance:
  - **p-value:** The probability of finding a gene with a given score if the labeling is random
- p-Values allow for higher level statistical assessment of data quality.
- p-Values provide a uniform platform for comparing observations, across data sets.
- p-Values enable class discovery

IDC HERZLIYA

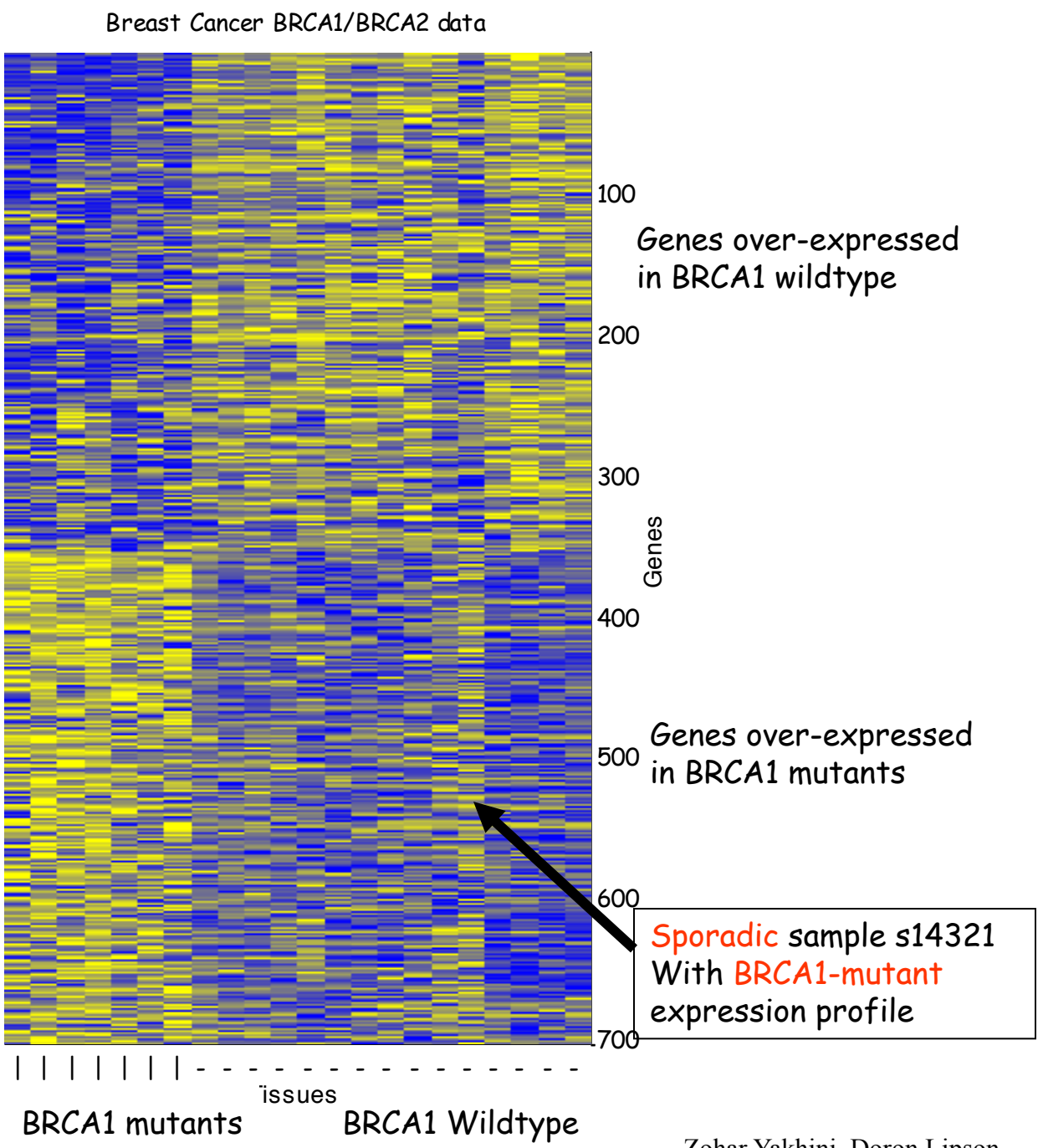
More on p-Values

- Student t-test affords an exact p-value under the equal means null model.
- Wilcoxon rank sum affords an exact p-value and a normal approximation
  (for large B and N)
- TNoM affords an exact p-value

IDC HERZLIYA

BRCA1 Differential Expression

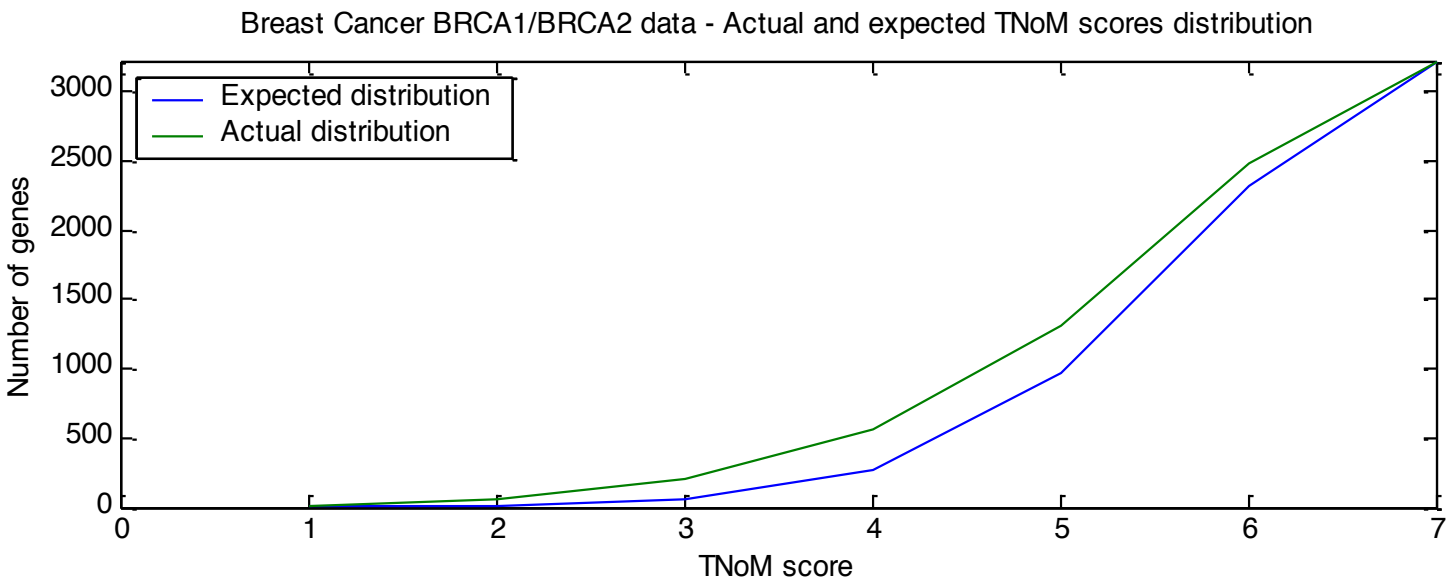
Collab with NIH; Hedenfalk et al, NEJM 2001

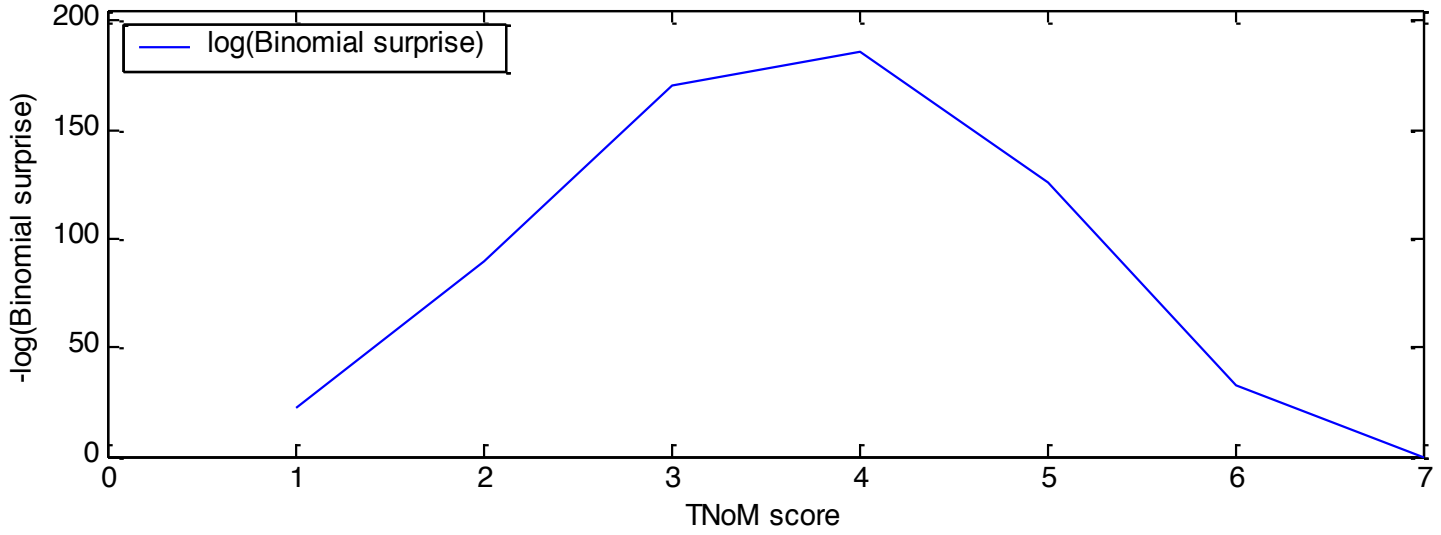


IDC HERZLIYA

Zohar Yakhini, Doron Lipson

BRCA1 Differential Expression: Overabundance Analysis

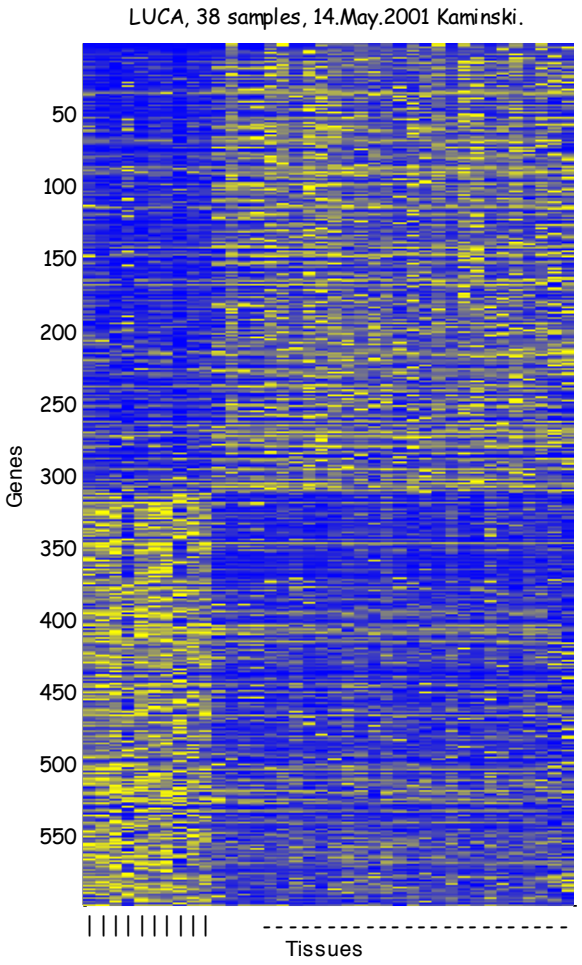




Lung Cancer Informative Genes

Data from Naftali Kaminski's lab, at Sheba.

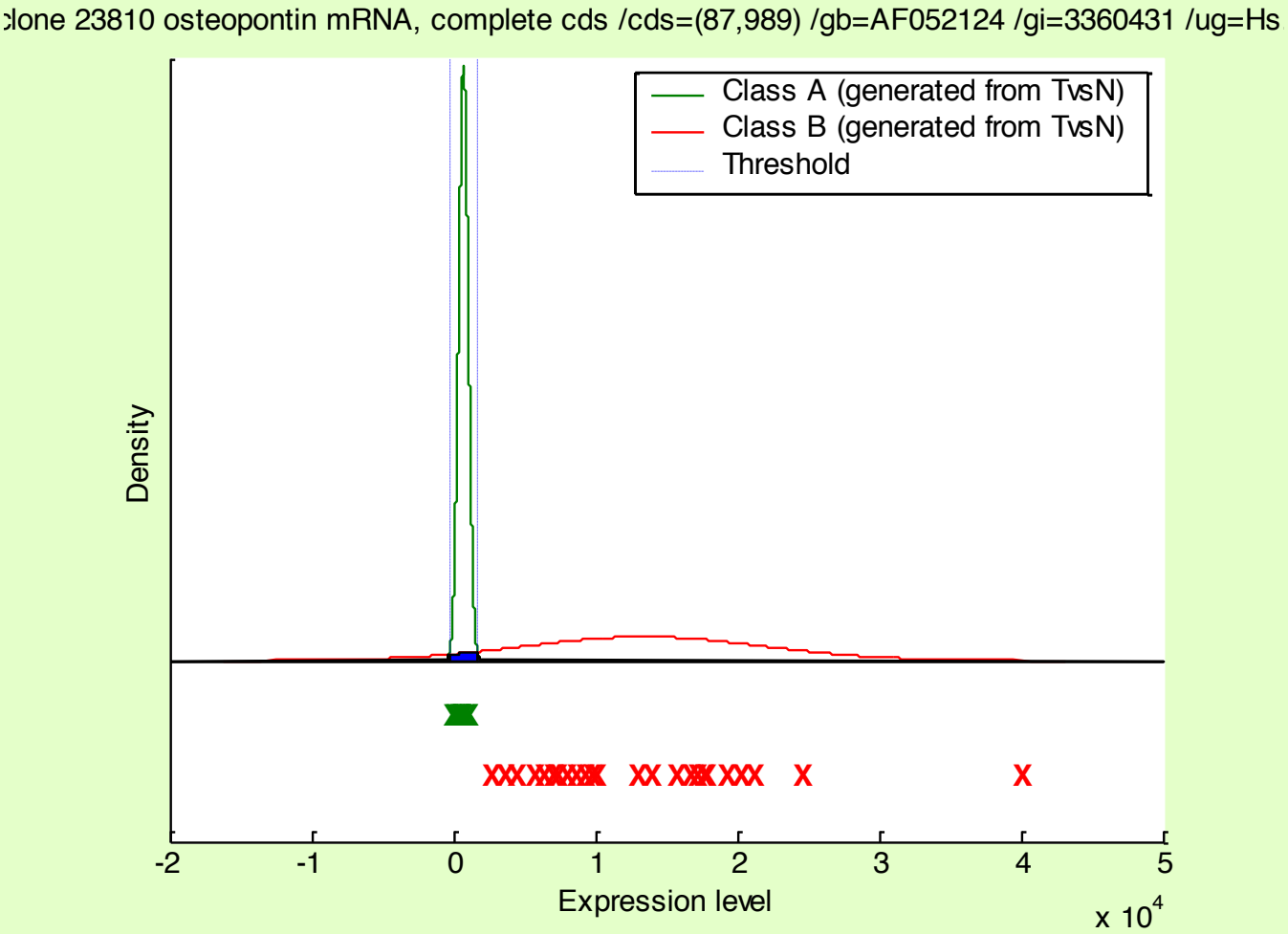
- 24 tumors (various types and origins)
- 10 normals (normal edges and normal lung pools)



IDC HERZLIYA

Zohar Yakhini, Itai Sharon

Ostopontin, Gaussian-Gram



Zohar Yakhini, Doron Lipson

More Applications of Microarrays

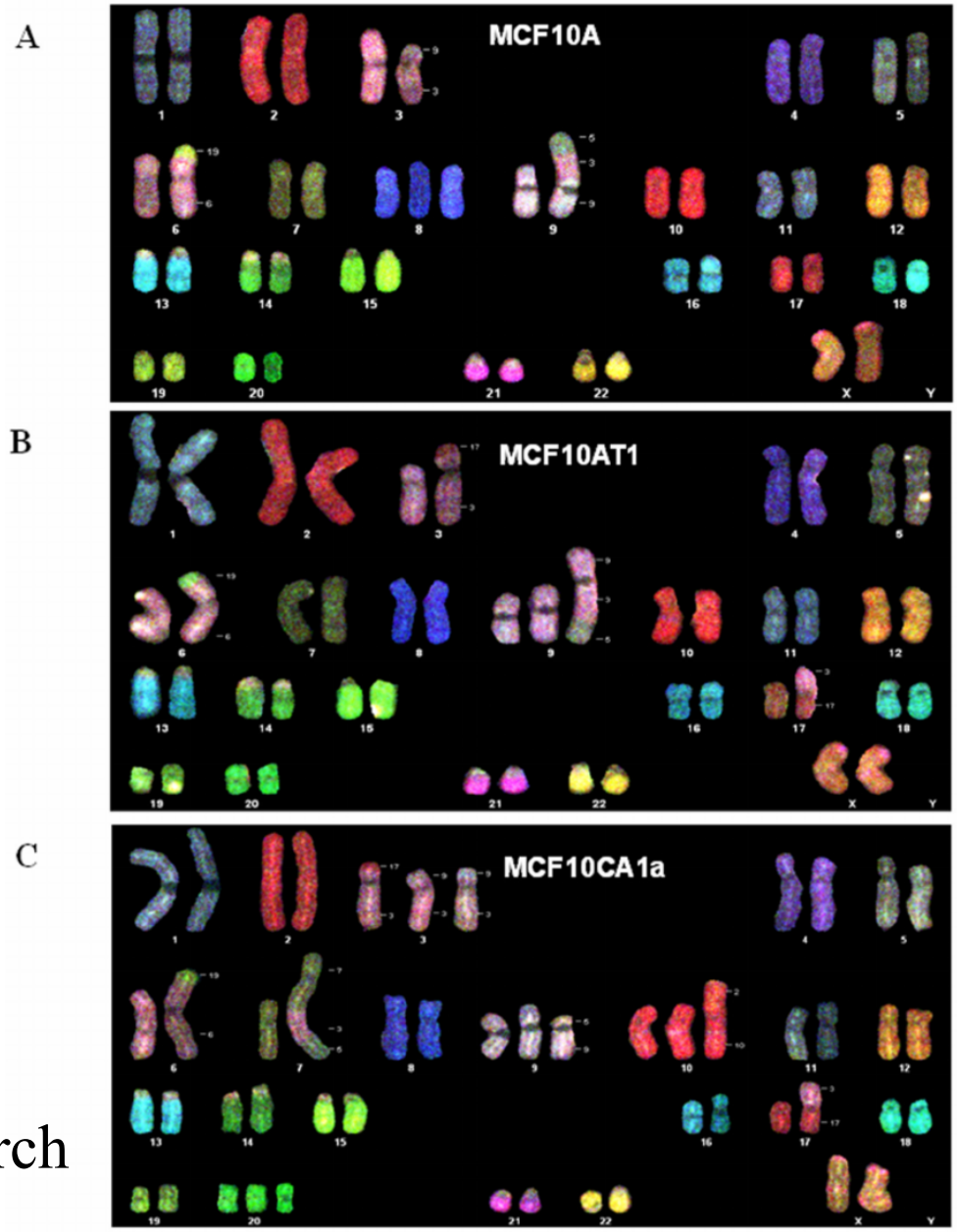
- Polymorphism analysis (genotyping)
- DCN profiling
- Location analysis (ChIP chip)
- Methylation analysis
- …

IDC HERZLIYA

Copy number changes in cancer

The genome structure in cancer cells is aberrant:

- Amplifications
- Deletions
- Translocations

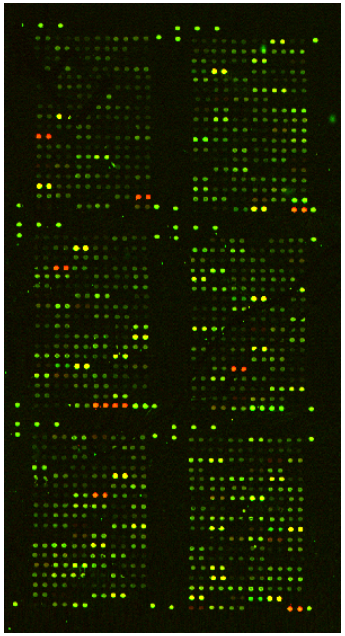


Marella et al, Cancer Research 2009

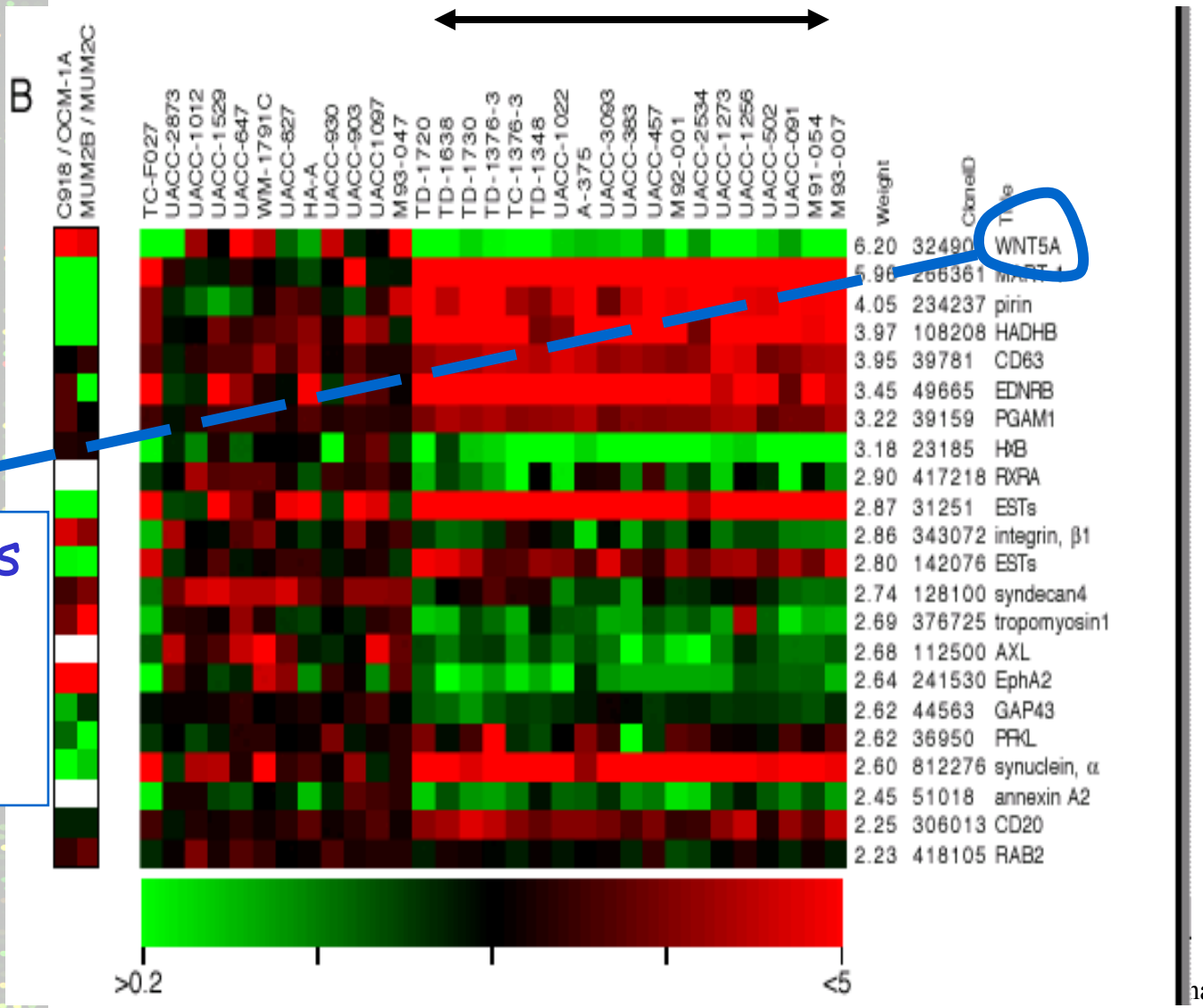
IDC HERZLIYA

Melanoma Sub-Classification

- **Expression profiling of melanoma** (Bittner et al, Nature 2000):
  - 31 melanoma samples, ~3600 genes
  - Found a subclass of 19 samples with a distinct characteristic expression profile
  - Computational method: clustering of the set of samples



Putative Mmelanoma Sub-Class



WNT5A plays a major role in tumor cell mobility

IDC HERZLIYA

haron

Biological Validation

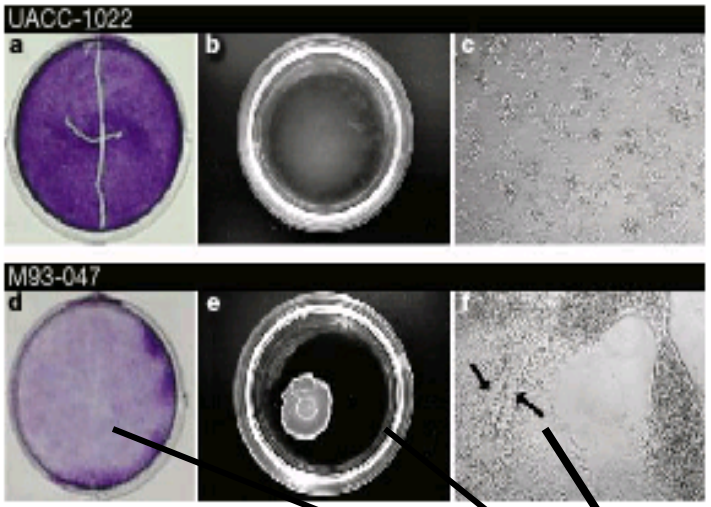


Figure 4 Variation in biological properties of melanoma clusters. **a**–**c**, A representative member of the major melanoma cluster (UACC-1022). **d**–**f**, A sample falling outside of the major cluster (M93-047). The two groups differ in the ability to migrate into a scratch wound (**a**, **d**), contract collagen gels (**b**, **e**) and form tubular networks (**c**, **f**). Results of these and additional cell mobility/invasion assays are included in Table 1. Tubular network formation (vasculogenic mimicry[5], **f**) and collagen gel contraction (related to the patterning of vascular channels, **e**) were observed only outside the major cluster (Table 1).

A sample from the putative sub-class

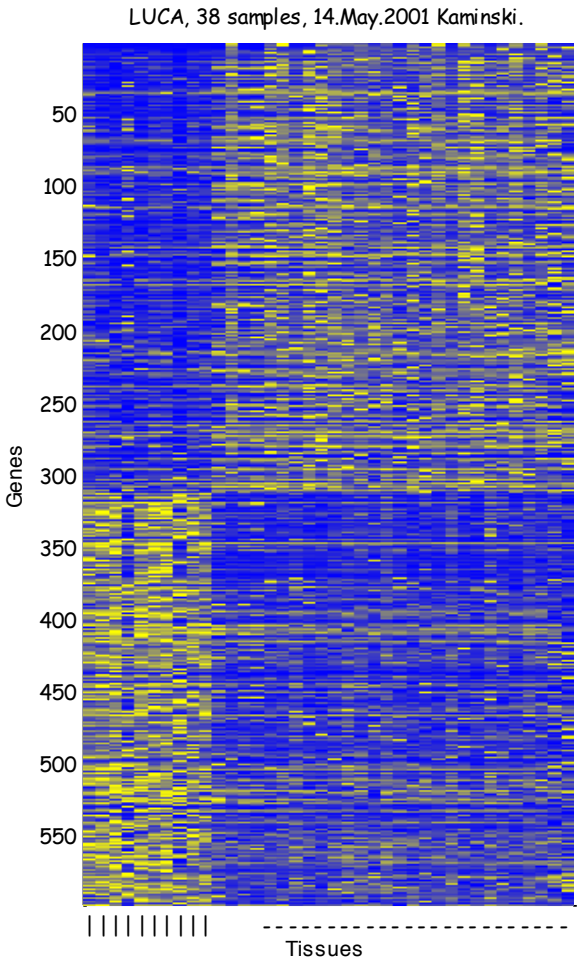
A sample NOT from the putative sub-class

- Migration into scratch wound
- Collagen contraction
- Tubular network formation

Lung Cancer Informative Genes

Dehan et al,
Lung Cancer 2007

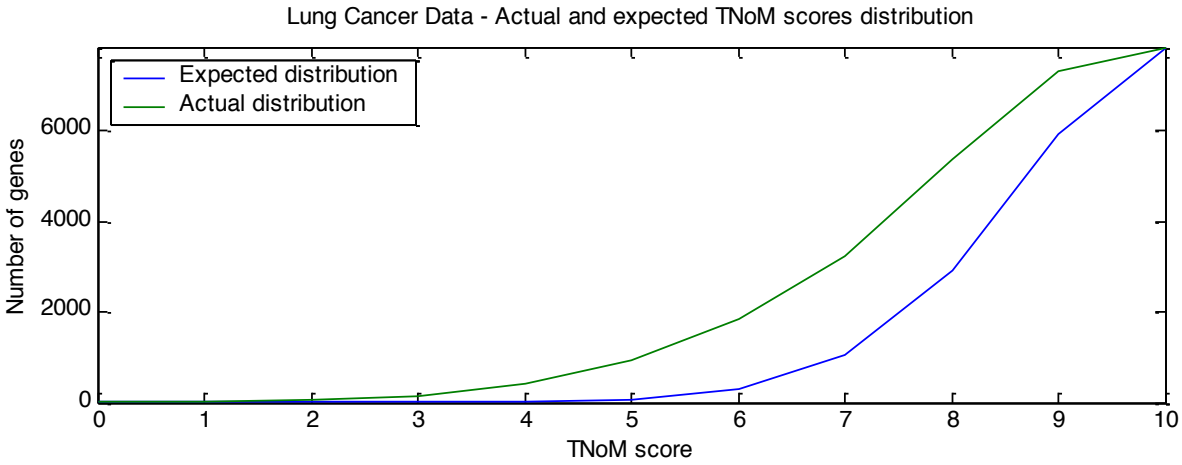
- 24 tumors (various types and origins)
- 10 normals (normal edges and normal lung pools)

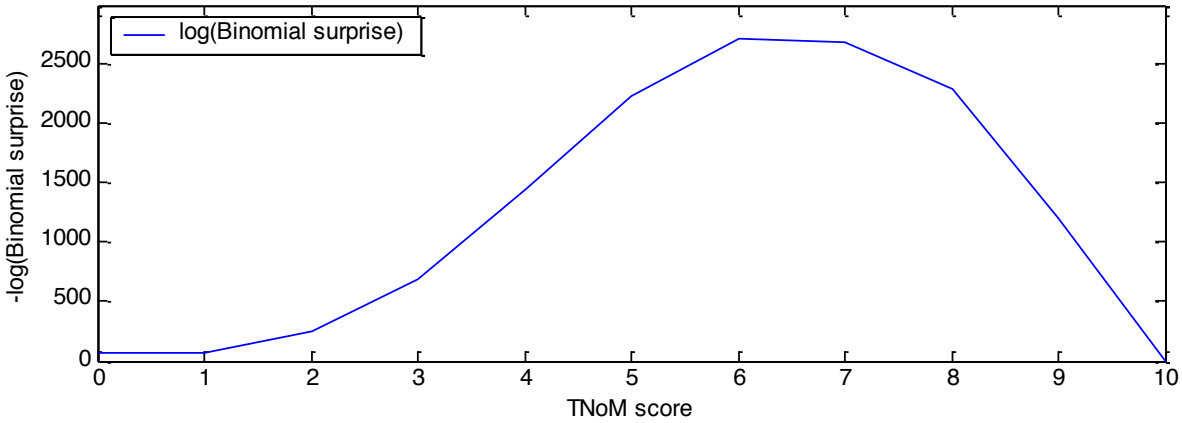


IDC HERZLIYA

Zohar Yakhini, Doron Lipson

Lung Cancer Overabundance Analysis





IDC HERZLIYA

Zohar Yakhini, Itai Sharon

How to correct for multiple testing

Bonferroni correction:
Multiply all p-values by the multiple testing factor (number of observations, comparisons, etc).
In the case of GE:

Bonferroni-p(g) = N*p-Val(g)

IDC HERZLIYA

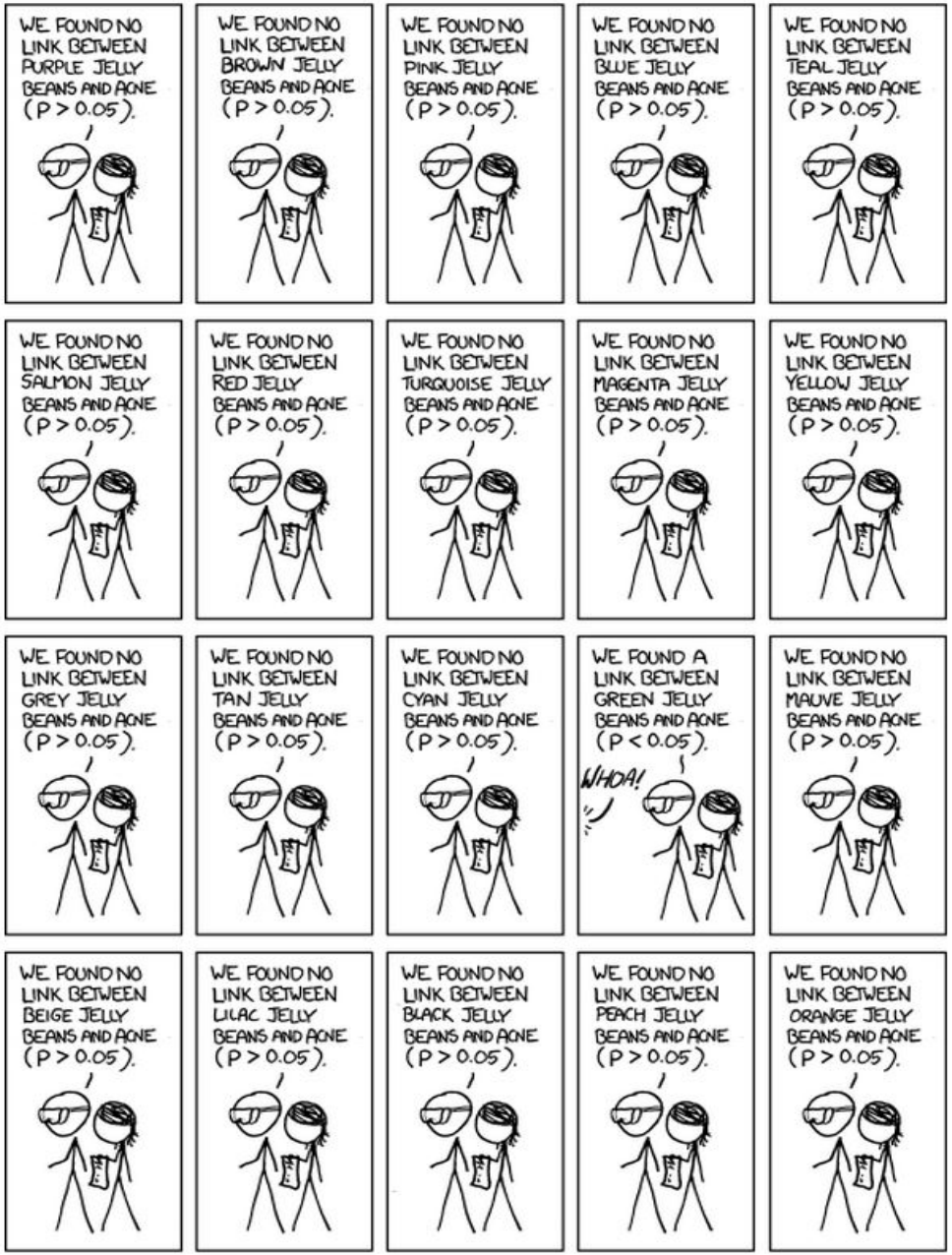
Tall people, Bonferroni

- 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?

IDC HERZLIYA

Bonferroni

What if two colors were linked at 0.05? Three?



IDC HERZLIYA

False Discovery Rate (FDR)

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

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

IDC HERZLIYA

Benjamini and Hochberg 1995

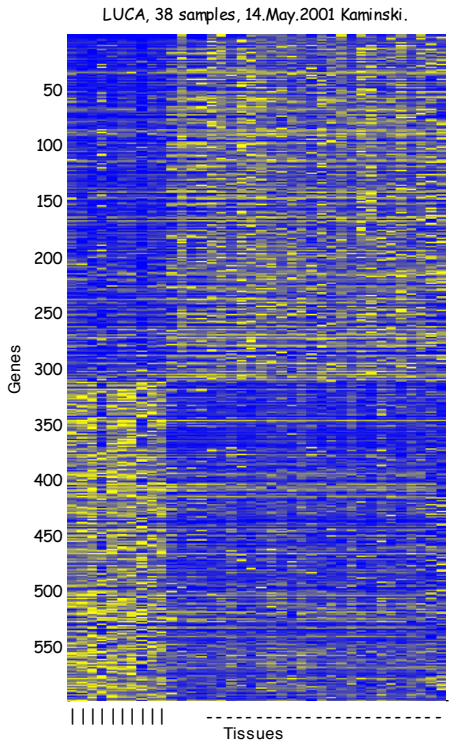
FDR - cont

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

  $FDR(i) = p(i)*N / i$

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

In the context of DGE …



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

IDC HERZLIYA

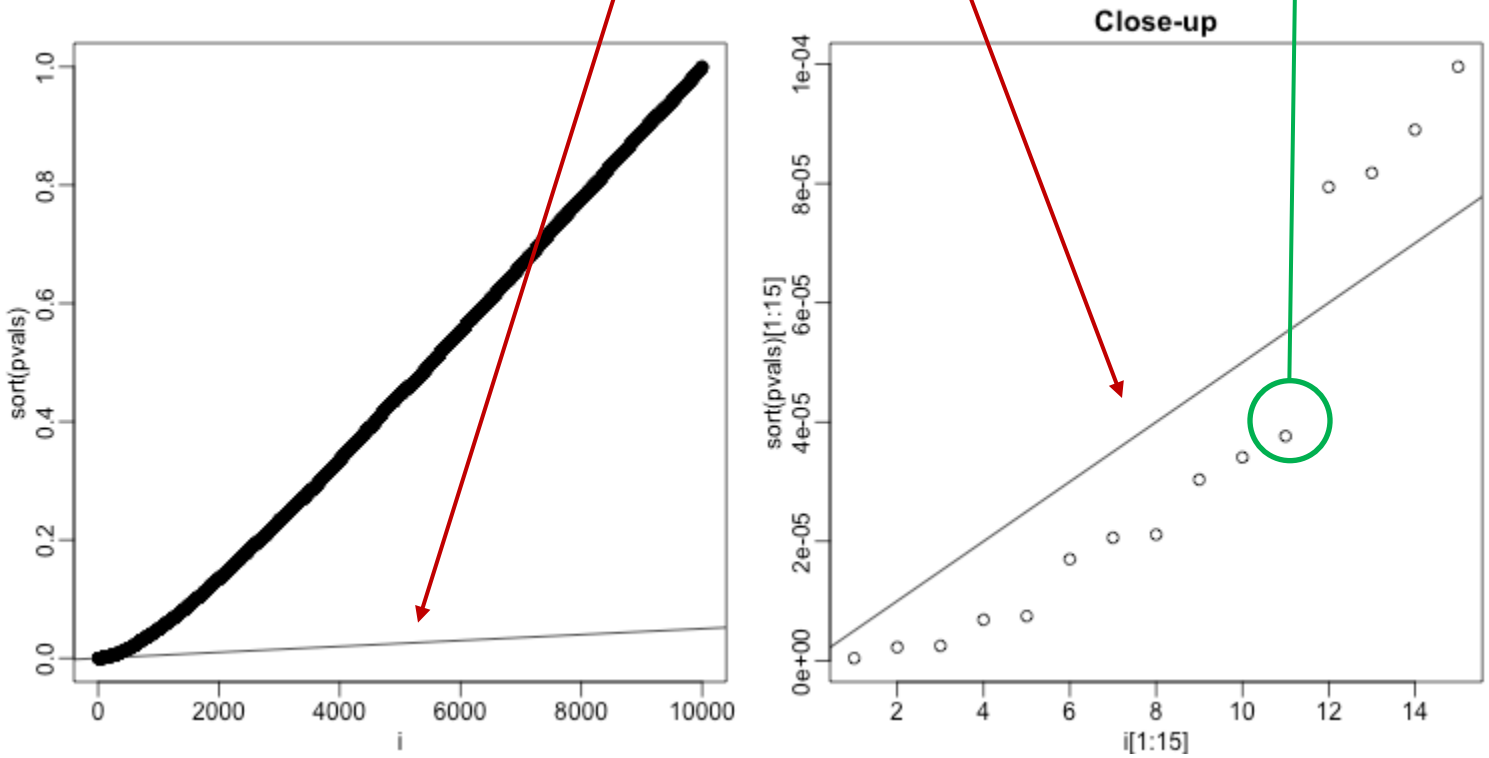
FDR illustrated

N = 10^4

$\tau = 0.05$

$y = \tau * (i/N)$

max i for which
$p(i)*N/i \leq \tau$



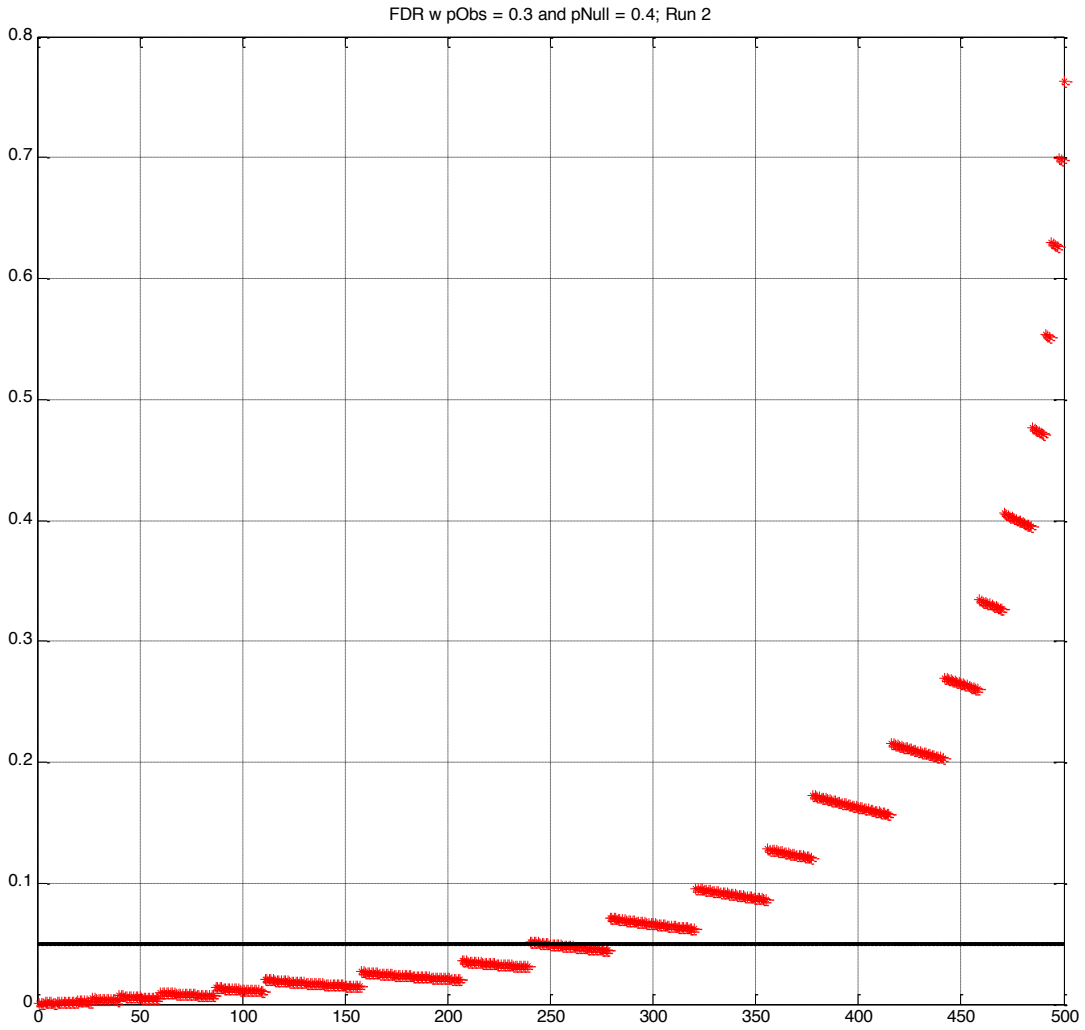
Zohar Yakhini, Doron Lipson

FDR, ex 7

Generate w p = 0.3
And compute FDR
against the null of
p = 0.4

>> F005 = nnz(FDR<0.05)

F005 = 270



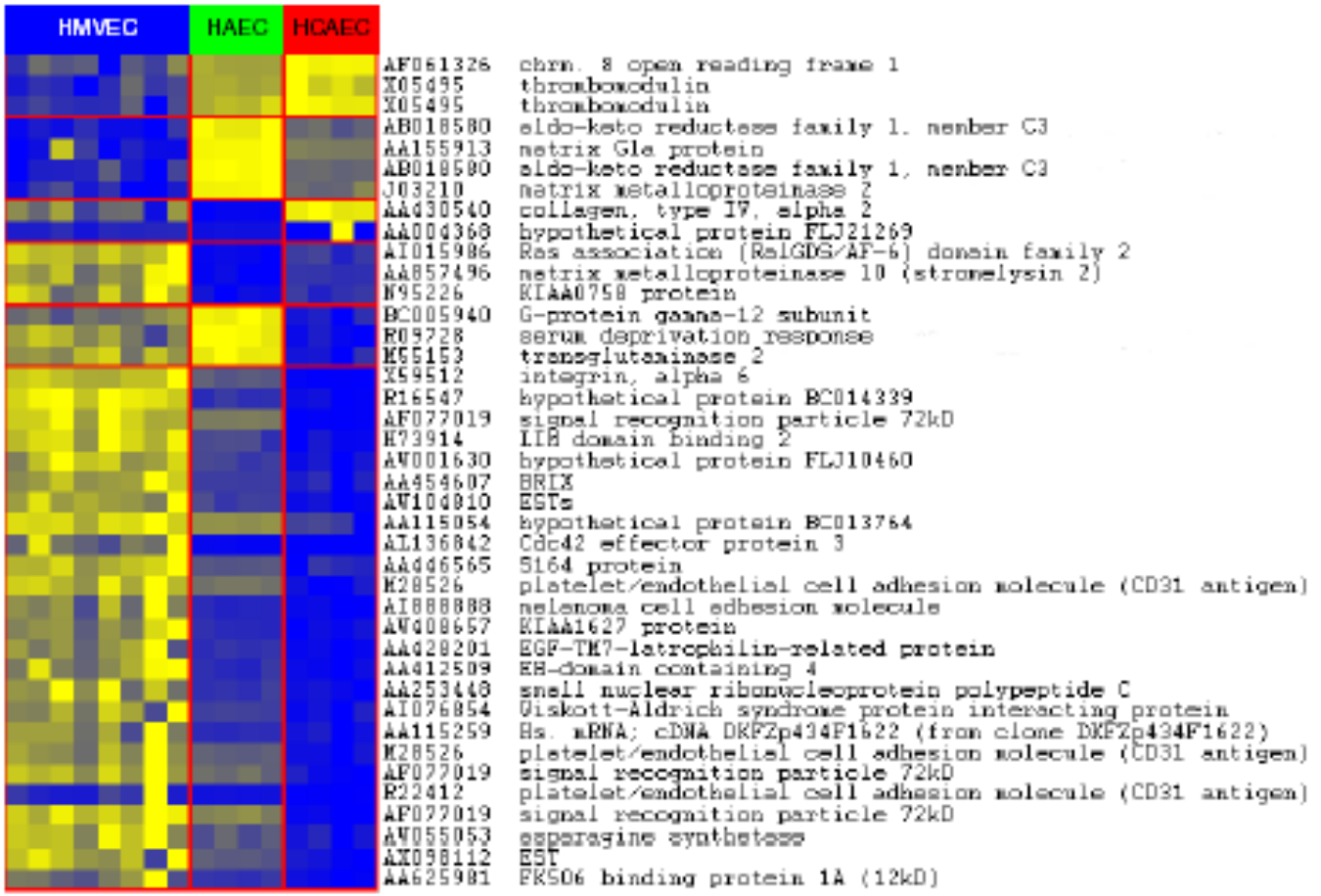
IDC HERZLIYA

A more subtle correction

$$\text{RobFDR}(i) = \min_{j \geq i} \left(p(j) * N / j \right)$$

IDC HERZLIYA

Multi-class DE scores

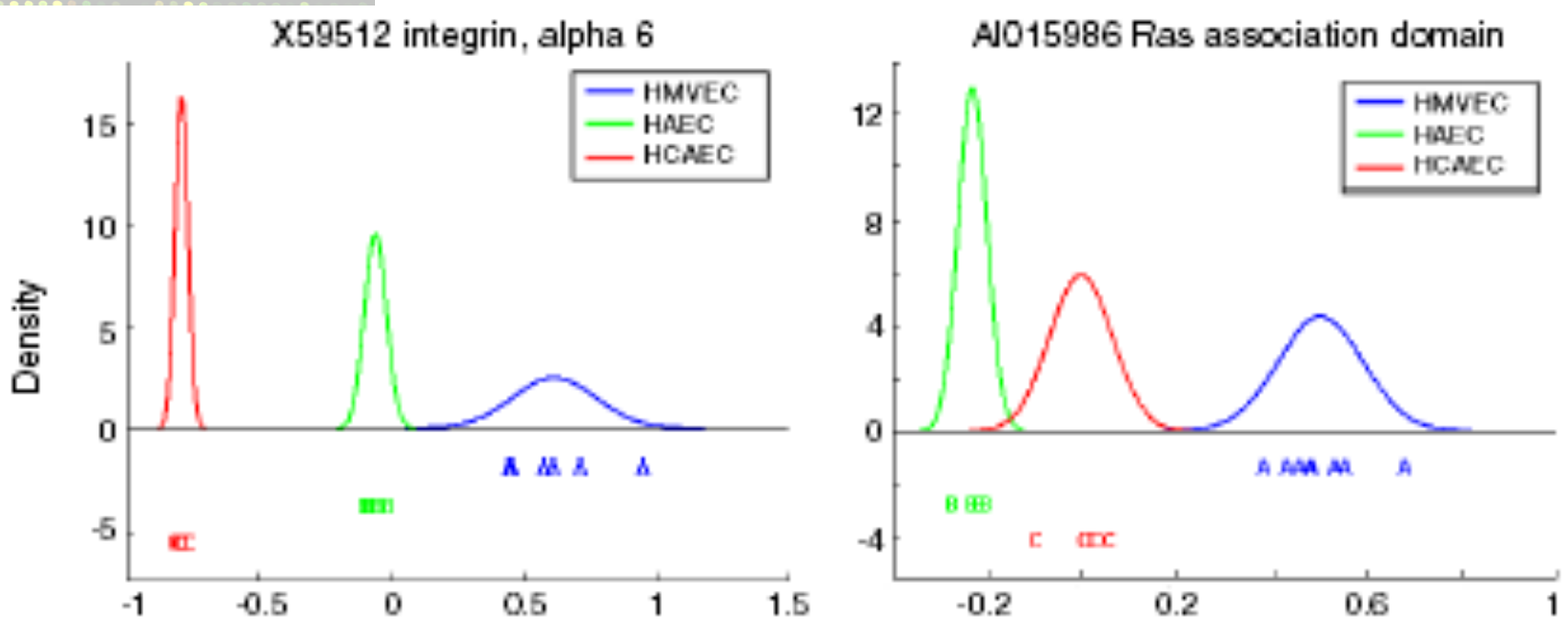


IDC HERZLIYA

Zohar Yakhini, Itai Sharon

Differential Expression in Endothelial Cells

(Ho et al, Phys Genomics 2003)



IDC HERZLIYA

Zohar Yakhini, Itai Sharon

Multi-class DE

ANOVA:

Compare

Variance between classes to Variance within classes

Gaussian Error Score:

Compute, for each tissue, t, of tissue class Γ(t) its error score

$$Err_g(t) = 1 - \frac{p(e_g(t) \mid \mu_{\Gamma(t)}, \sigma_{\Gamma(t)})\, p(\Gamma(t))}{\sum_{\Gamma} p(e_g(t) \mid \mu_{\Gamma}, \sigma_{\Gamma})\, p(\Gamma)}$$

Where $e_g(t)$ is the expression level of the gene g in tissue t, $p(\Gamma)$ is the prior probability of the class Γ and $p(x \mid \mu_{\Gamma}, \sigma_{\Gamma})$ is the best Gaussian density fit to the expression values of g in the class Γ.

The Gaussian Error score of g is now defined by

$$GER(g) = \sum_{j=1}^{m} Err_g(t_j)$$

Wilcoxon Rank Sum test

Consider two samplesets of independently acquired observations from two different labels.
Say safety test results for cars manufactured in the Randomistan VW factory vs those made in the German factory.

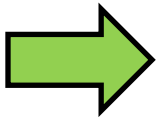
German factory	Stochastic Heights factory
3.2	3.7
4.5	8.5
8.1	6.1
9.9	9.3
4.1	9.1
7.3	4.3
5.2	7.2
6.0	
nG = 8	nG = 7

IDC HERZLIYA

Wilcoxon Rank Sum test

Null assumption:
When considering samples from both factories then all rank configurations are equiprobable.

German factory	Stochastic Heights factory
3.2	3.7
4.5	8.5
8.1	6.1
9.9	9.3
4.1	9.1
7.3	4.3
5.2	7.2
6.0	
nG = 8	nG = 7



9.9	G
9.3	R
9.1	R
8.5	R
8.1	G
7.3	G
7.2	R
6.1	R
6.0	G
5.2	G
4.5	G
4.3	R
4.1	G
3.7	R
3.2	G

N = nG + nR = 15

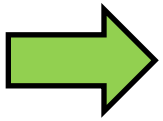
IDC HERZLIYA

Zohar Yakhini, Doron Lipson

Wilcoxon Rank Sum test

Null model, abstracted:
All N choose B binary configurations in
{0,1} ^(N,B)
are equiprobable.
Let T = sum of the ranks of the entries labeled 1.

9.9	G	1	0
9.3	R	2	1
9.1	R	3	1
8.5	R	4	1
8.1	G	5	0
7.3	G	6	0
7.2	R	7	1
6.1	R	8	1
6.0	G	9	0
5.2	G	10	0
4.5	G	11	0
4.3	R	12	1
4.1	G	13	0
3.7	R	14	1
3.2	G	15	0



N = nG + nR = 15

N = 15, B = 7
T = 50

IDC HERZLIYA

Wilcoxon Rank Sum test

Under the null model we have

$$E(T) = 56$$

The result here points to higher (lower numbers, better) ranks for R.

But is it significant?

1	0
2	1
3	1
4	1
5	0
6	0
7	1
8	1
9	0
10	0
11	0
12	1
13	0
14	1
15	0

N = 15, B = 7
T = 50

IDC HERZLIYA

Wilcoxon Rank Sum test

We want to compute

$P(T \leq 50)$,

under the null model.

1	0
2	1
3	1
4	1
5	0
6	0
7	1
8	1
9	0
10	0
11	0
12	1
13	0
14	1
15	0

$N = 15, B = 7$

$T = 50$

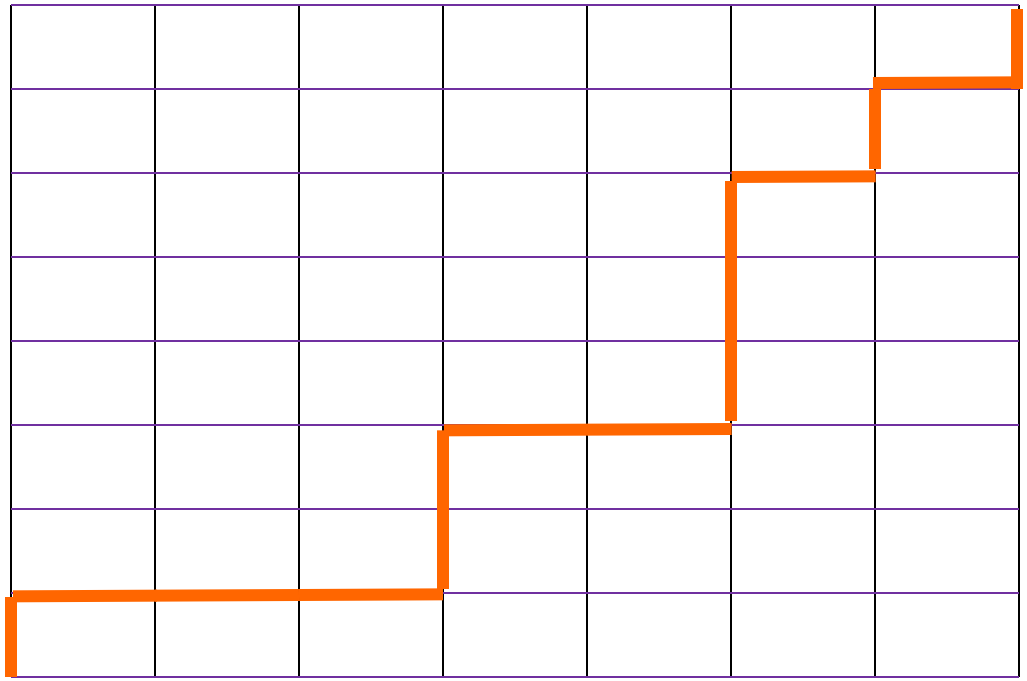
IDC HERZLIYA

Wilcoxon Rank Sum test; Exact p-values, the idea

$P_{Null}(T \leq 50)$?

Idea:
Map patterns to paths on the lattice
Use DP to count paths in which $T \leq 50$

0
1
1
1
0
0
1
1
0
0
0
1
0
1
0



IDC HERZLIYA

Wilcoxon Rank Sum test; Normal approximation

$P_{Null}(T \leq 50)$?

When B and N-B are sufficiently large … and depending on the desired accuracy

Let

$$\mu_T = \frac{B(N+1)}{2}$$

and

$$\sigma_T = \sqrt{\frac{B(N-B)(N+1)}{12}}$$

then

$$Z(T) = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

IDC HERZLIYA

In our case …

Using the normal approximation even though the numbers are not sufficiently large:

$Z \approx -0.7$

and therefore

$P_{Null}(T \leq 50) \approx 0.24$

and we do not have sufficient confidence for rejecting the hypothesis that the German factory is as good as the one in Stochastic Heights.

- Compare to internet calculators that compute the exact p, or Matlab, or equivalent.

IDC HERZLIYA

Computational Genomics

- Design and data analysis aspects are central to the science and practice of hybridization assays and of other molecular data intensive studies .
- Techniques:
  - Statistical benchmarking and combinatorics
  - Algorithmics
  - Optimization heuristics
  - Linear algebra
  - Graph theory
- There is growing science and technology interest in developing new methods and tools and in solving more design and analysis problems.
- Many of the computational aspects are also applicable in the context of more advanced measurement approaches such as sequencing.