

# MULTIPLE COMPARISON

ARNOLDO FRIGESSI  
FRIGESSI@MEDISIN.UIO.NO  
UNIVERSITY OF OSLO

(With material from lecture of T Speed and A Davison.)



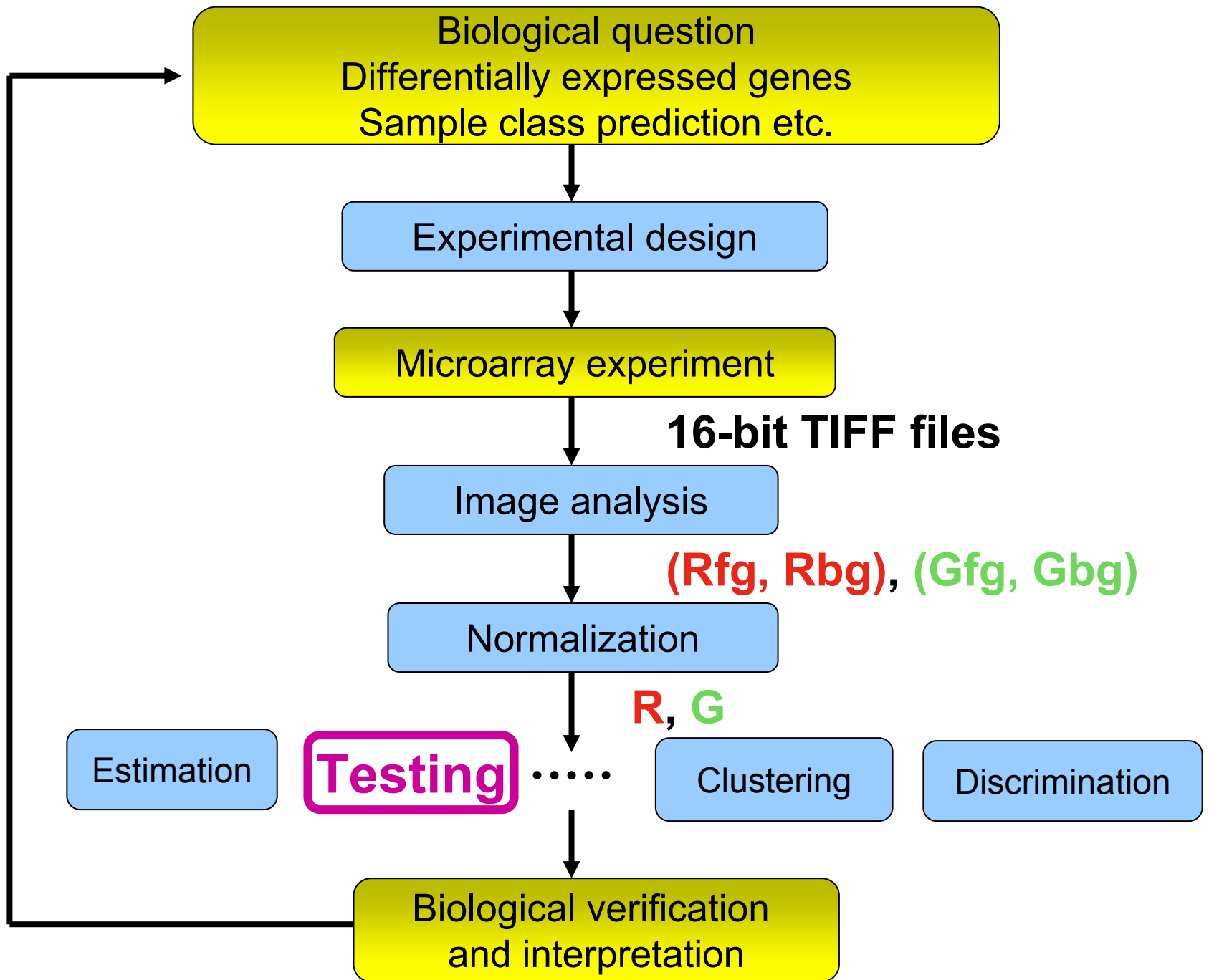
**EMBio** Styringsgruppen for forskning innen  
molekylærbiologi, bioteknologi og bioinformatikk ved UiO



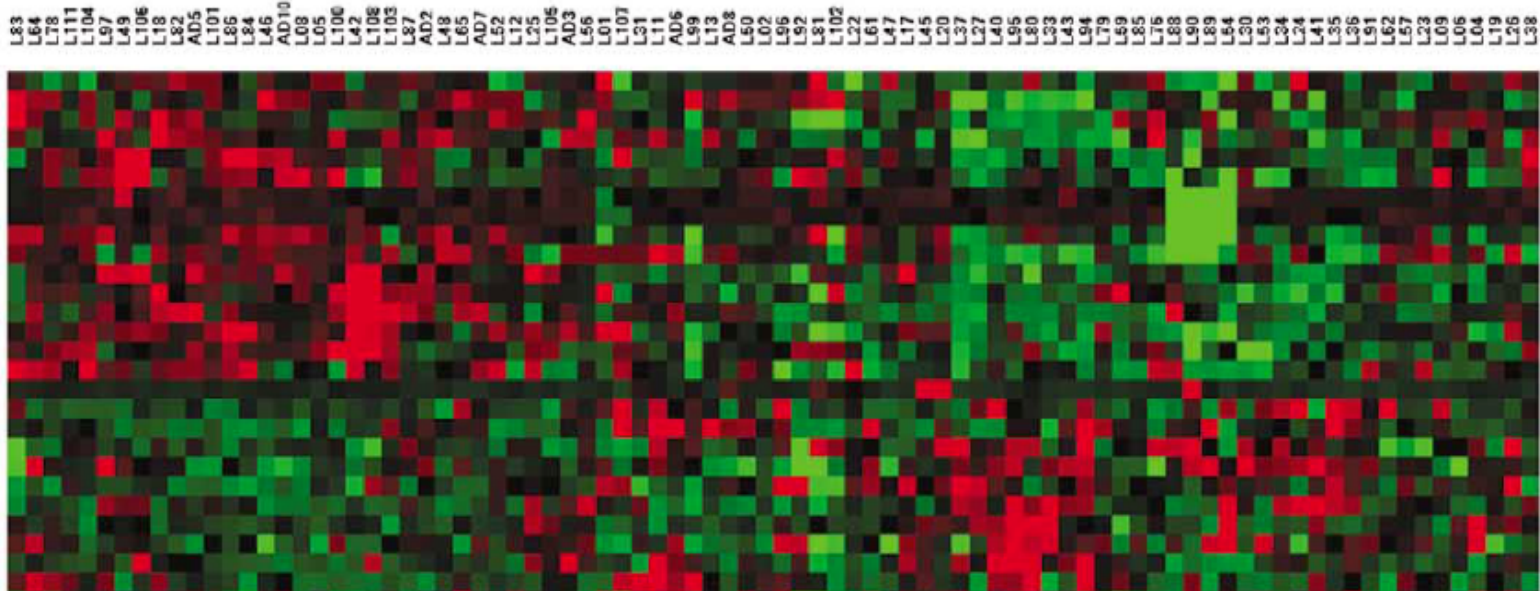
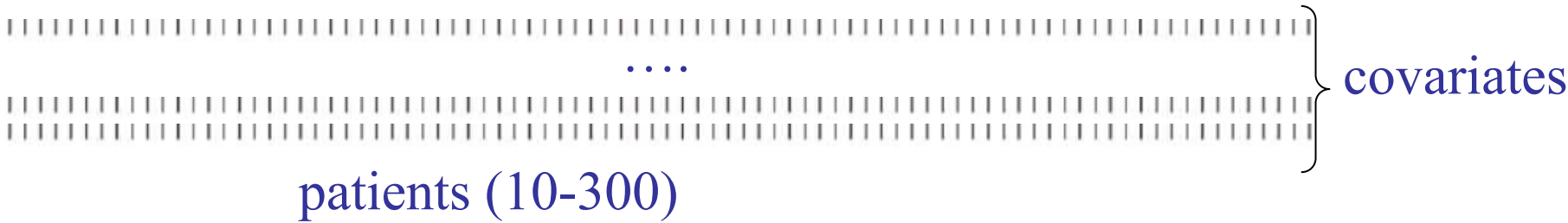
(sfi)<sup>2</sup>

**Statistics for Innovation**





# DATA



genes  
(20000)

- number of months to recovery
- yes/no distant metastases within 5 years
- other endpoints: censoring, end of study etc.

# THE PROBLEM IS UNKNOWN DEPENDENCE.

Huge amounts of simultaneous comparisons are necessary.

*Find differences between two (or more) varieties*

- *find differentially expressed genes*
- *find SNP patterns associated with one variety but not with the other*

Each single comparison is easy to do.

Problem:

Tests are dependent (co-regulation), but we do not know the dependency structure.

*The effective number of independent tests is unknown.*

Difficult to control multiplicity!

## SPARSITY OF THE SOLUTION

Often, we expect that only a small subset of comparisons will have a positive result: the solution is very sparse in the huge parameter space.

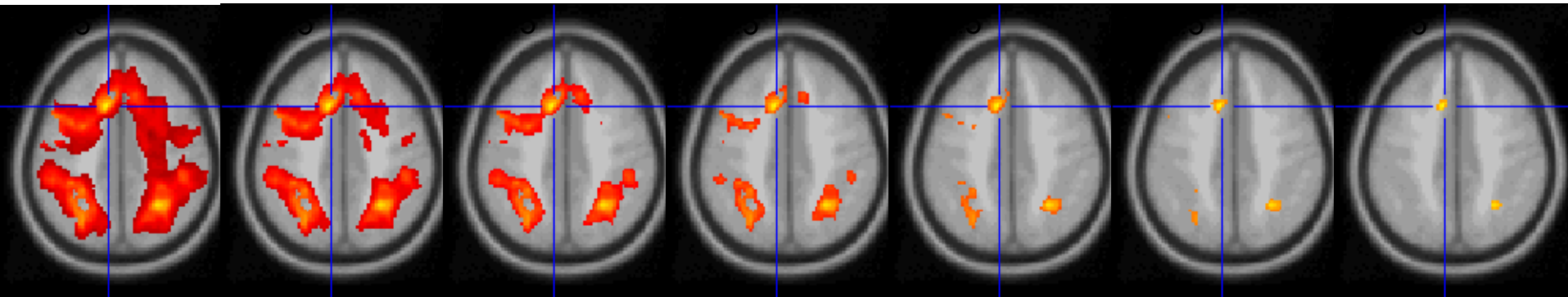
- FDR controls unstructured sparsity.
- Sometimes, additional information is available on structure of the sparse solution.

Then, one can develop methods that

- exploit available a priori knowledge,
- merge different data sets, each adding information.

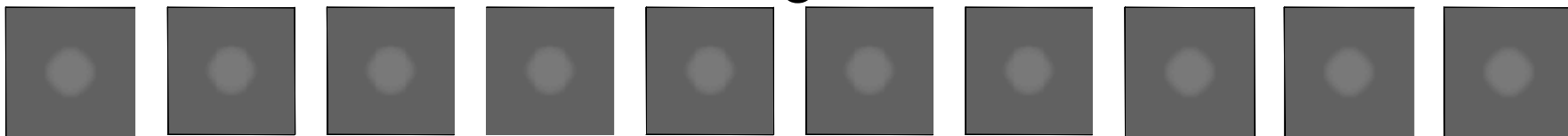
# fMRI & Multiple Comparisons

- Massively Univariate Modeling
  - Fit a model at each volume element (voxels)
  - Create images showing statistical significance
- Which of 100,000 voxels are significant?
  - For  $\alpha=0.05 \Rightarrow 5,000$  false positives!

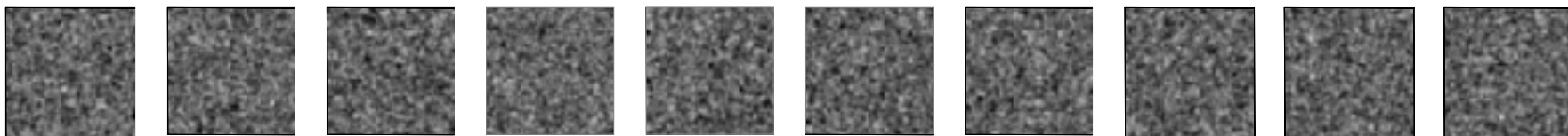


# Illustration:

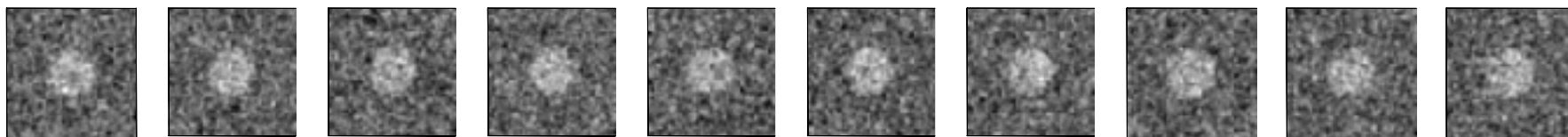
Signal



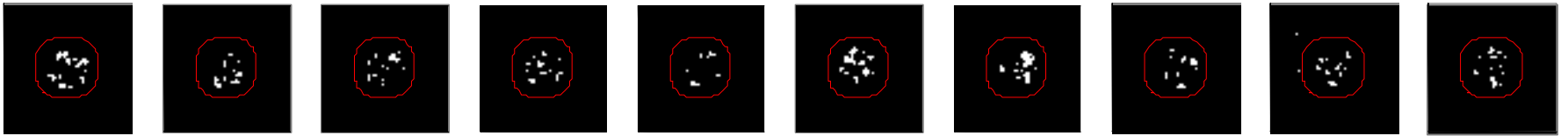
Noise



data = Signal+Noise

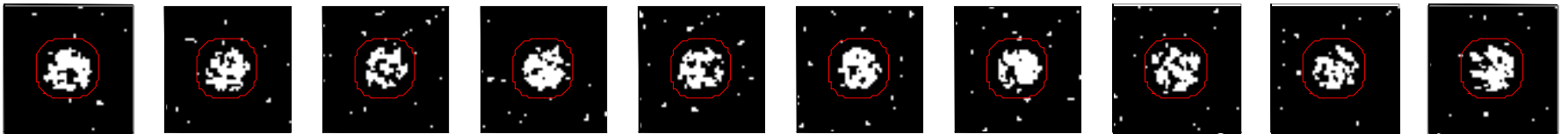


## Controlling by Bonferroni at $\alpha = 0.1$



very conservative, few false positives, many false negatives

## Controlling False Discovery Rate at $\alpha = 0.1$



More false positives, much less false negatives (lost voxels)

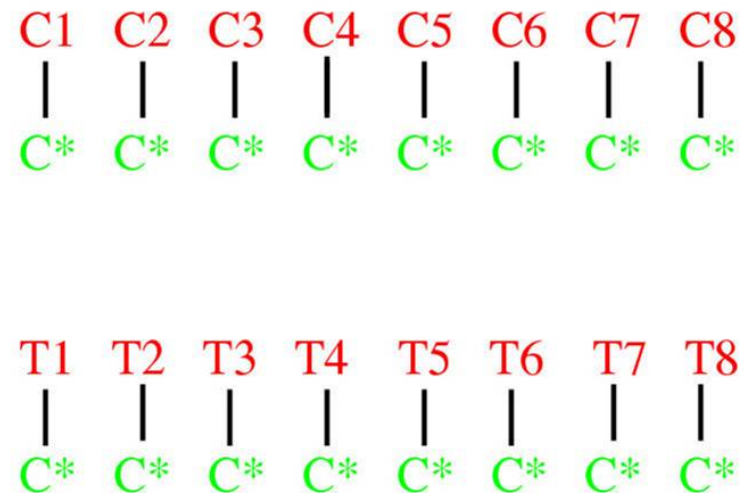


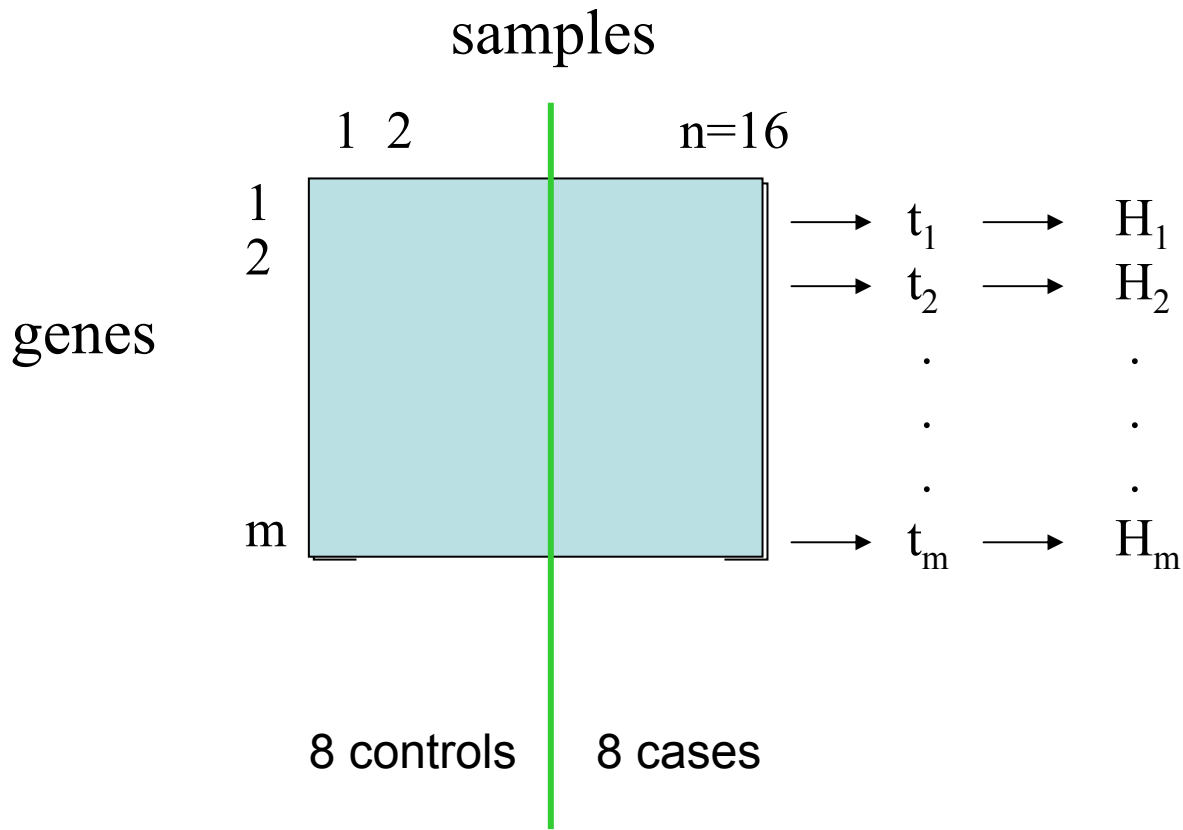
# Apo A1 experiment

(Matt Callow, Berkley; Circulation 1995)

**Goal:** To identify genes with altered expression in the livers of Apo A1 knock-out mice (T) compared to inbred control mice (C).

- 8 treatment and 8 control mice
- 16 hybridizations: liver mRNA from each of the 16 mice ( $T_i$ ,  $C_i$ ) labelled with Cy5, while pooled liver mRNA from the control mice ( $C^*$ ) is labelled with Cy3.
- Probes: ~ 6,000 cDNAs (genes), including 200 related to lipid metabolism.





$$M = \log_2\left(\frac{r_{ij}}{g_{ij}}\right)$$

## Which genes have changed?

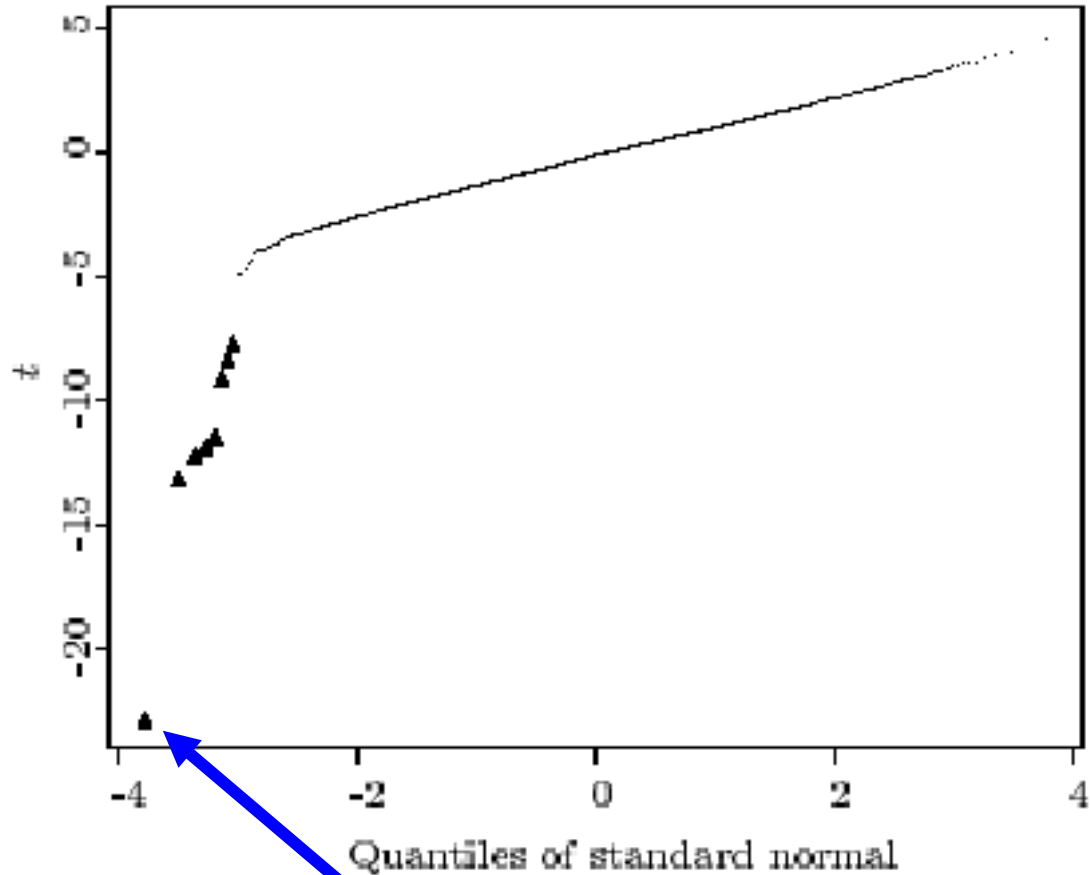
1. For each gene and each hybridization (8 ko + 8 ctl), use  $M = \log_2(R/G)$ .

2. For each gene compute the  $t$ -statistic:

$$\frac{\text{average of 8 ko Ms} - \text{average of 8 ctl Ms}}{\text{sqrt}(1/8 (\text{SD of 8 ko Ms})^2 + (\text{SD of 8 ctl Ms})^2)}$$

3. Do a normal  $qq$ -plot; look for values “off the line”.

# Normal *qq*-plot of *t*-statistics



**ApoA1**

# (What is a normal qq-plot?)

We have a random sample, say  $t_i, i=1, \dots, n$ , which we believe might come from a normal distribution.

If it did, then for suitable  $\mu$  and  $\sigma$ ,  $\Phi((t_i - \mu)/\sigma), i=1, \dots, n$  would be uniformly distributed on  $[0, 1]$ , where  $\Phi$  is the standard normal distribution. This means that for the order statistics of the  $t$ -sample,  $t_{(1)} < t_{(2)} < \dots < t_{(n)}$ ,  $\Phi((t_{(i)} - \mu)/\sigma)$  should be approximately equal to  $i/n$ . This is the same as saying that we expect  $t_{(i)}$  to be equal to  $\sigma\Phi^{-1}(i/n) + \mu$ .

Thus if we plot  $t_{(i)}$  against  $\Phi^{-1}(i/n)$ , we might expect to see a straight line of slope about  $\sigma$  with intercept about  $\mu$ .

This is our normal q-q plot.

When some of the samples do not seem to be on that line, these data points might NOT arise from the normal distribution.

# Why do a normal q-q plot?

One of the things we want to do with our  $t$ -statistics is to identify the *extreme* ones.

It is natural to rank them, but how extreme is extreme?

Converting ranked  $t$ 's into a normal qq-plot is a great way to **see the extremes: they are the ones that are “off the line”, at one end or another**. This technique is particularly helpful when we have thousands of values. Of course we can't expect all differentially expressed genes to stand out as extremes: many will be masked by even more extreme random variation.

## First 12 Largest T-Statistics

T-Statistic
-20.6
-12.5
-11.9
-11.7
-11.4
-11.3
-7.8
-7.4
5.0
-4.5
-4.5
-4.4

### Gene annotation

Apo AI

EST, weakly sim. to STEROL DESATURASE

CATECHOL O-METHYLTRANSFERASE

Apo CIII

EST, highly sim. to Apo AI

EST

Highly sim. to Apo CIII precursor

similar to yeast sterol desaturase

1. The t-statistics were ranked according to their absolute values.

**Discoveries are further studied; negative results are usually ignored**

# Steps to find diff. expressed genes

- 1. Formulate a single hypothesis testing strategy**
- 2. Construct a statistic for each gene**
- 3. Compute the raw p-values for each gene by permutation procedures or from (normal) distribution models**
- 4. Consider the multiple testing problem**
  - a. Find the maximum # of genes of interest**
  - b. Assign a significance level for each gene**



# ***p*-values**

The ***p*-value**  $p$  is the probability of getting a test statistic as or more extreme than the observed one, under the null hypothesis  $H$  of no differential expression.

In order to compute the *p*-value, we need to know the distribution of the statistics  $t$ .

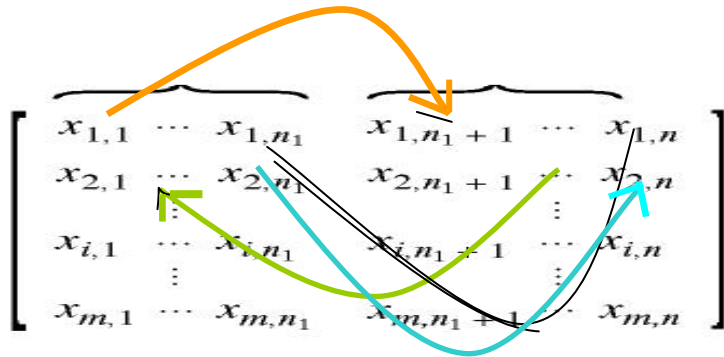
If we would have very many independent samples, then asymptotic theory would assure that  $t$  is *t*-distributed.

It is however unwise to assume that its null distribution is that of *Student's t*. We have another way to assign *p*-values: using permutations.

# Computing $p$ -values by permutations

We focus on one gene only. For the  $b$ th iteration,  $b = 1, \dots, B$ ;

1. Permute the  $n$  data points for the gene ( $x$ ). The first  $n_1$  are referred to as “treatments”, the second  $n_2$  as “controls”.



This is a wrong data set, but if there was no difference between treatments and controls, it would be ok.

2. For each gene, calculate the corresponding two sample t-statistic,  $t^b$ .

After all the  $B$  permutations are done;

3. Put  $p = \#\{b: |t_b| \geq |t_{real\ data}|\} / B$

# All permutations? How many permutations?

Combinatorics: too many!

Minimum number of permutations: equal to the square of the number of objects to be permuted.

A default value: I use the cube of the number of objects to be permuted, for example.

# Sequential Permutation

is an old idea of Besag and Clifford.

The method continues to sample until the sampled statistics  $T$  is  $w = 20$  of times larger (or smaller, depending on null hypothesis) than the observed value of the same statistics.

The choice of  $w$  can be changed and is critical.

Thus we use more times for large sampled statistics (which lead to small p-values) and do rapidly when the test is not leading to significance.

Sequential MC produces p-values that can be adjusted by FDR.

It is an open research problem to find optimal values of  $w$ , as this would depend on the critical value for significance.

Too large values of  $w$  would imply a too precise estimation of p-values which are very small. Too small values, would make the estimate inaccurate.

### First 12 Largest T-Statistics

Neglecting multiplicity issues, i.e. working at the individual 0.05 level, would identify, on the average,  $6000 \times 0.05 = 300$  differentially expressed genes, even if really no such gene exists.

Doing Bonferroni adjustment leads to 8 differentially expressed genes.

<b>T-Statistic</b>	<b>P-Value</b>
-20.6	$7.0 \times 10^{-12}$
-12.5	$5.6 \times 10^{-9}$
-11.9	$1.1 \times 10^{-8}$
-11.7	$1.3 \times 10^{-8}$
-11.4	$1.8 \times 10^{-8}$
-11.3	$1.9 \times 10^{-8}$
-7.8	$1.8 \times 10^{-6}$
-7.4	$3.6 \times 10^{-6}$
5.0	$1.8 \times 10^{-4}$
-4.5	$4.6 \times 10^{-4}$
-4.5	$4.9 \times 10^{-4}$
-4.4	$6.5 \times 10^{-4}$

1. The t-statistics were ranked according to their absolute values.

# Many tests: what is the problem?

## Simulation to illustrate it.

Example: assume we have 30 000 independent genes on a microarray and not a single gene is truly differentially expressed.

If we reject the null hypothesis at level 0.01, we still expect  $30000 \times 0.01 = 300$  to have by chance a p-value below 0.01.

We create a simulated data set, where nothing is differentially expressed, and then we compute the t statistics and the p-values. No gene should be found as differentially expressed.

Simulation of 6,000 genes with 8 treatments and 8 controls.

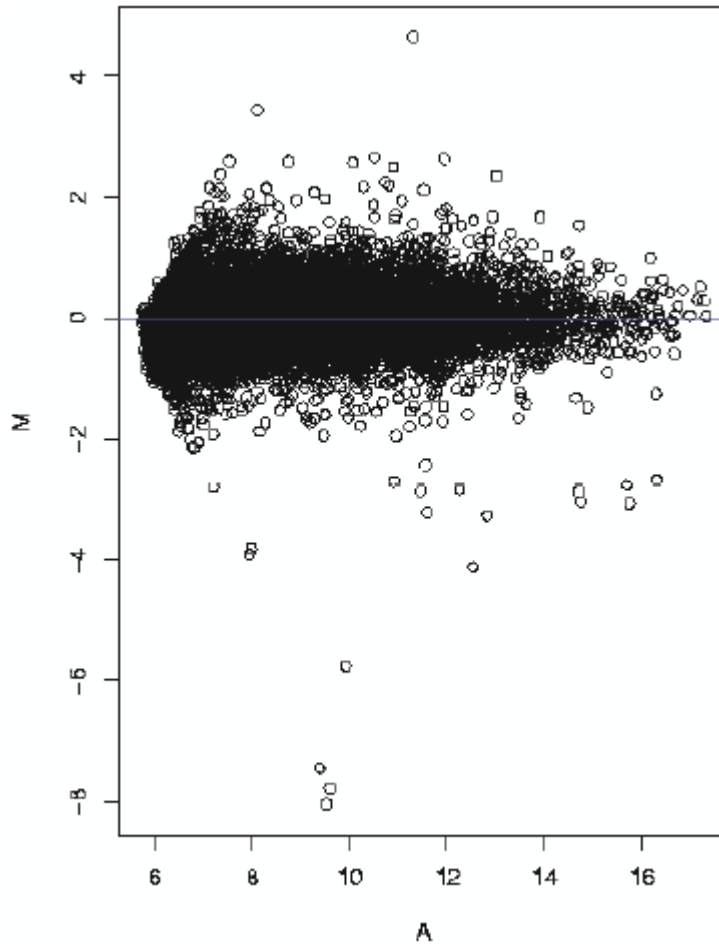
All the gene expression values were simulated *i.i.d* from a  $N(0,1)$  distribution, i.e. **NOTHING** is differentially expressed in our simulation.

We show the 10 smallest permutation p-values.

# Unadjusted p-values

<b>gene index</b>	<b>t value</b>	<b>p-value (unadj.)</b>
<b>2271</b>	<b>4.93</b>	<b><math>2 \times 10^{-4}</math></b>
<b>5709</b>	<b>4.82</b>	<b><math>3 \times 10^{-4}</math></b>
<b>5622</b>	<b>-4.62</b>	<b><math>4 \times 10^{-4}</math></b>
<b>4521</b>	<b>4.34</b>	<b><math>7 \times 10^{-4}</math></b>
<b>3156</b>	<b>-4.31</b>	<b><math>7 \times 10^{-4}</math></b>
<b>5898</b>	<b>-4.29</b>	<b><math>7 \times 10^{-4}</math></b>
<b>2164</b>	<b>-3.98</b>	<b><math>1.4 \times 10^{-3}</math></b>
<b>5930</b>	<b>3.91</b>	<b><math>1.6 \times 10^{-3}</math></b>
<b>2427</b>	<b>-3.90</b>	<b><math>1.6 \times 10^{-3}</math></b>
<b>5694</b>	<b>-3.88</b>	<b><math>1.7 \times 10^{-3}</math></b>

Clearly we can't just use standard p-value thresholds of 0.05 or 0.01.



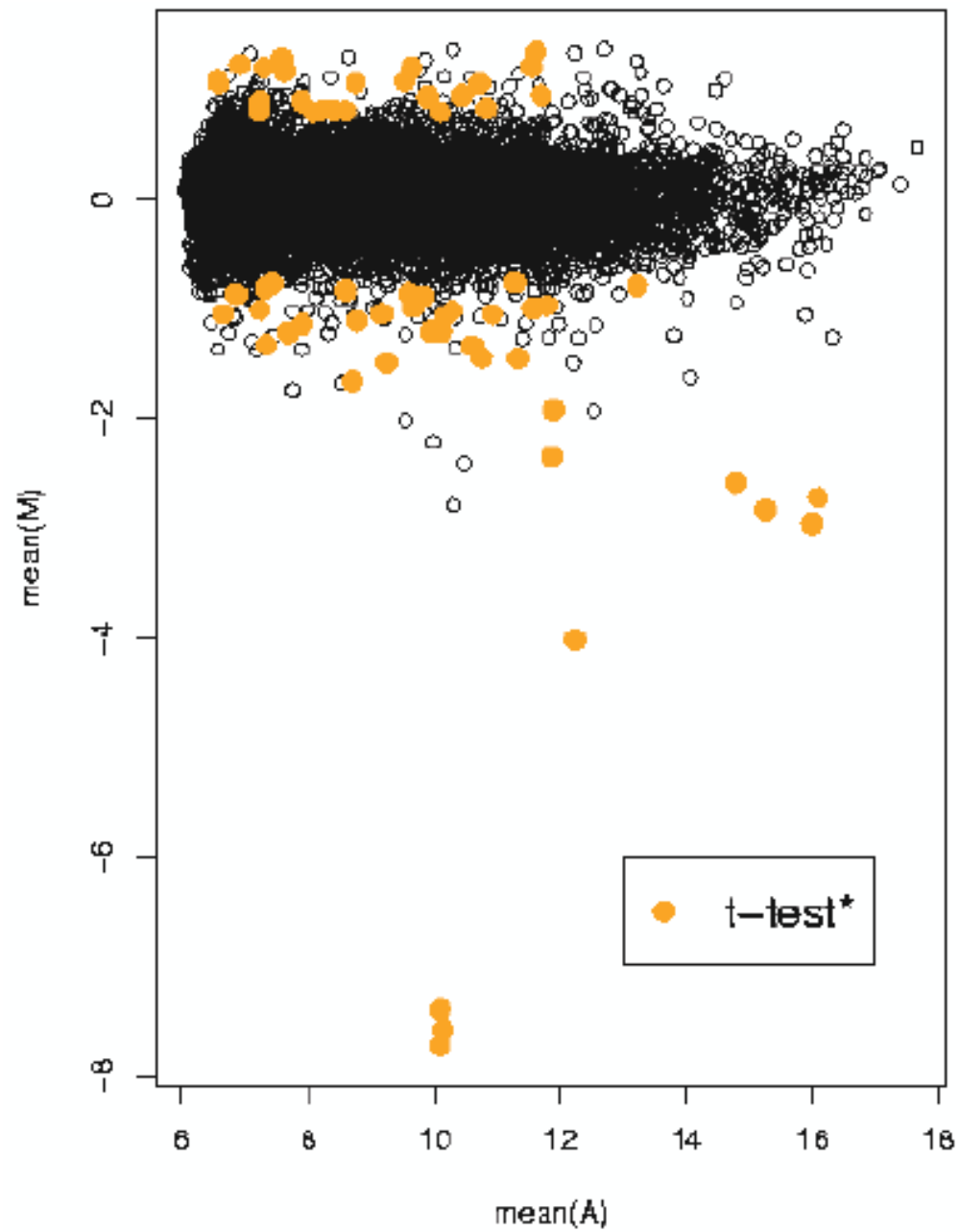
MA plot  
T-test

$$M = \log_2(Cy5/Cy3) = \log_2 Cy5 - \log_2 Cy3$$

$$A = \log_2 \sqrt{Cy5 \times Cy3} = (\log_2 Cy5 + \log_2 Cy3)/2$$

**Test**  $H_{0,g}$  :  $g$ th gene is not differentially expressed.  
 **$g = 1, \dots, G = 9279$**





# Multiple testing: Counting errors

Testing  $m$  genes:  $H^1, H^2, \dots, H^m$ .

$m_0 = \#$  of null hypotheses which are true

$R = \#$  of rejected null hypotheses

# Hypothesis Truth vs. Decision

Decision \ Truth	# not rejected	# rejected	totals
# true H	U	V	$m_0$
# non-true H	T	S	$m_1$
totals	$m - R$	R	$m$

V = # Type I errors [false positives]

T = # Type II errors [false negatives]

# Type I (False Positive) Error Rates

- Per-family Error Rate

$$\text{PFER} = E(V)$$

- Per-comparison Error Rate

$$\text{PCER} = E(V)/m$$

- Family-wise Error Rate

$$\text{FWER} = p(V \geq 1)$$

- False Discovery Rate

$$\text{FDR} = E(Q), \text{ where}$$

$$Q = V/R \text{ if } R > 0; Q = 0 \text{ if } R = 0$$

	# not rejected	# rejected	totals
# true H	U	V (F +)	$m_0$
# non-true H	T	S	$m_1$
totals	$m - R$	R	m

# Strong vs. Weak Control

- All probabilities are **conditional** on which hypotheses are true
- **Strong control** refers to control of the number of **F+**, ie. Type I error rate; under **any combination** of true and false null hypothesis
- **Weak control** refers to control of the number of **F+** only under the **complete null hypothesis** (i.e. **all** null hypothesis are simultaneously true, there is no interesting gene)
- In general, weak control without other safeguards is unsatisfactory

# Comparison of Type I Error Rates

- In general, for a given test and data set,

$$\text{PCER} \leq \text{FWER} \leq \text{PFER},$$

and

$$\text{FDR} \leq \text{FWER},$$

with  $\text{FDR} = \text{FWER}$  under the complete null

We will see that Bonferroni controls FWER.

# Adjusted p-values

$$\text{FWER} = \Pr(\# \text{ of false discoveries} > 0) = \Pr(V > 0)$$

- If interest is in controlling, e.g., the FWER, the **adjusted p-value** for hypothesis  $H_j$  is

$$p_j^*$$

such that if we reject hypothesis  $H_j$   
when  $p_j^* \leq \alpha$ , then overall FWER is equal to  $\alpha$

# Some Advantages of p-value Adjustment

- **Test level** (size)  $\alpha$  does not need to be determined in advance
- Some procedures **most easily described** in terms of their adjusted p-values
- Procedures can be **readily compared** based on the corresponding adjusted p-values



# A Little Notation

- For hypothesis  $H_j$ ,  $j = 1, \dots, m$   
observed test statistic:  $t_j$   
observed unadjusted p-value:  $p_j$
- Ordering of observed (absolute)  $t_j$ :  $\{r_j\}$   
such that  $|t_{r_1}| \geq |t_{r_2}| \geq \dots \geq |t_{r_G}|$
- Ordering of observed  $p_j$ :  $\{r_j\}$   
such that  $|p_{r_1}| \leq |p_{r_2}| \leq \dots \leq |p_{r_G}|$

# Control of the FWER

- **Bonferroni single-step** adjusted p-values

$$p_j^* = mp_j$$

- **Holm (1979) step-down** adjusted p-values

$$p_{r_j}^* = \max_{k=1 \dots j} \{ (m-k+1)p_{r_k} \}$$

- **Hochberg (1988) step-down** adjusted p-values (Simes inequality)

$$p_{r_j}^* = \min_{k=j \dots m} \{ (m-k+1)p_{r_k} \}$$

# Control of the FWER

- Westfall & Young (1993) step-down minP adjusted p-values

$$p_{r_j}^* = \max_{k=1 \dots j} \{ \text{prob} ( \max_{l \in \{r_k \dots r_m\}} P_l \leq p_{r_k} \mid H_0^{\text{COM}} ) \}$$

- Westfall & Young (1993) step-down maxT adjusted p-values

$$p_{r_j}^* = \max_{k=1 \dots j} \{ \text{prob} ( \max_{l \in \{r_k \dots r_m\}} |T_l| \geq |t_{r_k}| \mid H_0^{\text{COM}} ) \}$$

# Westfall & Young (1993)

## Adjusted p-values

- Step-down procedures: successively **smaller adjustments** at each step
- Take into account the **joint distribution** of the test statistics
- **Less conservative** than Bonferroni, Holm, or Hochberg adjusted p-values
- Can be estimated by **resampling** but computer-intensive (especially for minP)

<b>gene index</b>	<b>t statistic</b>	<b>unadj. p (<math>\times 10^4</math>)</b>	<b>minP adjust.</b>	<b><i>p</i>lower</b>	<b>maxT adjust.</b>
<b>2139</b>	<b>-22</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>2 \times 10^{-4}</math></b>
<b>4117</b>	<b>-13</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>5 \times 10^{-4}</math></b>
<b>5330</b>	<b>-12</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>5 \times 10^{-4}</math></b>
<b>1731</b>	<b>-11</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>5 \times 10^{-4}</math></b>
<b>538</b>	<b>-11</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>5 \times 10^{-4}</math></b>
<b>1489</b>	<b>-9.1</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>1 \times 10^{-3}</math></b>
<b>2526</b>	<b>-8.3</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>3 \times 10^{-3}</math></b>
<b>4916</b>	<b>-7.7</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b><math>8 \times 10^{-3}</math></b>
<b>941</b>	<b>-4.7</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b>0.65</b>
<b>2000</b>	<b>+3.1</b>	<b>1.5</b>	<b>.53</b>	<b><math>8 \times 10^{-5}</math></b>	<b>1.00</b>
<b>5867</b>	<b>-4.2</b>	<b>3.1</b>	<b>.76</b>	<b>0.54</b>	<b>0.90</b>
<b>4608</b>	<b>+4.8</b>	<b>6.2</b>	<b>.93</b>	<b>0.87</b>	<b>0.61</b>
<b>948</b>	<b>-4.7</b>	<b>7.8</b>	<b>.96</b>	<b>0.93</b>	<b>0.66</b>
<b>5577</b>	<b>-4.5</b>	<b>12</b>	<b>.99</b>	<b>0.93</b>	<b>0.74</b>

FDR



**EMBIO** Styringsgruppen for forskning innen  
molekylærbiologi, bioteknologi og bioinformatikk ved UiO



$(sfi)^2$

**Statistics for Innovation**



## False discovery rates (FDR)

$$Q = \frac{\# \text{ of false discoveries}}{\# \text{ of discoveries}} = \frac{V}{R}$$

**FDR = E(V/R | R>0)**

**We know R but not V!**

**FDR is not exactly computed, but estimated.**

Benjamini and Hochberg (1995)

## Estimate the FDR

Rank the p-values  $p_{r_1} \leq p_{r_2} \leq \dots \leq p_{r_m}$ .

The following adjusted p-values  $\pi_{r_i}$  control the FDR (when the unadjusted p-values  $p_i$  are **independently distributed**):

$$\pi_{r_i} = \min_{k \in \{i, \dots, m\}} \{ m p_{r_k} / k \}.$$

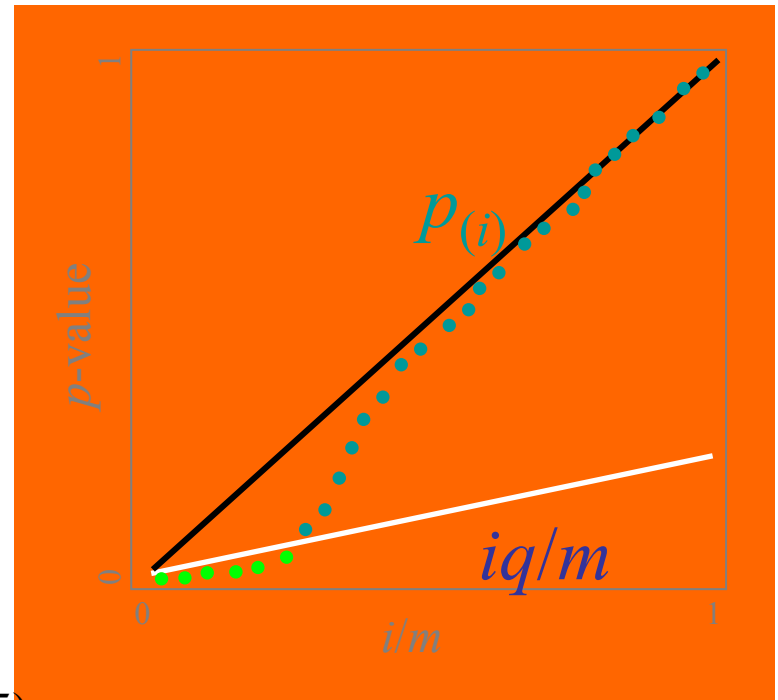


- Select desired limit  $q$  on FDR
- Order p-values,  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$
- Let  $r$  be largest  $i$  such that

$$p_{(i)} \leq i/m \cdot q$$

- Reject all hypotheses corresponding to  $p_{(1)}, \dots, p_{(r)}$ .

**This keeps the FDR  $\leq q$  under independence.**



# BRCA1 versus BRCA2-mutation positive tumours (Hedenfalk et al., 2001)

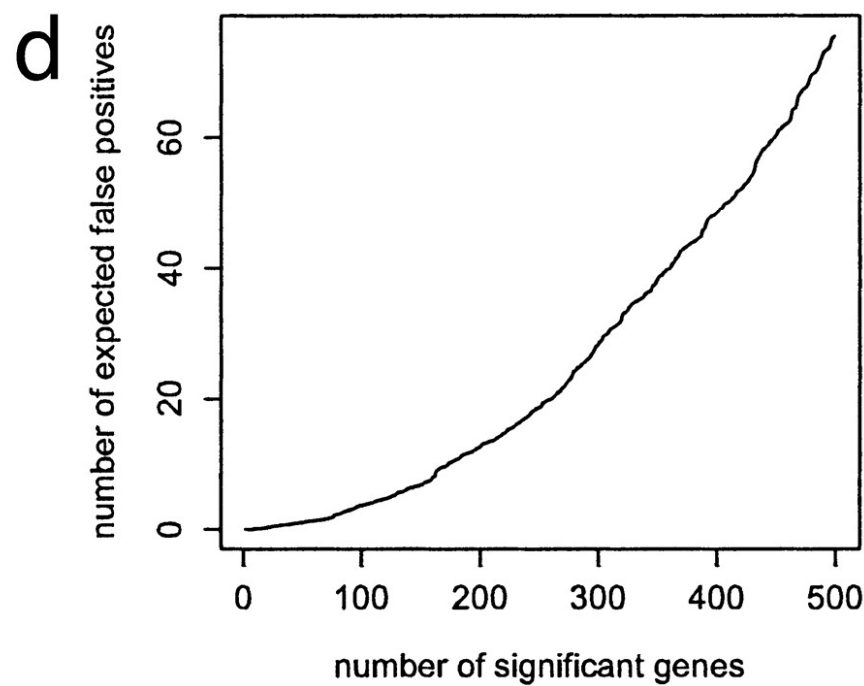
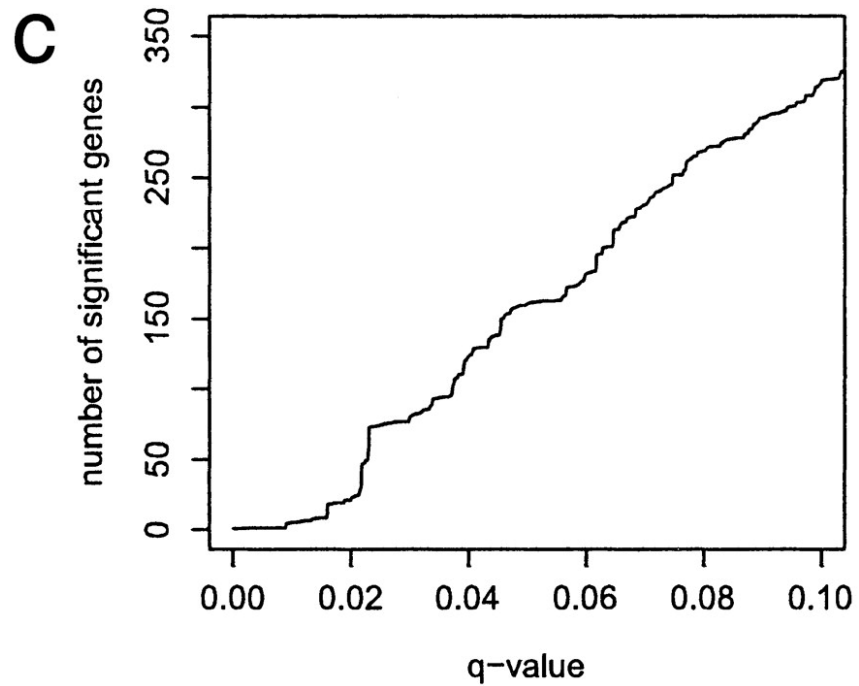
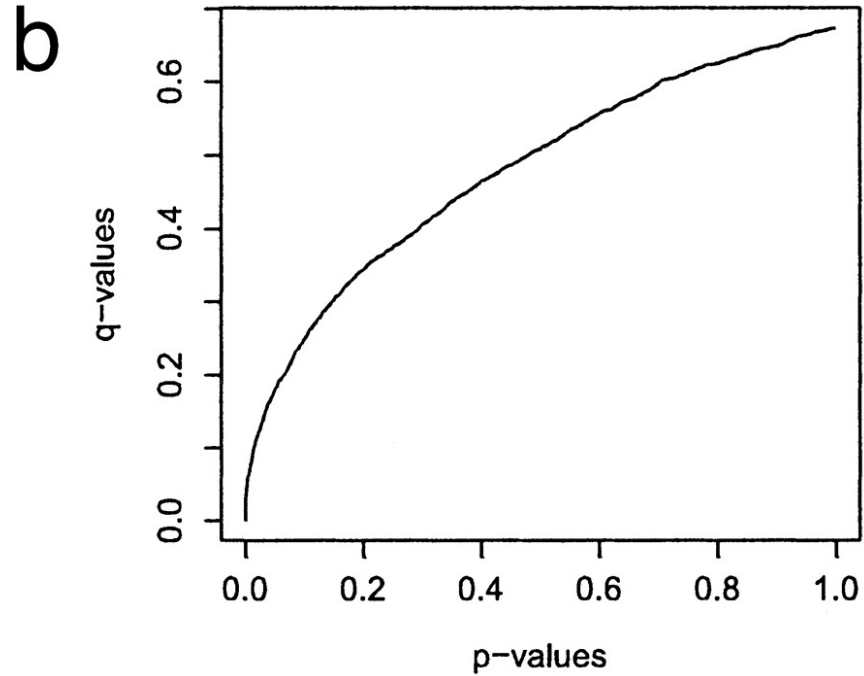
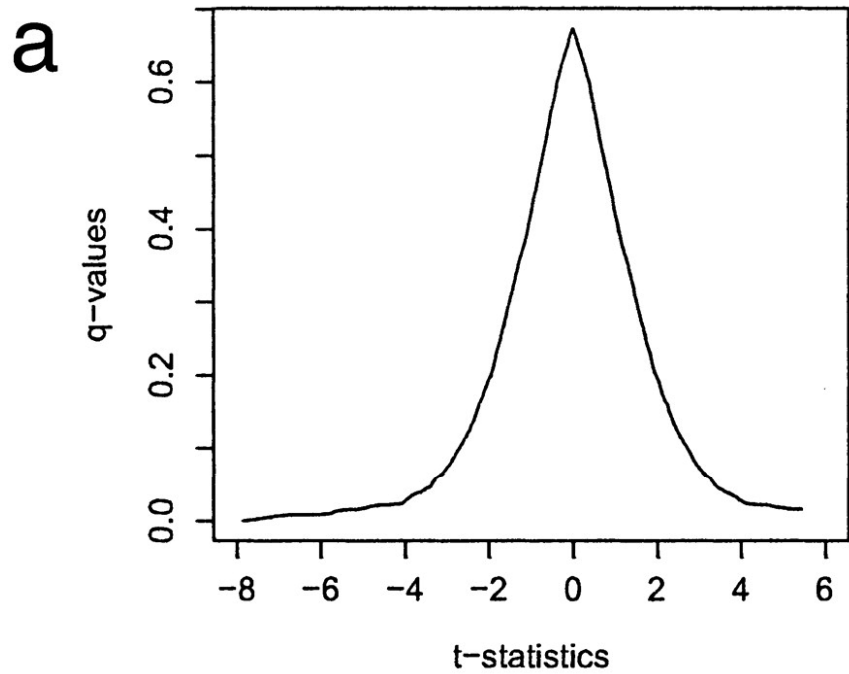
BRCA1 (7) versus BRCA2-mutation (8) positive tumours,  
 $p=3226$  genes

$P=.001$  gave 51 genes differentially expressed

$P=0.0001$  gave 9-11 genes

Using  $q<0.05$ , gives 160 genes taken to be significant.

It means that approx. 8 of these 160 genes are expected to be false positives.



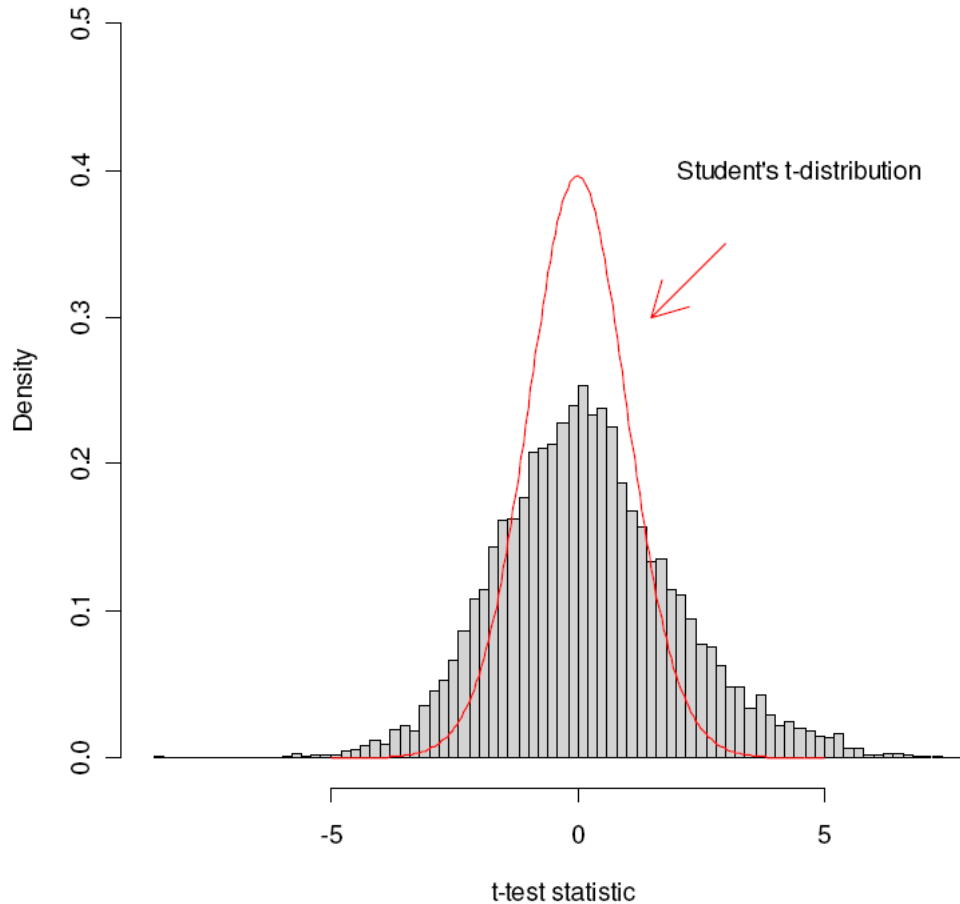
# Golub *et al* (1999) experiments

**Goal. To identify genes which are differentially expressed in acute lymphoblastic leukemia (ALL) tumours in comparison with acute myeloid leukemia (AML) tumours.**

- **38 tumour samples: 27 ALL, 11 AML.**
- **Data from Affymetrix chips, some pre-processing.**
- **Originally 6,817 genes; 3,051 after reduction.**

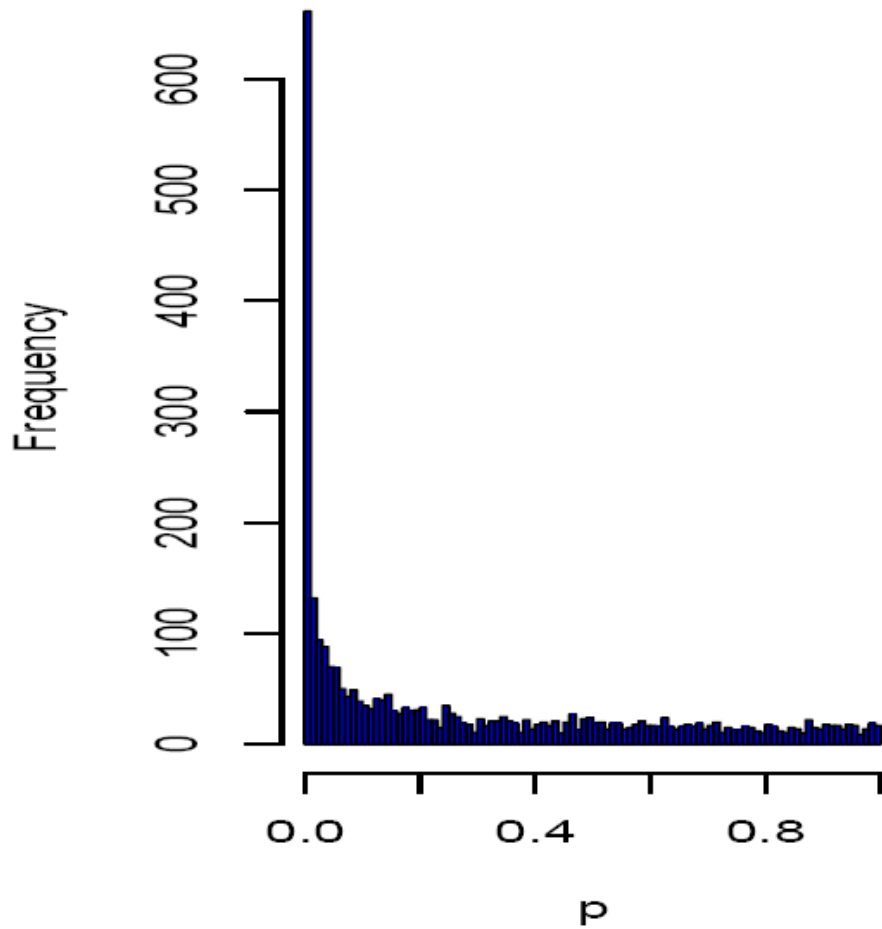
**Data therefore a  $3,051 \times 38$  array of expression values.**

# Golub *et al* (1999)

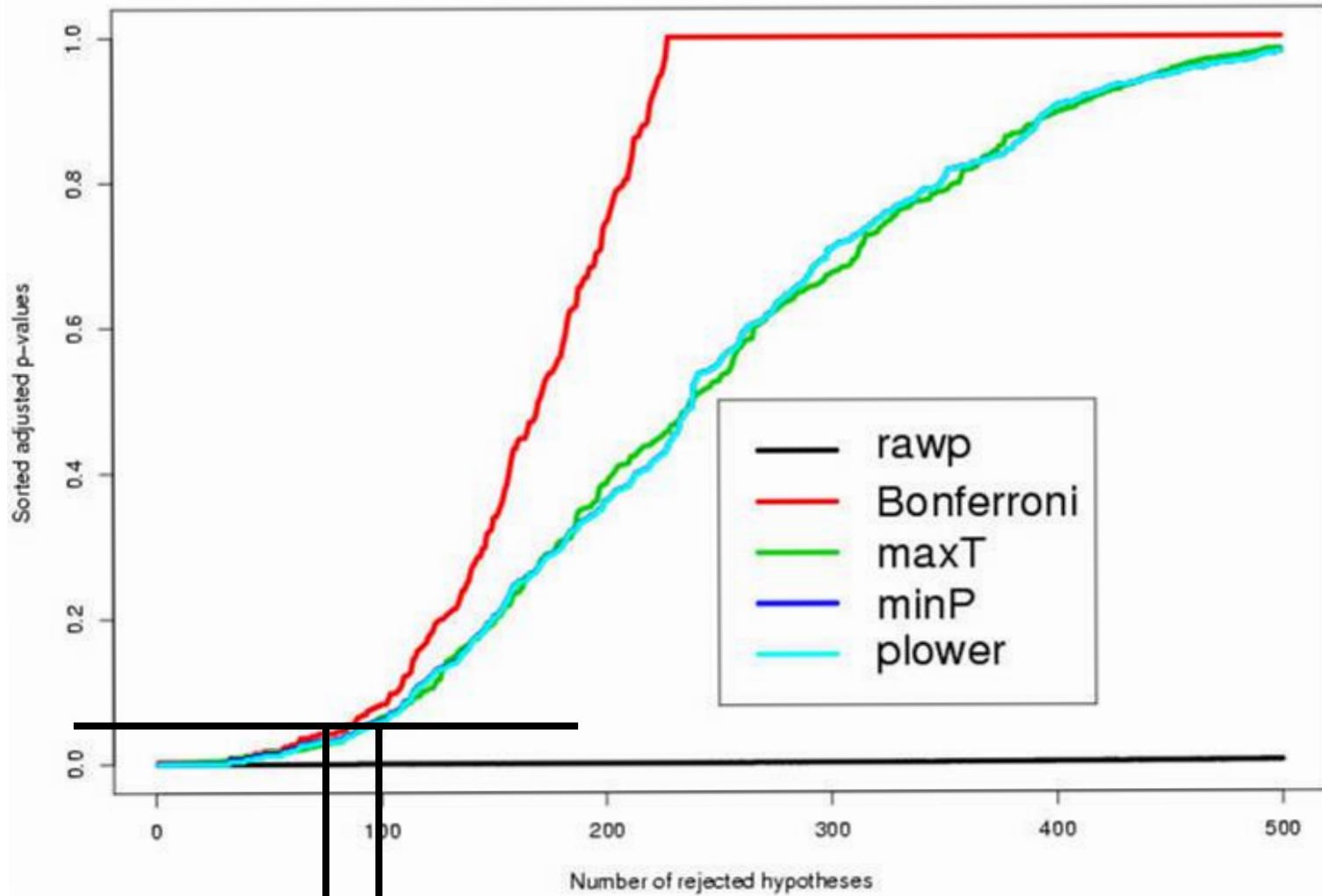


The empirical distribution of the  $t$ -test statistic, that compares the sample types AML and ALL. A comparison with the theoretical null distribution, Student's  $t$ -distribution, hints that there are too many extreme values to be accounted for by pure chance.

**histogram of p-values**



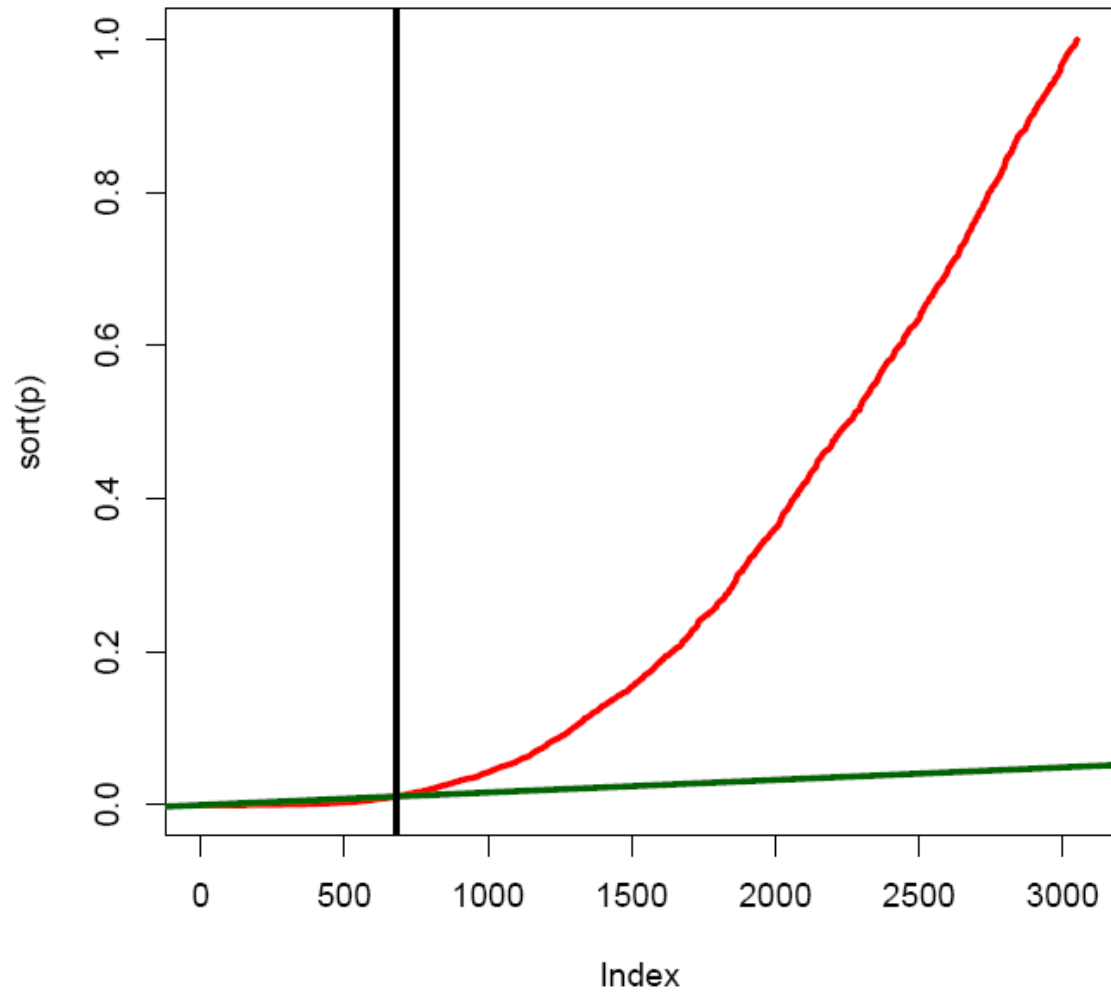
# Golub's data---1M simulations



Bonferroni

others

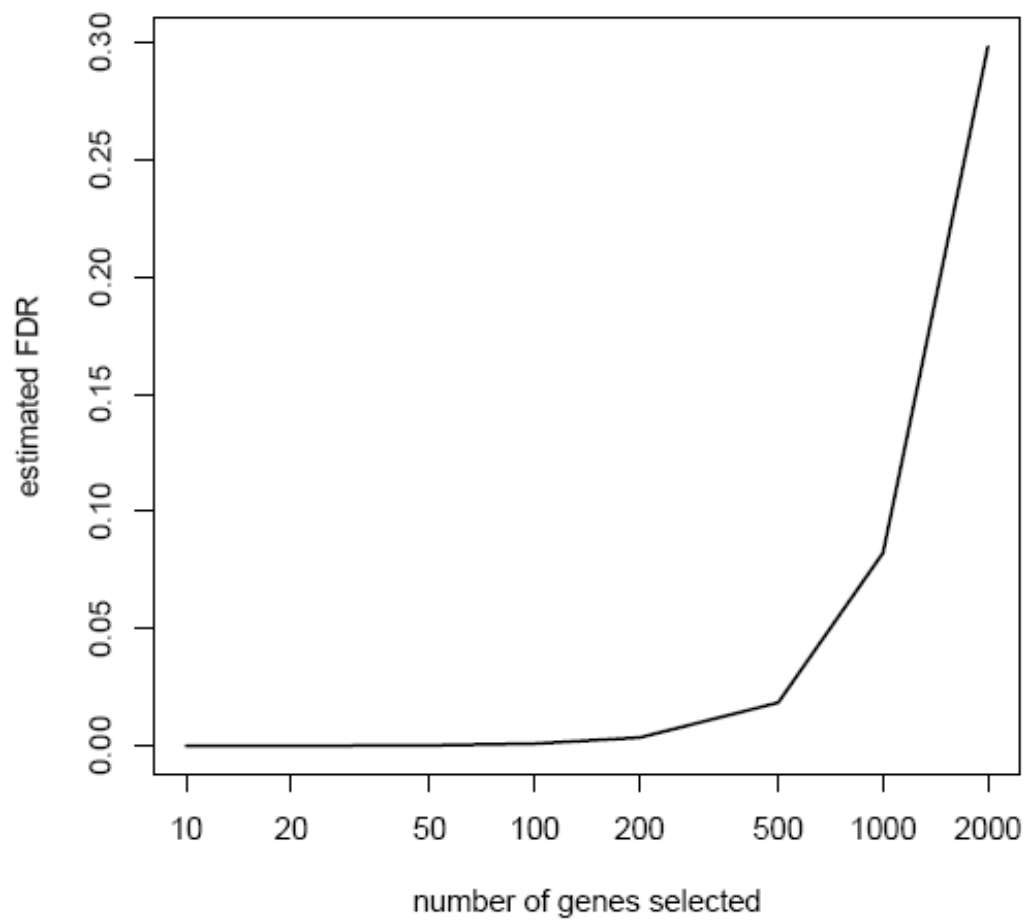
FDR



Golub data: 681 genes with BH-adjusted  $p < 0.05$ .



False discovery rate, Golub data



# Identification of Genes Associated with Survival

- Data: survival  $y_i$  and gene expression  $x_{ij}$  for individuals  $i = 1, \dots, n$  and genes  $j = 1, \dots, m$
- Fit Cox model for each gene singly:

$$h(t) = h_0(t) \exp(\beta_j x_{ij})$$

- For any gene  $j = 1, \dots, m$ , can test  $H_j: \beta_j = 0$
- Complete null  $H_0^{\text{COM}}: \beta_j = 0$  for all  $j = 1, \dots, m$
- The  $H_j$  are tested on the basis of the Wald statistics  $t_j$  and their associated p-values  $p_j$

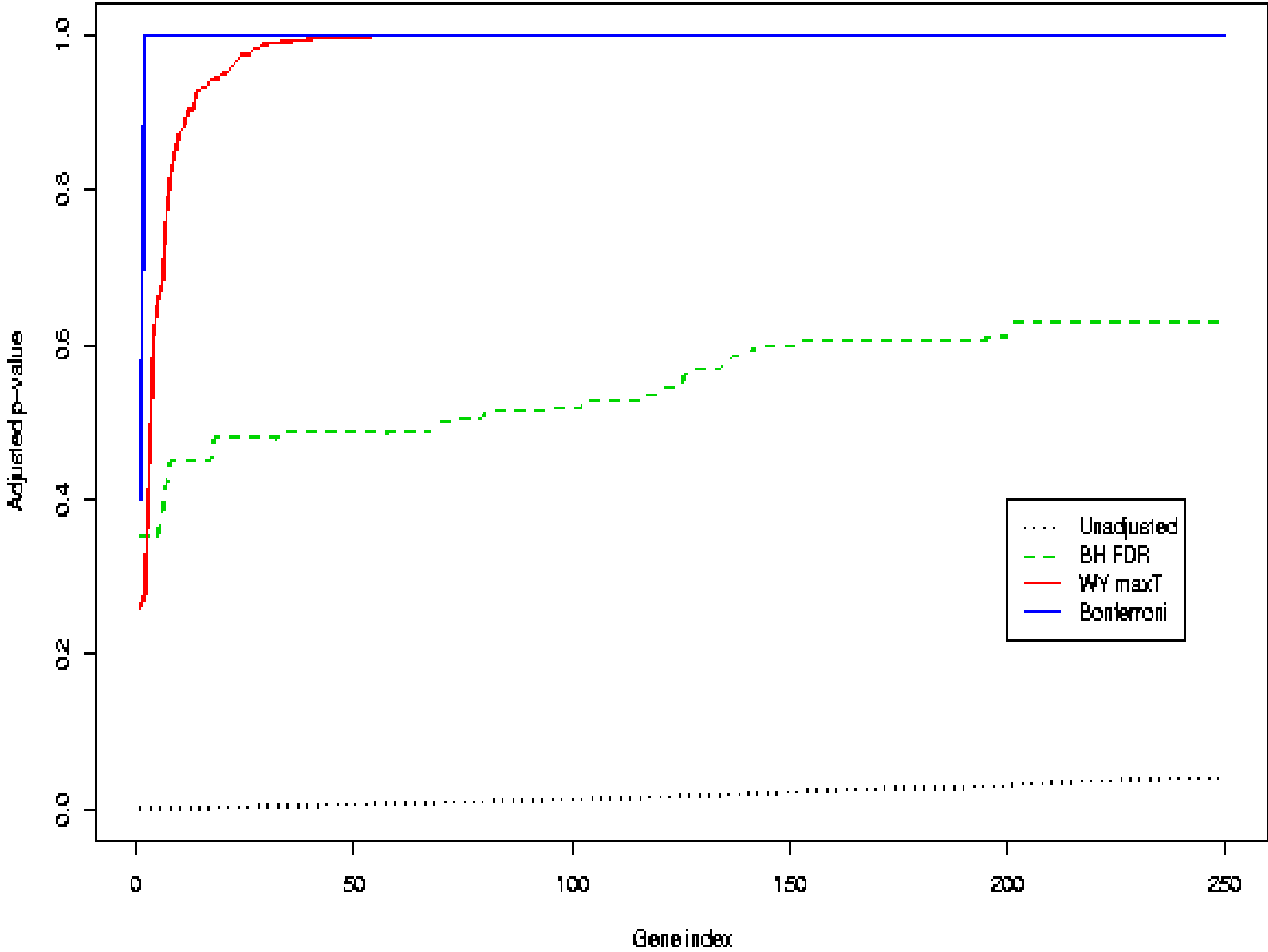
# Datasets

- **Lymphoma** (Alizadeh et al.)  
40 individuals, 4026 genes
- **Melanoma** (Bittner et al.)  
15 individuals, 3613 genes
- Both available at  
<http://lpgprot101.nci.nih.gov:8080/GEAW>

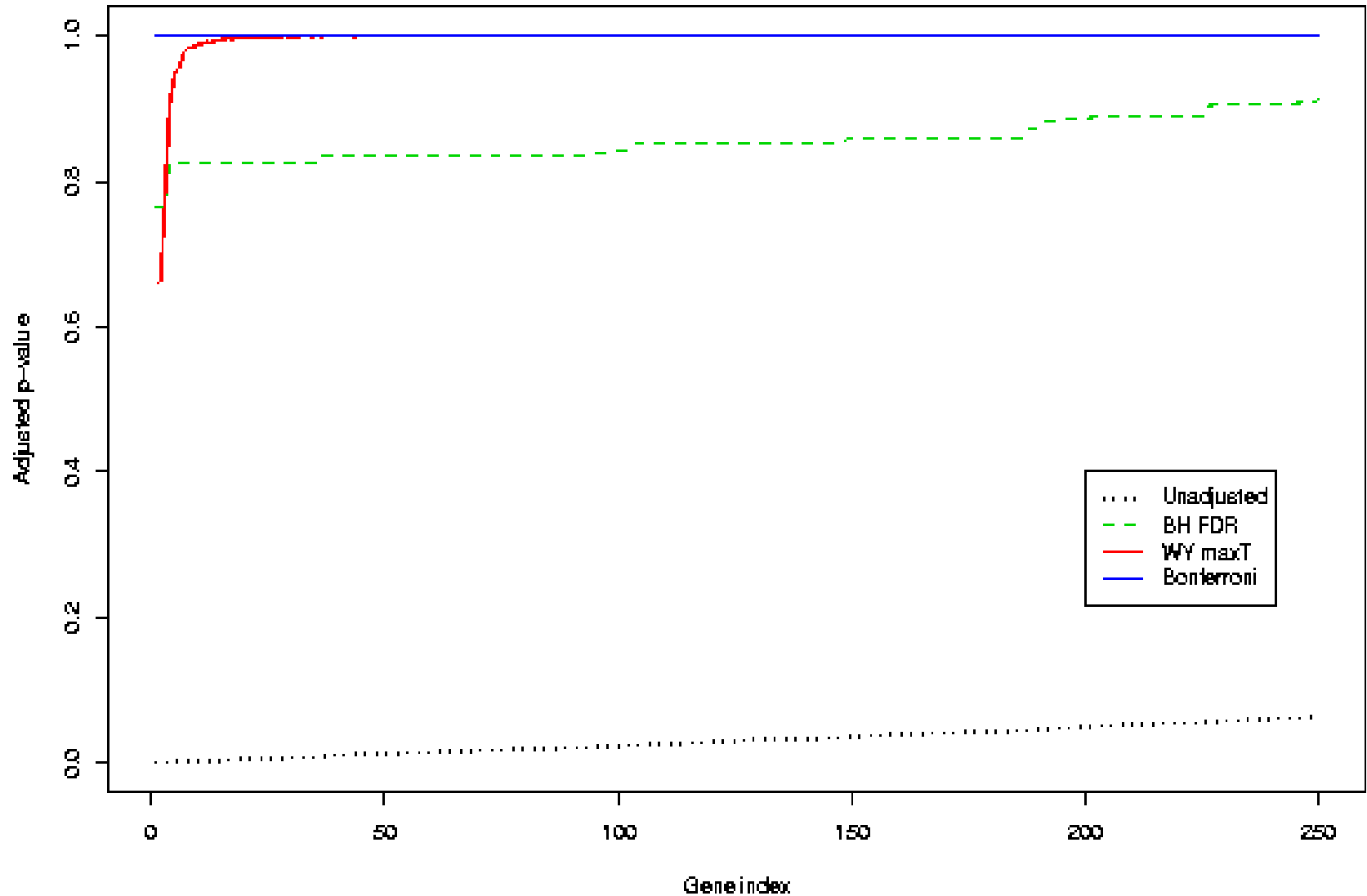
Bittner et al. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406: 536-540

Alizadeh et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503-511

# Results: Lymphoma



# Results: Melanoma



## Other Proposals from the Microarray Literature

- ‘**Neighborhood Analysis**’, Golub et al.
  - In general, gives only weak control of FWER
- ‘**Significance Analysis of Microarrays (SAM)**’ (2 versions)
  - Efron et al. (2000): weak control of PFER
  - Tusher et al. (2001): strong control of PFER
- SAM also estimates ‘FDR’, but this ‘FDR’ is defined as  $E(V|H_0^{\text{COM}})/R$ , not  $E(V/R)$

# Controversies

- **Whether** multiple testing methods (adjustments) should be applied at all
- **Which tests** should be included in the **family** (e.g. all tests performed within a single experiment; define 'experiment')

- It is plausible that **all nulls may be true**
- A **serious claim** will be made whenever any  $p < .05$  is found
- **Much data manipulation** may be performed to find a 'significant' result
- The analysis is planned to be **exploratory** but wish to claim 'sig' results are real
- Alternatives
  - Bayesian approach
  - Meta-analysis



# Some references

- Benjamini and Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB* 57: 289-200
- Benjamini and Yekutieli (2001) The control of false discovery rate in multiple hypothesis testing under dependency. *Annals of Statistics*
- Hochberg (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800-802
- Holm (1979) A simple sequentially rejective multiple testing procedure. *Scand. J Statistics* 6: 65-70
- Ihaka and Gentleman (1996) R: A language for data analysis and graphics. *J Comp Graph Stats* 5: 299-314
- Tusher et al. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *PNAS* 98: 5116 -5121
- Westfall and Young (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley
- Yekutieli and Benjamini (1999) Resampling based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Inf* 82: 171-196

# LOCAL FDR AND COVARIATE MODULATED FDR



**EMBIO** Styringsgruppen for forskning innen  
molekylærbiologi, bioteknologi og bioinformatikk ved UiO



$(sfi)^2$

**Statistics for Innovation**



# A Microarray Example: The Prostate Data

Singh et al., 2002

- *102 Subjects*: 50 normal, 52 cancer
- $N = 6033$  genes :  $X$   $6033 \times 102$
- *Which genes are “non-null”?*  
i.e. expressed differently in cancer vs normal subjects?

# t-statistics and z-scores

- $i^{\text{th}}$  row of  $X$   
normals  $(x_{i1}, x_{i2}, \dots, x_{i50})$   
cancer  $(x_{i51}, x_{i52}, \dots, x_{i102})$  } “ $t_i$ ”
- $t_i$  = two-sample  $t$ -stat, cancer vs normals

z-scores

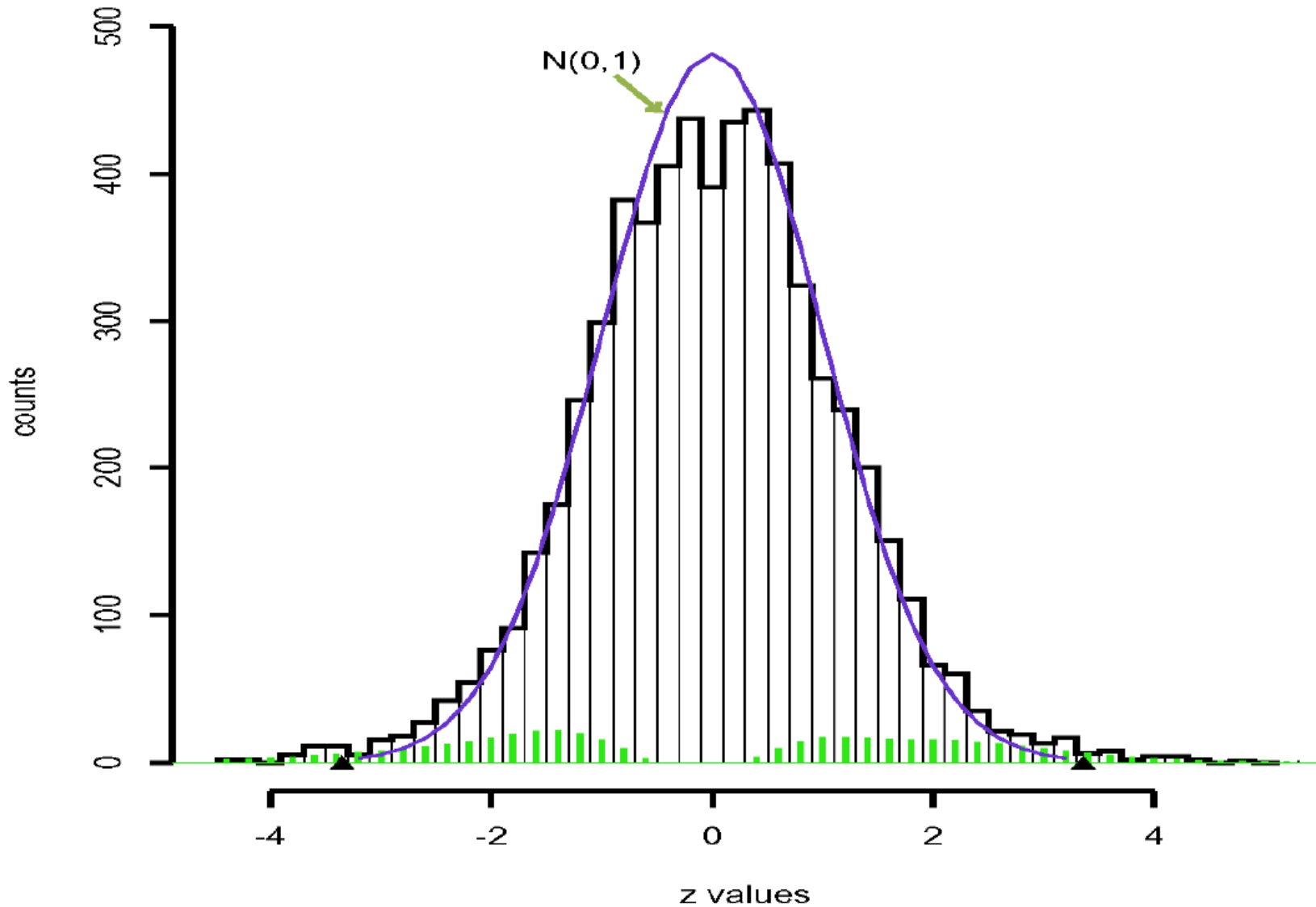
$$z_i = \Phi^{-1}(F_{100}(t_i))$$

where

$\phi$  is  $N(0, 1)$  cdf, and  $F_{100}$  is cdf for  $t_{100}$ .

- **Theoretical Null**      $z_i \sim N(0, 1)$

# Prostate Data, Singh et al: z-values for 6033 genes, Comparing 50 normals with 52 cancer subjects



*Short vertical bars are estimated counts of non-null genes*

## Two Classes of Genes “null”, “non-null”

- $\begin{cases} p_0 = \text{Prob}\{\text{null}\}, & f_0(z) \text{ density if null} \\ p_1 = \text{Prob}\{\text{non-null}\}, & f_1(z) \text{ density if non-null} \end{cases}$
- $p_0$  large ( $\geq 0.90$ )
- **Theoretical Null**  $f_0 \sim N(0, 1)$

(fits center of histogram)

- $f_1(z)$  ”long tailed” ( $z_i$ 's far from zero)
- *Mixture density*

$$f(z) = p_0 f_0(z) + p_1 f_1(z)$$

## EMPIRICAL Bayesian testing

Assume a prior probability on each null hypothesis

$$\text{Prob}(H_0 \text{ is true}) = \pi_0$$

$$\text{Prob}(H_0 \text{ is true} \mid \text{data}) = \frac{\overset{\text{GIVEN}}{P(\text{data} \mid H_0)} \pi_0}{P(\text{data})}$$

$$P(\text{data}) = \pi_0 P(\text{data} \mid H_0) + (1 - \pi_0) P(\text{data} \mid H_1)$$

$$\text{Prob}(H_0 \text{ is true}) = \pi_0$$

$$\text{Prob}(H_0 \text{ is true} \mid \text{data}) = \frac{\pi_0 P(\text{data} \mid H_0)}{P(\text{data})}$$

Test statistics  $\mathbf{z} = (z_1, \dots, z_i, \dots, z_m)$

Hypothesis:  $H_{0i}$  vs.  $H_{1i}$ ,  $i = 1, \dots, m$ .

**Local false discovery** **PROBABILITY**

$$fdr(z_i) = \text{Prob}(H_0 \text{ true} \mid Z = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}$$

If we decide to reject null hypothesis  $H_{0i}$  based on the test statistics  $z_i$ , then  $fdr(z_i)$  is probability that we make the wrong decision.



## Empirical Bayes Methods and False Discovery Rates for Microarrays

Bradley Efron<sup>1,\*</sup> and Robert Tibshirani<sup>2</sup>

# Empirical Bayes Analysis of a Microarray Experiment

Bradley EFRON, Robert TIBSHIRANI, John D. STOREY, and Virginia TUSHER

---

Microarrays are a novel technology that facilitates the simultaneous measurement of thousands of gene expression levels. A typical microarray experiment can produce millions of data points, raising serious problems of data reduction, and simultaneous inference. We consider one such experiment in which oligonucleotide arrays were employed to assess the genetic effects of ionizing radiation on seven thousand human genes. A simple nonparametric empirical Bayes model is introduced, which is used to guide the efficient reduction of the data to a single summary statistic per gene, and also to make simultaneous inferences concerning which genes were affected by the radiation. Although our focus is on one specific experiment, the proposed methods can be applied quite generally. The empirical Bayes inferences are closely related to the frequentist false discovery rate (FDR) criterion.

---

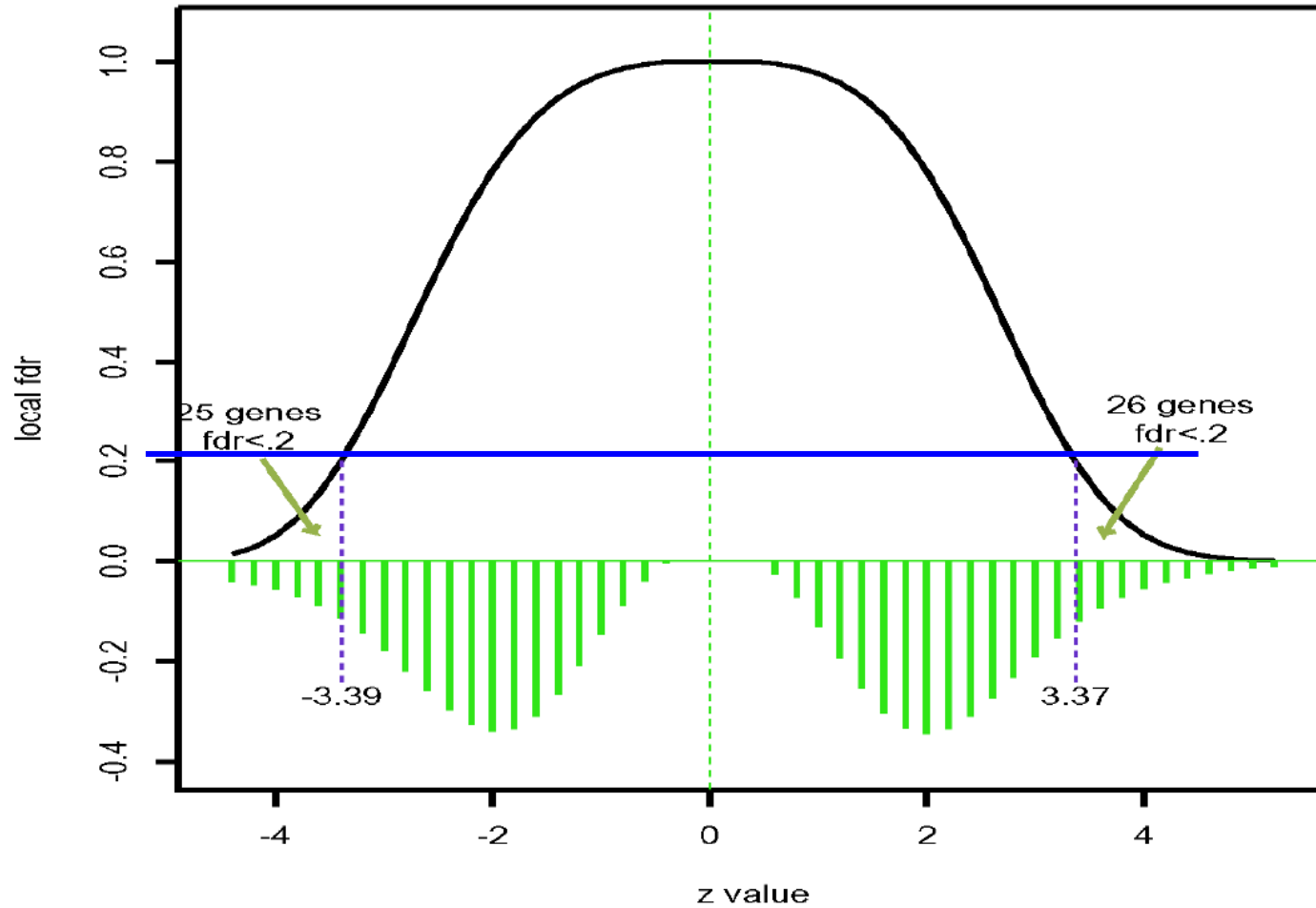
# Empirical Bayes

- Estimate mixture density  $f(z)$  from observed  $z$ -values  $z_1, z_2, \dots, z_N$  :

$$\widehat{fdr}(z) = p_0 f_0(z) / \widehat{f}(z)$$

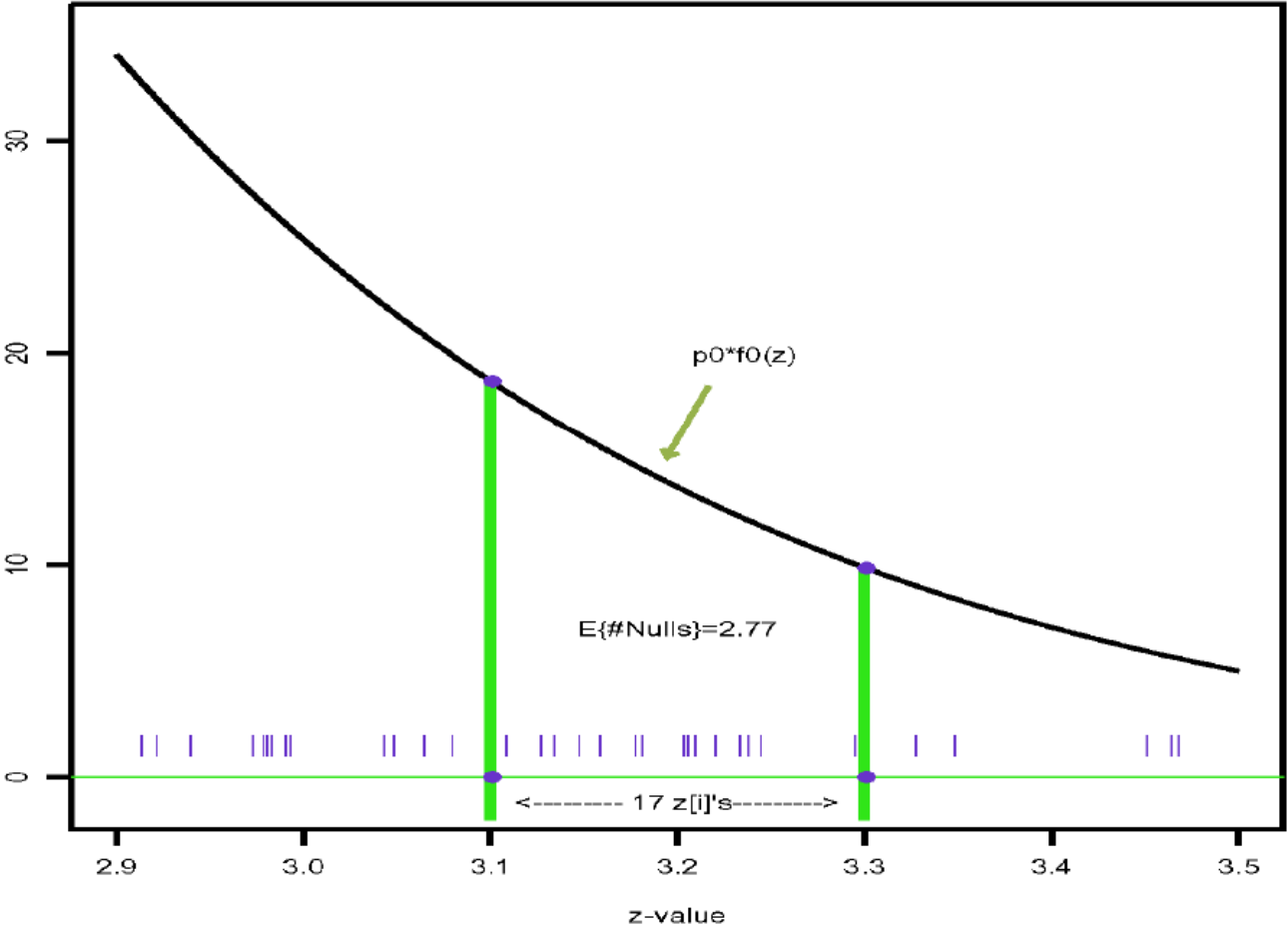
- $BH : p_0 = 1$  ( $\widehat{p}_0 = 0.94.$ ) Also estimated from data.
- *Don't need:*  $z_i$ 's independent,  $t$ -tests ...

Estimated fdr(z) for prostate data (solid curve)  
Bars proportional to non-null histogram



*Short vertical bars are estimated counts of non-null genes*

close-up of histogram:  $fdr(3.2) = 2.77/17$



## How to compute the local FDR?

- Histogram has 49 bins, width  $\Delta = .2$
- $\text{bin}_{39} = [3.1, 3.3]$ ; count  $y_{39} = 17$
- $y_{39}^{(o)} = \# \{ \text{null genes in } \text{bin}_{39} \} = ??$
- $\hat{y}_{39}^{(o)} = E \# \{ \text{null genes in } \text{bin}_{39} \} = N \Delta \hat{p}_0 f_0(3.2)$   
 $= 2.77$

- $\widehat{fdr}(3.2) = 2.77/17 = .16$

- *About one sixth of the 17 genes in  $\text{bin}_{39}$  are false discoveries*

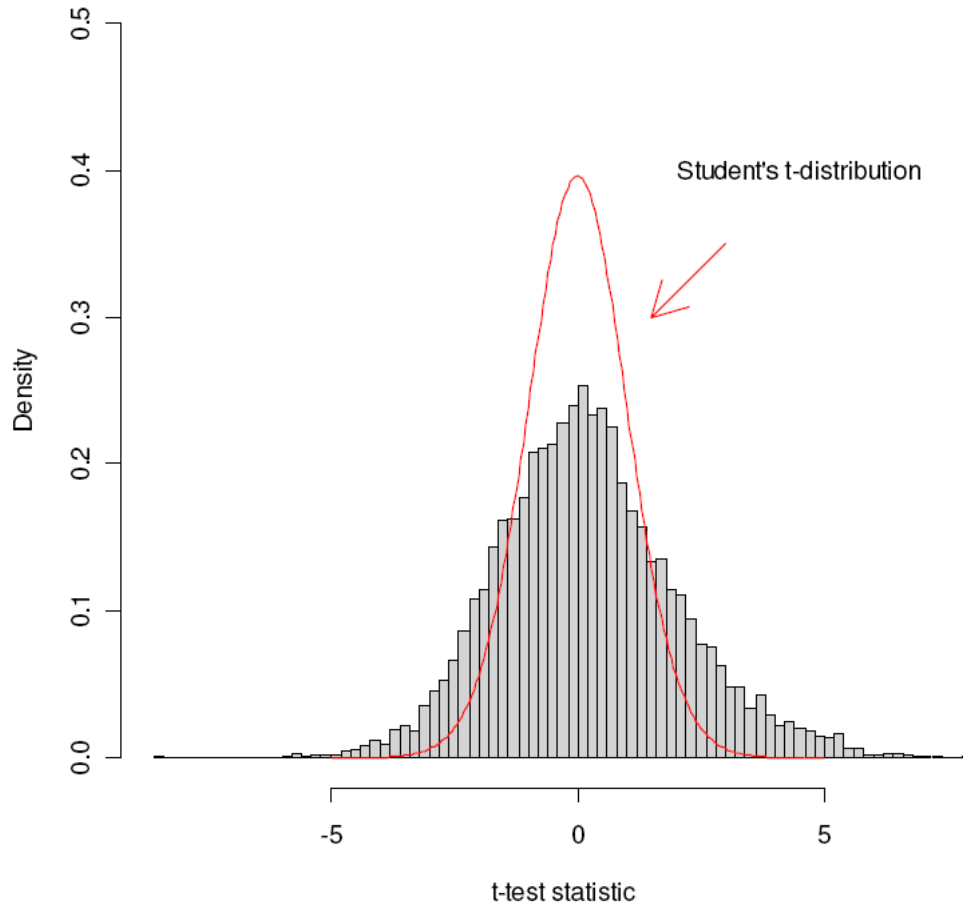
## How to compute the green bars? (estimated counts of non-null genes)

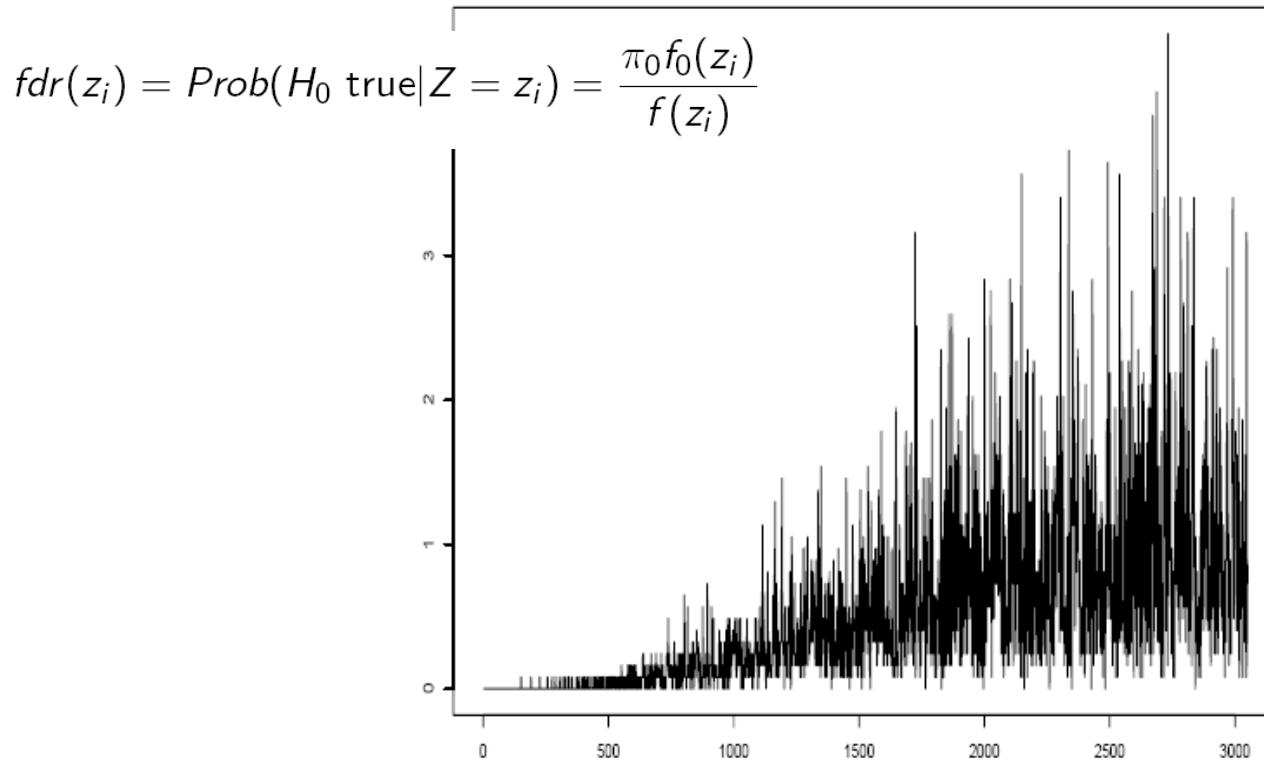
- $y_k$  = observed count in  $k^{th}$  bin
- $1 - fdr(z_i) = Prob\{gene\ i\ non\text{-}null | z_i = z\}$
- So estimated number non-nulls in  $bin_k$  is

$$\hat{y}_k^{(1)} = [1 - \widehat{fdr}_k] y_k$$

where  $\widehat{fdr}_k = \widehat{fdr}(z = \text{midpoint } bin_k)$ .

# Golub *et al* (1999)



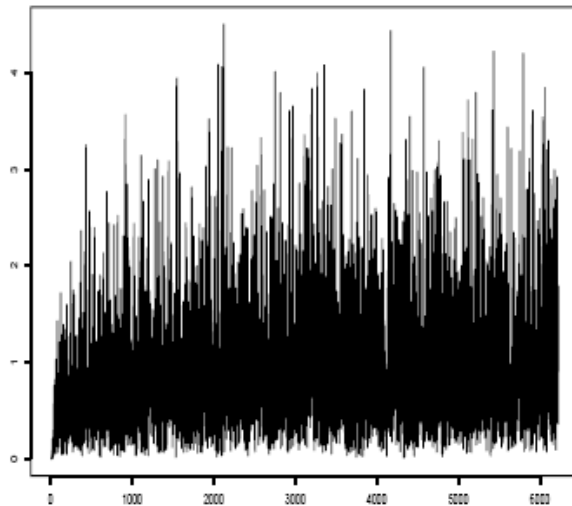


Genes, order with increasing un-adjusted p-value

The lists based on un-adjusted p-values and **local FDR can be very different**, while a list based on FDR-adjusted p-values would not be re-ordered.



APO data!



(a)

## BMC Bioinformatics

BioMed Central

Research article

Open Access

**Determination of the differentially expressed genes in microarray experiments using local FDR**

J Aubert, A Bar-Hen, J-J Daudin\* and S Robin

# Covariate-modulated false discovery rates

The covariate-modulated false discovery rate

- takes advantage of prior information on the probability of each null hypothesis being true
- based on external additional data,
- to produce a more precise list of selected genes.

*Submitted to the Annals of Applied Statistics*

## UNSUPERVISED EMPIRICAL BAYESIAN MULTIPLE TESTING WITH EXTERNAL COVARIATES

BY EGIL FERKINGSTAD,\* ARNOLDO FRIGESSI, HÅVARD RUE,  
GUDMAR THORLEIFSSON AND AUGUSTINE KONG

*University of Oslo and Centre for Integrative Genetics,  
(sfi)<sup>2</sup> — Centre for Statistics for Innovation,  
Norwegian University of Science and Technology,  
Decode Genetics and Decode Genetics*

## Additional prior information:

Assume that for each null hypothesis  $H_{0i}$  there is a corresponding covariate  $X_i$  which influences the probability of  $H_{0i}$  being true.

Null hypotheses with different corresponding values of  $x_i$  will have different probabilities of being true.

$$Prob(H_{0i} \text{ true} | X_i = x_i) = \pi_0(x_i)$$

$$Prob(H_{0i} \text{ true} | X_i = x_i) = \pi_0(x_i)$$

Distribution of the test statistics  $z_i$  :

$$g(z_i | x_i) = \underbrace{\pi_0(x_i)}_{\text{modulates the null distribution}} g(z_i | H_{0i}) + (1 - \pi_0(x_i)) g(z_i | H_{1i}, x_i)$$

... modulates the null distribution

### Covariate modulated false discovery rate

$$P(H_{0i} | z_i, x_i) = \frac{g(z_i | H_{0i}) \cdot \pi_0(x_i)}{g(z_i | x_i)},$$

where the unknowns are :

$$\pi_0(x_i)$$

$$g(z_i | H_{1i}, x_i)$$

## p-value model

We can write the model in terms of p-values instead of Z-test scores:

$$g(z_i|x_i) = \pi_0(x_i)g(z_i|H_{0i}) + (1 - \pi_0(x_i))g(z_i|H_{1i}, x_i)$$

corresponds to

$$f(p_i | x_i) = \pi_0(x_i) + (1 - \pi_0(x_i))f(p_i | H_{1i}, x_i)$$

where the unknowns are:

$$\pi_0(x_i) \text{ and } f(p_i | H_{1i}, x_i)$$

( $f(p_i|H_{0i}) = 1$  since  $p_i$  is a p-value).

We have data  $(p_i, x_i)$  and need to estimate  $\pi_0(x_i)$  and  $f(p_i|H_{1i}, x_i)$ .

## Covariate-modulated FDR

$$cmfdr(p_i|x_i) = P(H_{0i}|p_i, x_i) = \frac{P(H_{0i}|x_i)}{f(p_i|x_i)} = \frac{\pi_0(x_i)}{f(p_i|x_i)}$$

$$f(p_i | x_i) = \pi_0(x_i) + (1 - \pi_0(x_i)) f(p_i | H_{1i}, x_i)$$

How to model  $\pi_0(x_i)$  and  $f(p_i | H_{1i}, x_i)$ ?

Diaconis and Ylvisaker (1985) show that any distribution on  $[0, 1]$  can be modelled as mixture of beta distributions.

Allison et al (2002) investigate this further and apply it to estimating the density  $f$  of a sample of p-values.

In their experience with several sets of data, the simplest possible model, which is a mixture of a  $U[0, 1]$  corresponding to the true null hypotheses and one single beta component corresponding to the false null hypotheses, seems always to be sufficient.

Hence we use

$$f(p_i | x_i) = \pi_0(x_i) \text{UNIFORM}[0,1] + (1 - \pi_0(x_i)) \text{BETA}(\xi(x_i), \theta(x_i))$$

(formal checking of the fit of the alternative model can be done)

$$f(p_i | x_i) = \pi_0(x_i) \text{UNIFORM}[0,1] + (1 - \pi_0(x_i)) \text{BETA}(\xi(x_i), \theta(x_i))$$

Next we bin the p-values into B sets:  $B_1, B_2, \dots, B_B$  for increasing  $x_i$ .

In each bin we assume the dependence on  $x_i$  to be constant.

We drop the dependence on  $x_i$  within each bin but allow for bin specific parameters  $\theta_j$  and  $\xi_j$ .

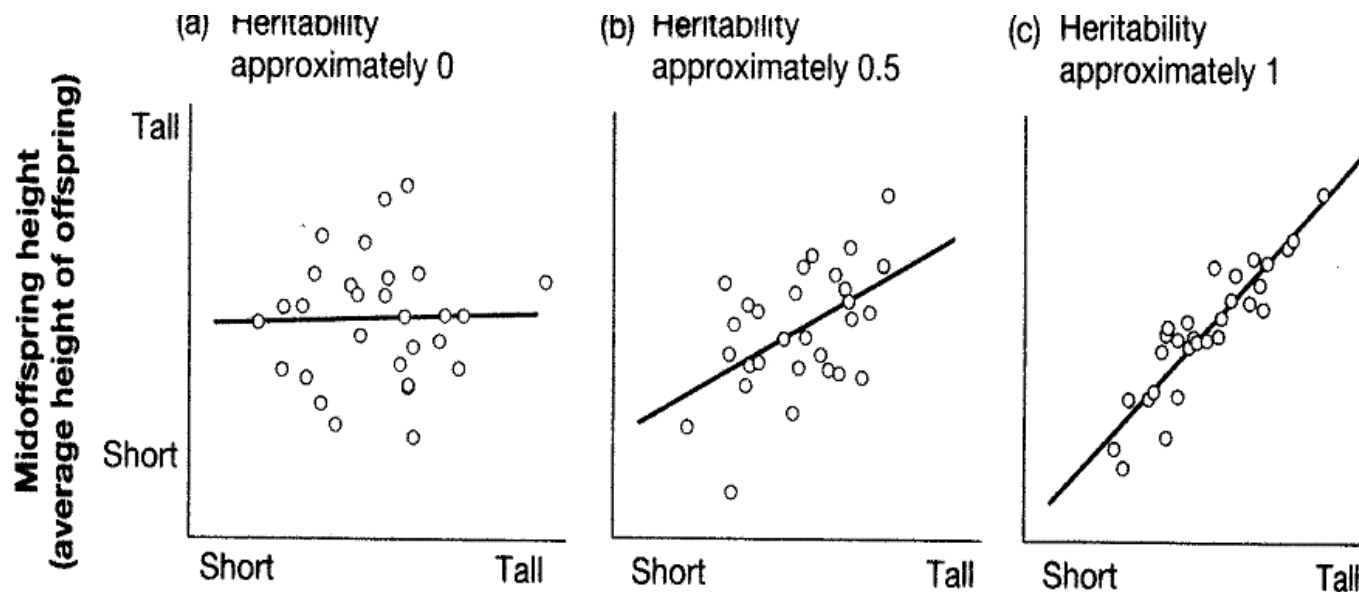
In bin  $B_j$  :

$$f_j(p_i) = \pi_{0j} + (1 - \pi_{0j}) \text{BETA}(p_i | \xi_j, \theta_j)$$



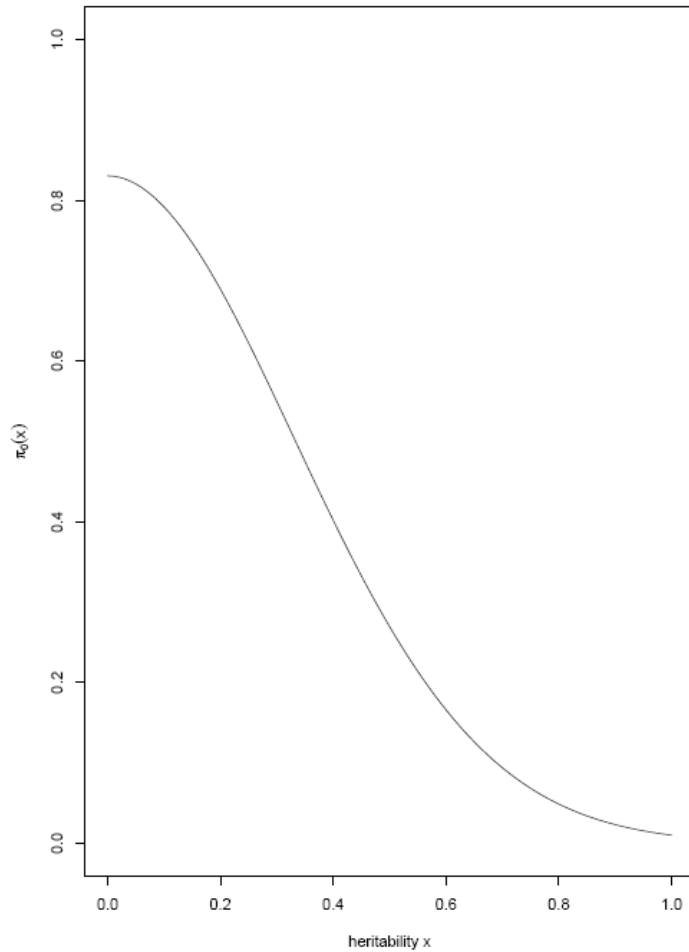
Use family data, to study the tendency for the expression of each gene (seen as the phenotype) to be attributable to genetic factors.

Compute for each gene its **heritability**, the proportion of variation in a trait that is attributable to genetic factors, here defined as the ratio between the genetic variance and the total phenotypic variance in a classical simple additive model.



Genes with higher heritability  $x$  have a larger probability to be linked with the phenotype, so smaller  $P(H_0|x) = \pi_0(x)$ .

$\pi_0(x)$



heritability x

A reasonable parametric form

$$\pi_0(x) = e^{-\alpha - (\beta - \alpha)x^\gamma}$$

could be used instead than binning.

## Data

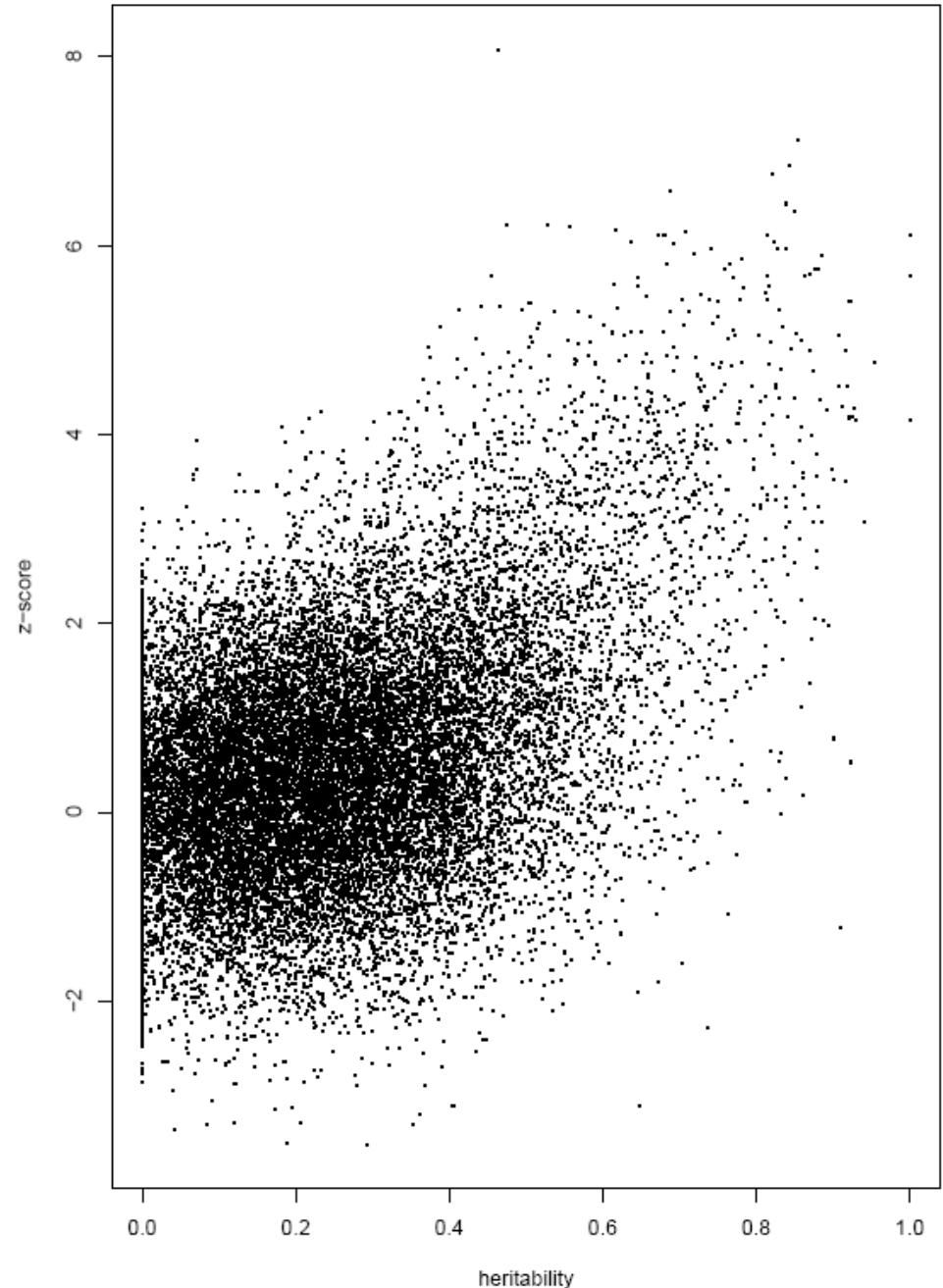
22317 pairs of z-scores and heritabilities

(test for linking certain QTL markers to expressions – here only cis regulation)

If we use the parametric model of  $\pi_0(x)$ , we obtain estimates

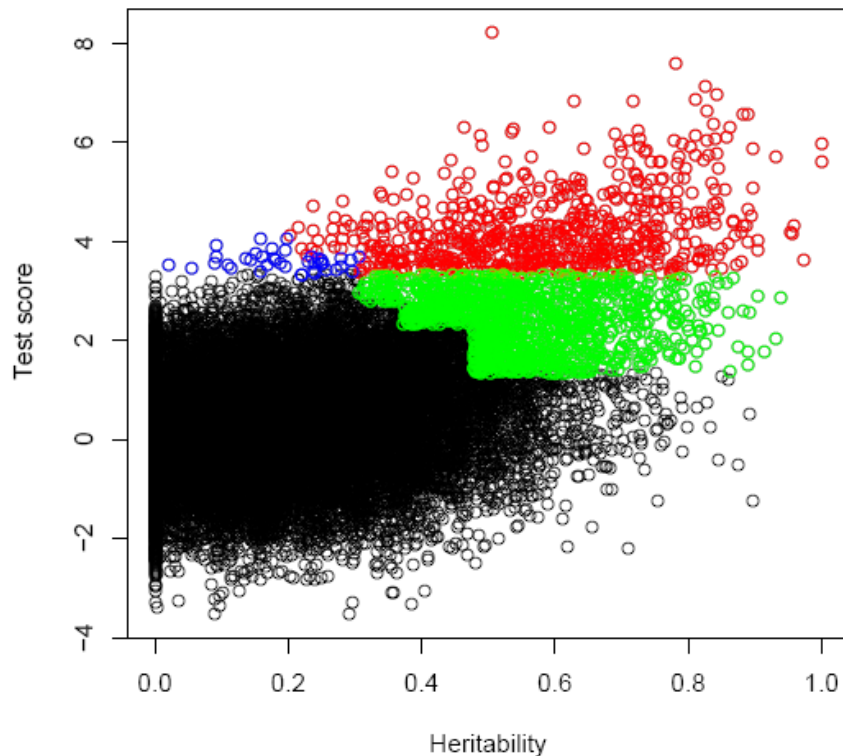
$$\pi_0(0) = 0.83$$

$$\pi_0(1) = 0.01.$$



# Comparison to Efron's local fdr method

fdr cutoff level 0.05



Each point is a gene. In total:  
22317 genes.

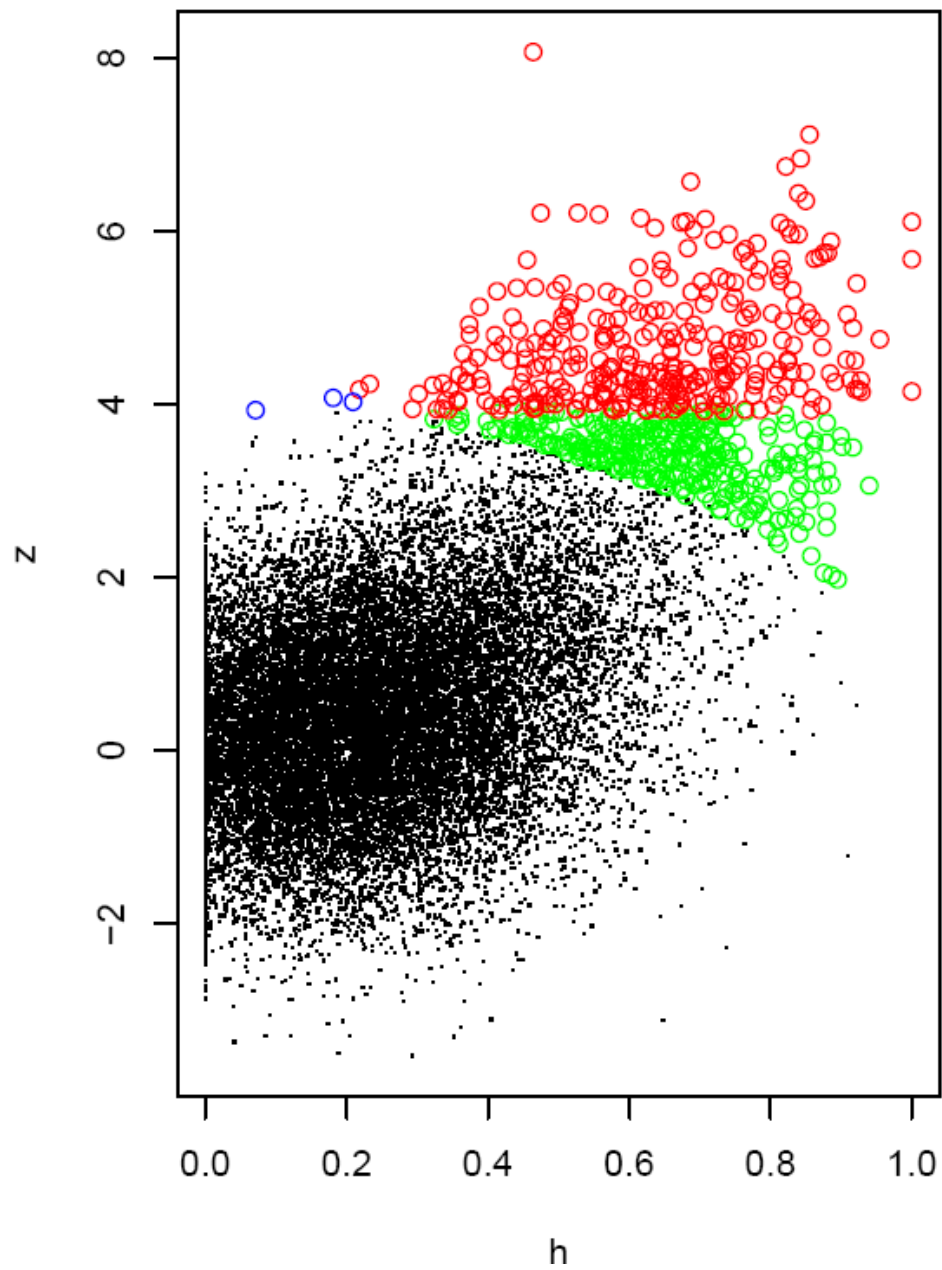
**Green:** cmfdr significant, but lfdr not significant, 1158 genes

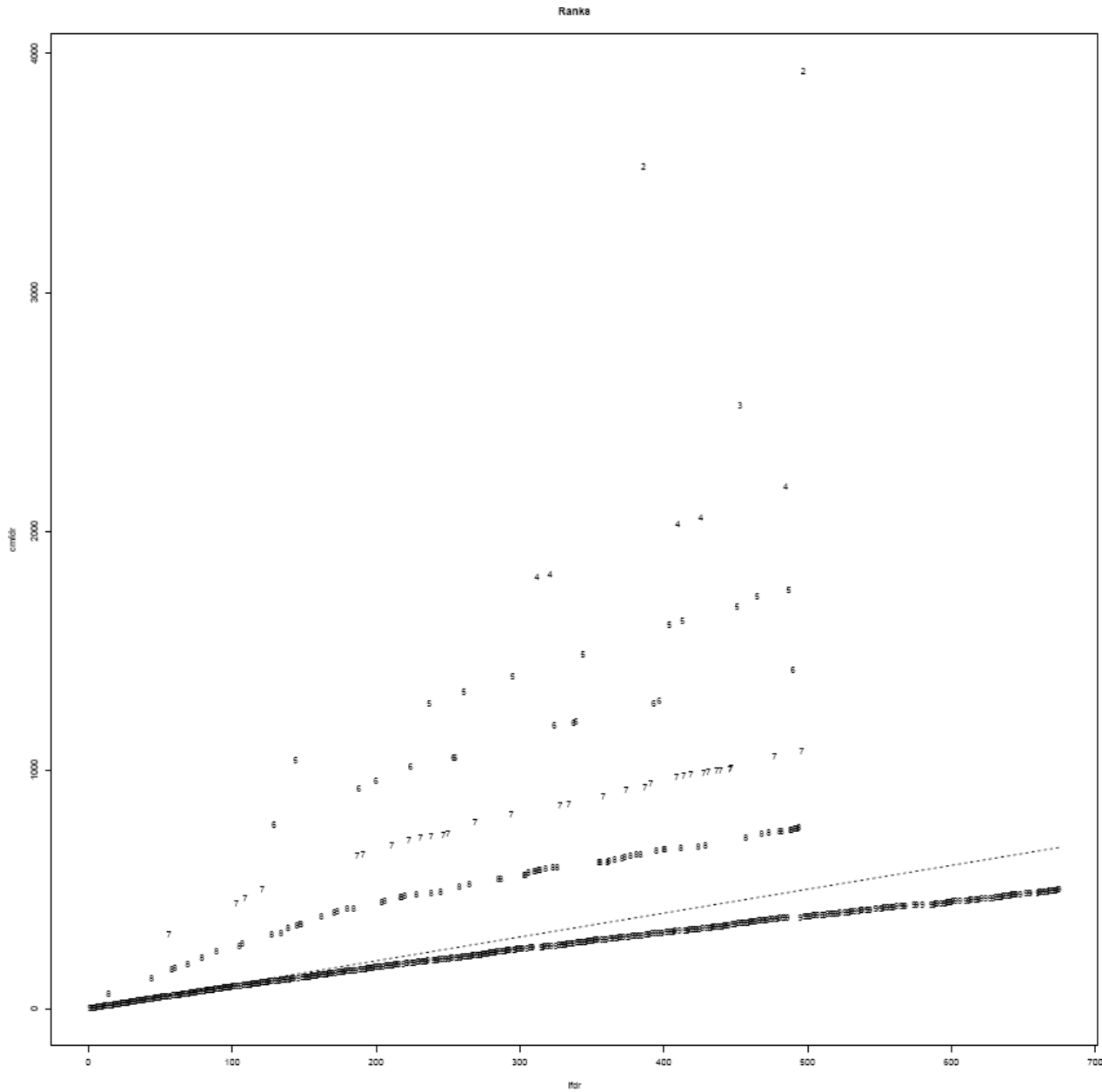
**Blue:** lfdr significant, but cmfdr not significant, 39 genes

**Red:** in agreement, 665 genes

NB: The ranks of the **red** genes may be different between lfdr and cmfdr.

fdr level: 0.01





Different ranks between the Efron based list and the cmfdr based list. for the common (red) Genes.

# CGH and Expressions

Jonathan R. Pollack, Therese Sorlie, Charles M. Perou, Christian A. Rees, Stefanie S. Jerey, Per E. Lonning, Robert Tibshirani, David Botstein, Anne-Lise Borresen-Dale, and Patrick O. Brown.

*Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.*

PNAS, 99(20):12963-12968, 2002.

In this study, measurements of DNA copy number and mRNA expressions (relative to a single reference sample) on 6095 genes were done for 4 cell lines and 37 breast tumors.

“A strong influence of DNA copy number on gene expression” (p. 12965) was observed

Testing for differentially expressed genes between the two types.  
Simple t-test on normalised expression data.

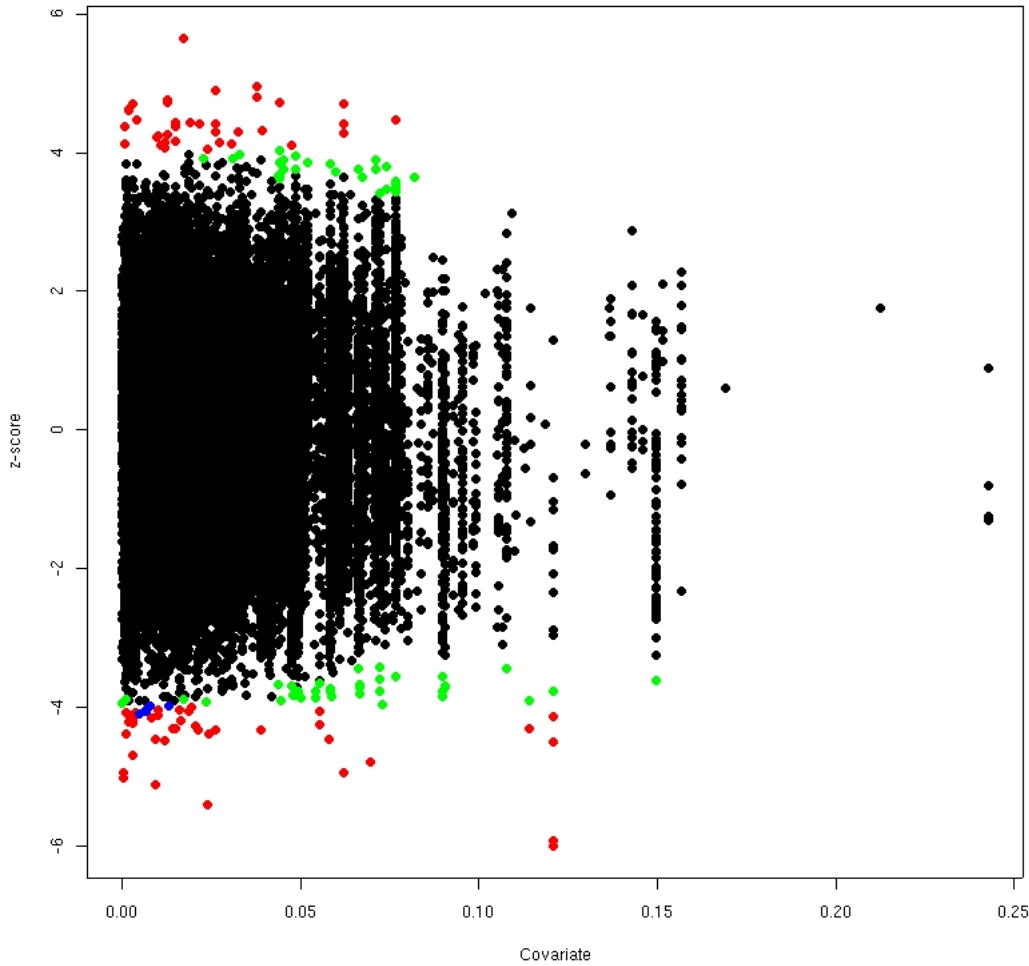
External covariate: DNA copy number available for each gene, for each sample.

- When the mean copy number for a gene is very different in LUMINAL and BASAL, then we can expect a larger probability that the gene expression will differ between LUMINAL and BASAL.
- But expressions can differ also when the copy numbers do not: however, for this to be a real effect, the difference in expression must be “really big”, with respect to normal variability.

So, for a gene  $i$ , the natural covariate  $x_i$  is simply the difference of the average copy number in the LUMINAL samples minus the average copy number in the BASAL samples.



# t statistics



cmfdr = 5%






76 genes sign. with Efron and cmfdr

65 genes cmfdr sign. but not Efron.

5 genes Efron sign. but not cmfdr.  
(cluster around  $x=0$ ,  $z=-4$ ).

No difference

Large difference

-  Allison *et al.*, *A mixture model approach for the analysis of microarray gene expression data*, *Comp. Stat. Data Anal.*, **39**, 1–20 (2002).
-  Benjamini and Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, *J. R. Statist. Soc. B*, **57**, 289–300 (1995).
-  Diaconis and Ylvisaker, *Quantifying prior opinion*, *Bayesian Statistics* **2**, 133–156 (1985).
-  Efron, Tibshirani, Storey, and Tusher, *Empirical Bayes analysis of a microarray experiment*, *J. Am. Stat. Assoc.*, **96**, 1151–1160 (2001).
-  Langaas, Ferkingstad and Lindqvist, *Estimating the proportion of true null hypotheses, with application to DNA microarray data*, *J. R. Statist. Soc. B*, **67**, 555–572 (2005).

Statistical methods and bioinformatics for the analysis of microarray data

# Themes from the Literature

Arnoldo Frigessi  
Department of Biostatistics  
University of Oslo  
frigessi@medisin.uio.no



radium.no

Cancer Research at the Norwegian Radium Hospital  
Comprehensive Cancer Center

**(sfi)<sup>2</sup> Statistics for  
Innovation**

**BMMS**

Thematic Research Area at the Faculty of Medicine  
University of Oslo, Norway

**Finding errors in the papers of others...**

Dave et al.

*"Prediction of survival in follicular lymphoma based on molecular features of tumor infiltrating cells".*

NEJM Nov. 18, 2004 vol 351:2159-2169,

I think it is useful to determine the degree to which an analysis is **fragile**. Even if an analysis produces small p-values, a scientist should be concerned if small but reasonable changes in the analysis strategy cause large changes in the results.

With microarray analyses, there are many choices that one has to make, and one hopes that the results are not too sensitive to these choices.

Dave et al. derive a model for predicting patient survival from gene expression data using two "immune response" clusters, IR1 (good prognosis) and IR2 (poor prognosis). A Cox model using expression averages from the IR1 and IR2 clusters was constructed, and this model had a highly significant p-value (0.003 or less) in an independent test set.

RE-ANALYSIS (Tibshirani et al.)

When their equal-sized training and test sets are swapped, and their model-building procedure is re-applied, their finding disappears and virtually nothing is significant in the test set.

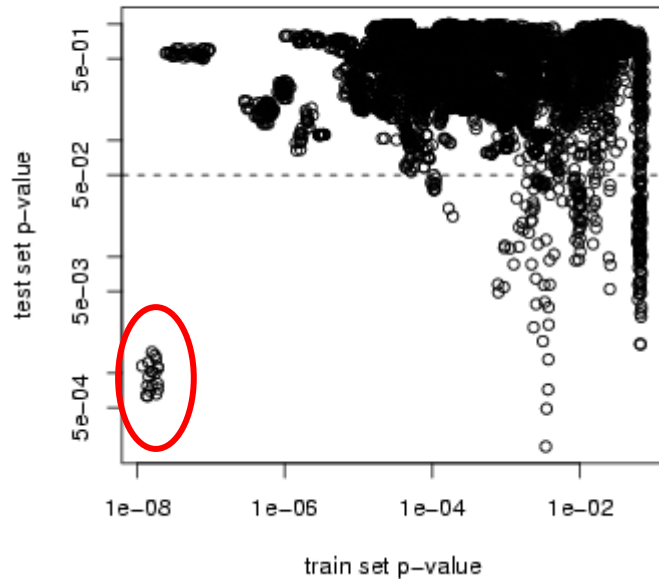
Also, when a small change is made to their model-building recipe (changing the allowable cluster size range from [25,50] to [30,60]) with either the original or swapped datasets, again, their finding disappears and very little of significance emerges.

This and other analyses suggest that their result occurred by chance and is not robust or reproducible.

Other analyses suggest that there is little or no correlation between gene expression and patient survival in this dataset.

The steps in the authors' modelling procedure were:

0. Divide the data randomly into training and test sets of approximately equal numbers of patients. Apply the following recipe [steps 1--5] to the training set.
1. Choose all genes with univariate Cox score  $> 1.5$  in absolute value. This reduced the number of genes from roughly 49,000 to roughly 3,200, with about a 50-50 split between good prognosis genes (negative scores) and poor prognosis genes (positive scores).
2. Do separate hierarchical clusterings of the good and poor prognosis genes.
3. Find all clusters in the dendrograms (clustering trees) containing between 25 and 50 genes, with internal correlation at least 0.5. Represent each cluster by the average expression of all genes in the cluster-- a "supergene". (ca. 200)
4. Try every pair of supergenes as predictors in Cox models for predicting survival.
5. Choose the most significant pair from this process.  
The authors call the resulting pair of clusters IR1 (good prognosis) and IR2 (poor prognosis).
6. Finally use the model (IR1, IR2) in a Cox model to predict survival in the test set.

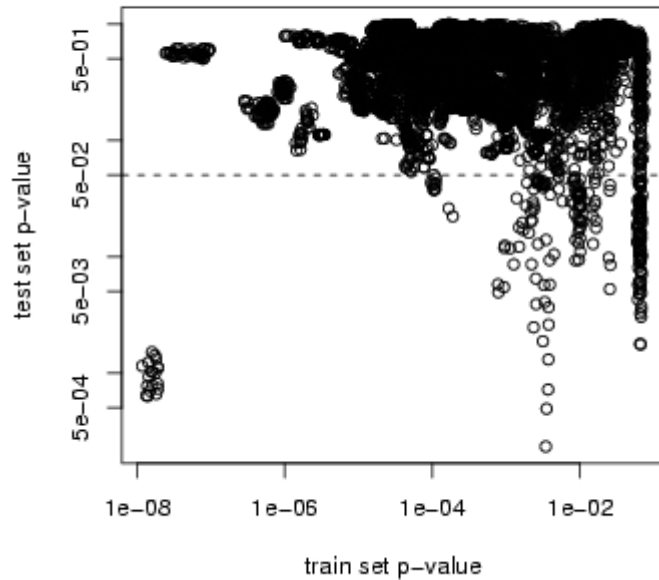


*Training and test p-values for all cluster pairs (8930).*

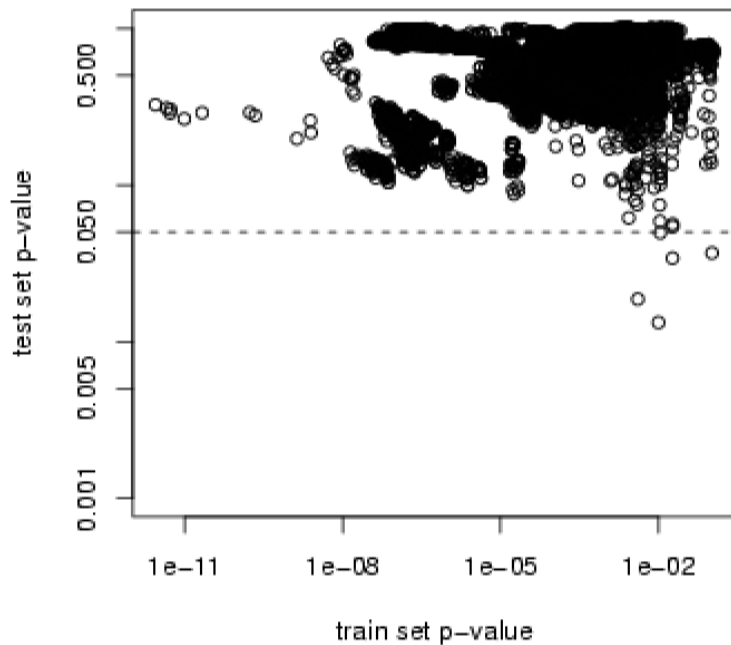
Each point represents the training and test p-values from a Cox model containing two clusters from the set of all clusters that passed the filtering in the first part of the analysis. All possible two-cluster models are represented in the plot.

The **(IR1, IR2) cluster model** corresponds to the island of points in the bottom left. All of these pairs use the IR2 cluster, and variations on the IR1 cluster. The (IR1, IR2) model has the smallest training set p-value-





But we also note that the total number of points (cluster pairs) with p-values less than 0.05 (239) is far fewer than we'd expect to see by chance (735), even if there was no correlation between gene expression and survival. This suggests that there may be no overall significance in this dataset.



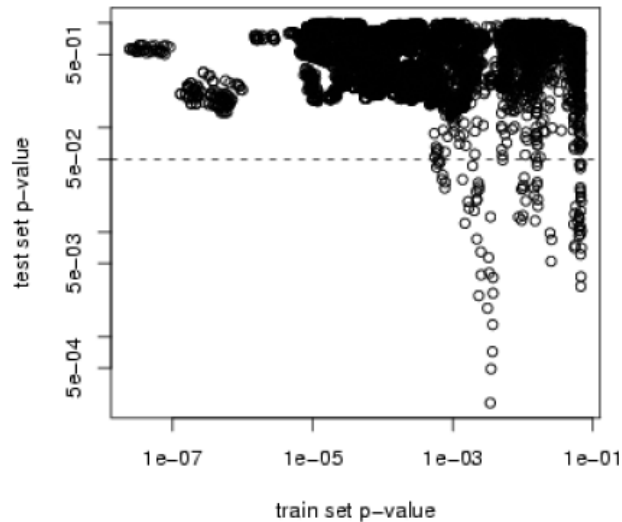
*Training and test p-values for all cluster pairs, with training and test sets swapped.*

The original training and test sets were of approximately equal size, and were chosen at random. Hence it seems reasonable to swap them, train on the original test set and test on the original training set.

We see that the authors' finding does not appear, even approximately.

Notice also that only 5 pairs out of 8930 have test set p-values less than 0.05.

Even if there was no correlation between gene expression and survival, we'd expect  $8930 \times .05 = 447$  significant pairs.



*Training and test p-values for all cluster pairs (with original training and test datasets), using a cluster size range of [30,60] genes instead of [25,50].*

In choosing this range, I have intentionally ruled out the strong IR2 cluster, which has 27 genes. But one would expect that if the authors' finding was robust, we would find some other cluster with significant overlap with the IR2 cluster. But the authors' finding does not appear, even approximately.

And there are only 85 pairs out of 11628 that are significant in the test set at the 0.05 level, while we would expect  $11628 \cdot 0.05 = 581$  pairs just by chance.

Authors' reply:

They randomly selected new equal-sized training and test sets from the data, and reapplied their original model (IR1-IR2) to each new half-set.

They found that every resulting p-value was less than 0.011, with a median of 0.001.

**This is not surprising and tells us nothing.**

The original model was highly significant ( $p < 10e-8$ ) on the original training set, simply as a result of the fitting process.

And we already know that it is significant on the original test set ( $p=.003$ ).

Therefore it was very significant on the whole data and there it must be significant on any half of the data that we choose.

To learn about the robustness or fragility of their model, one must go through the entire model building process from scratch.

Sjoblom, T., Jones, S., Wood, L., Parsons, D., Lin, J., Barber, T., Mandelker, D., Leary, R., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S., Willis, J., Dawson, D., Willson, J., Gazdar, A., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B., Bachman, K., Papadopoulos, N., Vogelstein, B., Kinzler, K., and Velculescu, V.

**The consensus coding sequence of human breast and colorectal cancers.**

Science, 2006, pages 268–274.



## Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers"

Gad Getz, *et al.*

*Science* **317**, 1500b (2007);

DOI: 10.1126/science.1138764

**S**jöblom *et al.* (1) reported the first genome-wide effort to identify genes mutated in cancer. They also introduced a two-stage design in which they screened a large set of genes (13,023) for somatic mutations in a discovery set (11 breast and 11 colorectal cancers) and then screened only the small subset of genes that harbored at least one somatic mutation in a validation set (24 breast or colorectal tumors). They identified genes as candidate cancer genes (*CAN* genes) by applying a statistical model designed to assess the likelihood that the observed somatic nonsynonymous mutations would occur by chance. The approach employed the false discovery rate (FDR) approach of Benjamini and Hochberg (2) and used an assumed background mutation rate of  $\mu = 1.2 \times 10^{-6}$ .

The Sjöblom *et al.* analysis yielded rank-ordered lists of candidate genes with 122 and 69 genes in breast and colorectal cancers, respectively. These genes were said to have a 90% chance of being true cancer genes, that is, harboring mutations at a frequency significantly greater than expected by chance, based on the FDR approach (that is,  $FDR \leq 10\%$ ). Reassuringly, 6 genes known to be mutated in these cancer types appear at the top of these lists (1 in breast and 5 in colon cancer). Extrapolating from these lists to the entire genome, the authors estimate that the total number of genes harboring important somatic mutations in breast and colon cancer, respectively, exceeds 189 and 107 genes, with the typical tumor carrying 14 and 20 mutations. These observations are of great interest because the number of genes is much higher than previously thought. However, this analysis raises two

Identify genes in tumors that have an increased mutation rate.

[Sjoblom et al., 2006] sequenced 13,023 CCDS genes in breast and colorectal cancer tumors. CCDS genes are protein encoding genes and represent the most highly curated gene set currently available. The data collection phase consisted of two main parts in which genes that were deemed not to have an increased mutation rate were eliminated.

These two parts were:

### Discovery screen:

All genes were sequenced in 11 breast and 11 colorectal cancer tumors.

Initially 816,986 mutations were identified. In order to find true somatic mutations (i.e. present in the tumor but not present in the germline of the patient), a complex set of filtering steps was used and all but 1,307, mutations in 1,149 genes were discarded.

The next step of the data collection was only performed on those genes that contained at least one of these mutations.



## Validation screen:

Genes with mutations in the Discovery screen were sequenced in additional 24 breast and 24 colorectal cancer tumors.

Through a similar system as before, 133,693 initially identified mutations were filtered down to 365 in 236 genes.

Only genes with at least one mutation in the Discovery as well as the Validation screen were used in the subsequent statistical data analysis. These genes were called "validated".

Among the **validated genes**, those that have a significantly increased mutation rate have to be identified.

Sjoblom et al proposed to use the Benjamini-Hochberg procedure to deal with multiple hypothesis testing and control the False Discovery Rate (FDR). In order to do this, they defined the CaMP score. A validated gene is determined to be significant at an FDR level of 0.1 if its CaMP score is  $> 1$ .

CaMP is roughly the probability of exactly having the observed number of mutations under the background mutation rate.

Using this score, 122 genes in breast and 69 genes in colon cancer were identified as significant.

However, they made an error!

When this error is corrected only 2 significant genes in breast and 28 genes in colorectal cancer are discovered.

First, the authors incorrectly apply the FDR formula. The formula requires the tail probabilities [ $\text{Prob}(X \geq T)$ ] as input, but Sjöblom *et al.* instead use the point probabilities [ $\text{Prob}(X = T)$ ]. Consequently, their probabilities are smaller than they should be and therefore falsely appear to be more significant. When  $P$  values rather than point probabilities are used, the number of candidate genes falls from 122 to 6 in breast cancer and from 69 to 28 in colorectal cancer.

Second, the analysis is highly sensitive to the background mutation rate  $\mu$  used in the statistical model (see Supporting Online Material). Different tumors and cell lines may have different background mutation rates, and accurate estimation of  $\mu$  requires large amounts of sequence data generated from the same tumor population. Sjöblom *et al.* estimated  $\mu$  based on a different, smaller data set. However, an estimate based on their own data yields substantially higher mutation rates—by factors of about 1.9 and 1.4 in breast and colorectal cancers, respectively (estimated in two ways; see SOM). If these rates are inserted into the analysis, the number of candidate genes falls to only 1 for breast cancer and 11 for colorectal cancer. Only four of these genes were not previously reported as mutated in cancer.

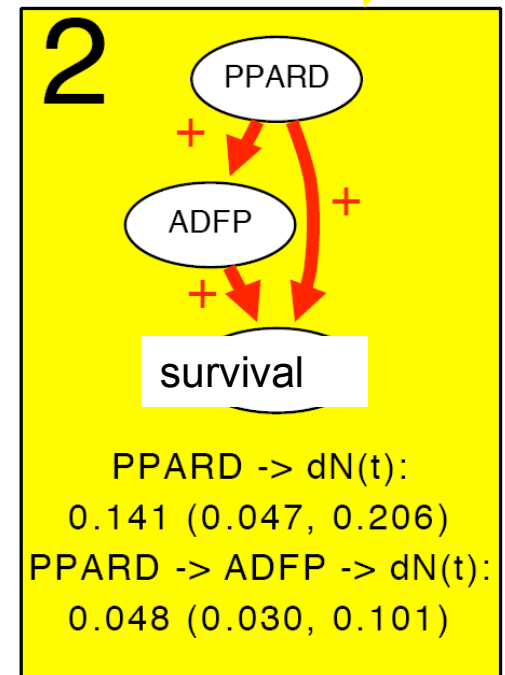
After correcting the statistical analysis and using a background mutation rate that better fits the data, one cannot conclude that the ~200 candidate genes reported in Sjöblom *et al.* have >90% probability of being cancer-related. The issue is simply one of statistical power: Much larger sample sizes are required to detect cancer genes. With smaller sample sizes, most candidate genes are expected to be false positives. Nevertheless, we strongly support the authors' experimental approach and urge its adoption in future large-scale cancer genome sequencing efforts. We suspect that there are indeed many more important cancer genes waiting to be discovered, some of which may well be on the lists of Sjöblom *et al.* In the end, statistical validation of a candidate gene will require study of large samples to show such properties as a high frequency of mutations and a high ratio of nonsynonymous to synonymous mutations.

# Direct and indirect genomic effects on survival by dynamic path analysis

# Direct and indirect genomic effects on survival by dynamic path analysis

How can we detangle direct from indirect effects of genes on survival of cancer patients?

- A gene has an indirect effect on survival if its expression influences survival through one or more other prognostic genes present in the data.
- A gene has a direct effect on survival if its expression influences survival but no gene is found in the data that mediates this effect.



# Aalen's additive regression model

- Intensity process for individual  $i$ :

$$\lambda_i(t) = \alpha_i(t)R_i(t),$$

where  $\alpha_i(t)$  is the hazard rate and  $R_i(t)$  an at risk indicator.

- $x_{i1}(t), \dots, x_{ip}(t)$ : (possibly) time-dependent covariates for individual  $i$  (assumed predictable)
- Aalen's non-parametric additive model is given by

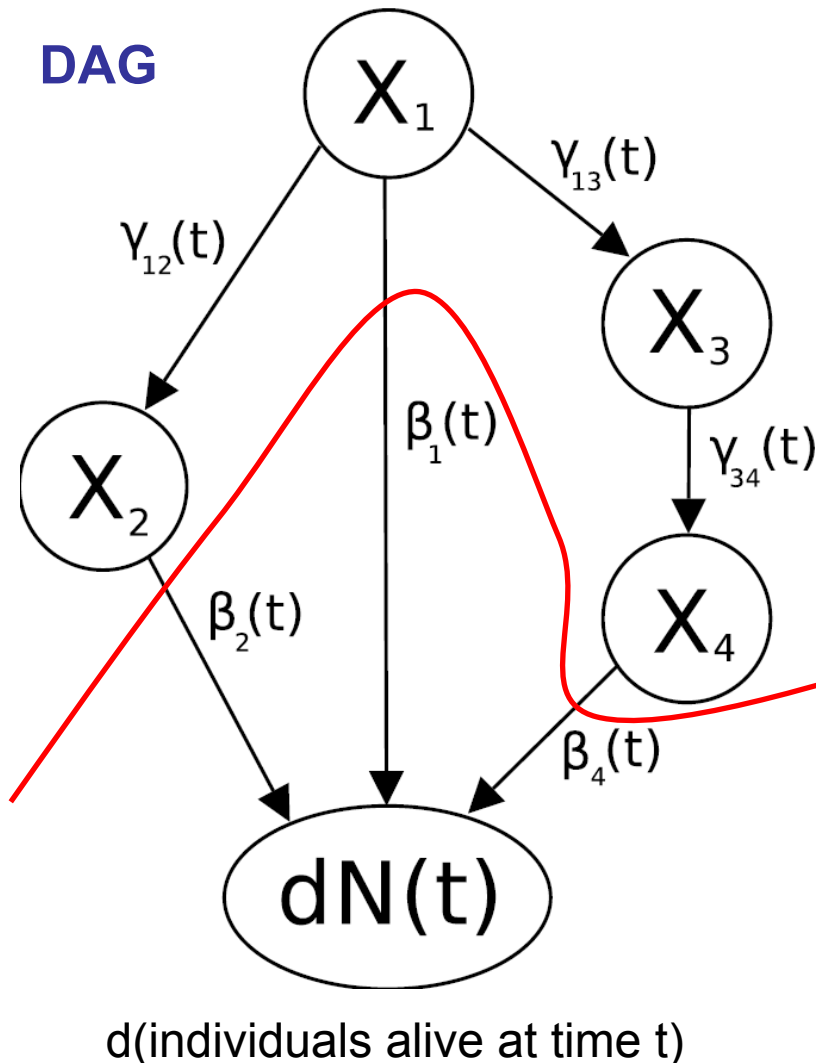
$$\alpha_i(t) = \beta_0(t) + \beta_1(t)x_{i1}(t) + \dots + \beta_p(t)x_{ip}(t).$$

Here,  $\beta_0(t)$  is the baseline hazard, while  $\beta_j(t)$  is the excess risk at  $t$  per unit increase of  $x_{ij}(t)$  for  $j = 1, \dots, p$ .

Estimate the  $\beta_j(s)$  by ordinary least squares at each time  $s$  when a failure occurs.



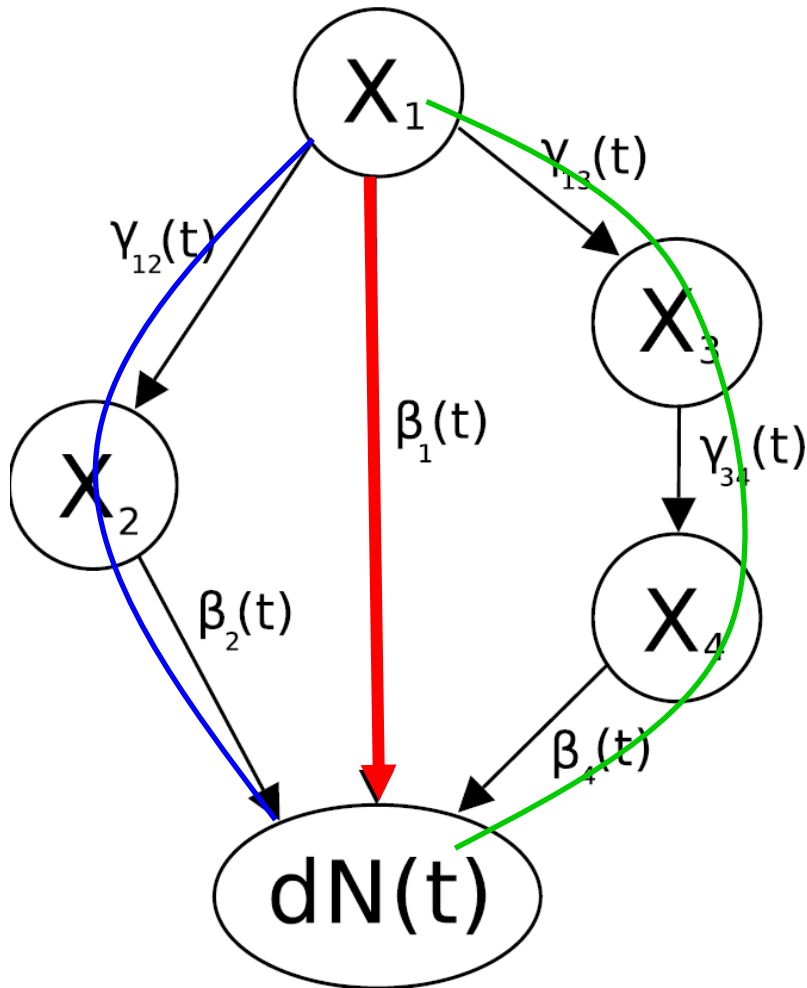
# Dynamic path modelling



- The edges are interactions between the genes.
- There is an edge from  $X_j$  to  $X_h$ , if the expression of gene  $j$  influences the expression of gene  $h$ .
- The  $\gamma_{hj}$  are called path coefficients.
- Regression of all parents on each child

Aalen's additive model

- Each  $\gamma_{ij}(t)$  is an ordinary least squares regression coefficient at time  $t$
- Each  $\beta_j(t)$  is an additive hazard regression coefficient at time  $t$ .



- **Direct effect on survival at time  $t$  for gene 1:**

$$\beta_1(t)$$

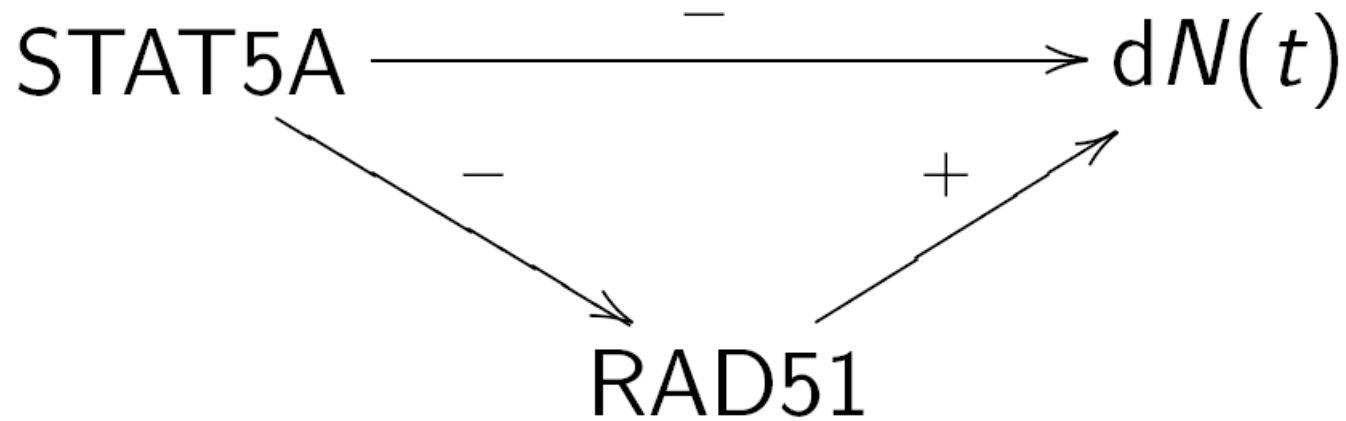
- **Indirect effect on survival at time  $t$  for gene 1 mediated through gene 2:**

$$\gamma_{12}(t)\beta_2(t)$$

- **Total indirect effect on survival at time  $t$  for gene 1:**

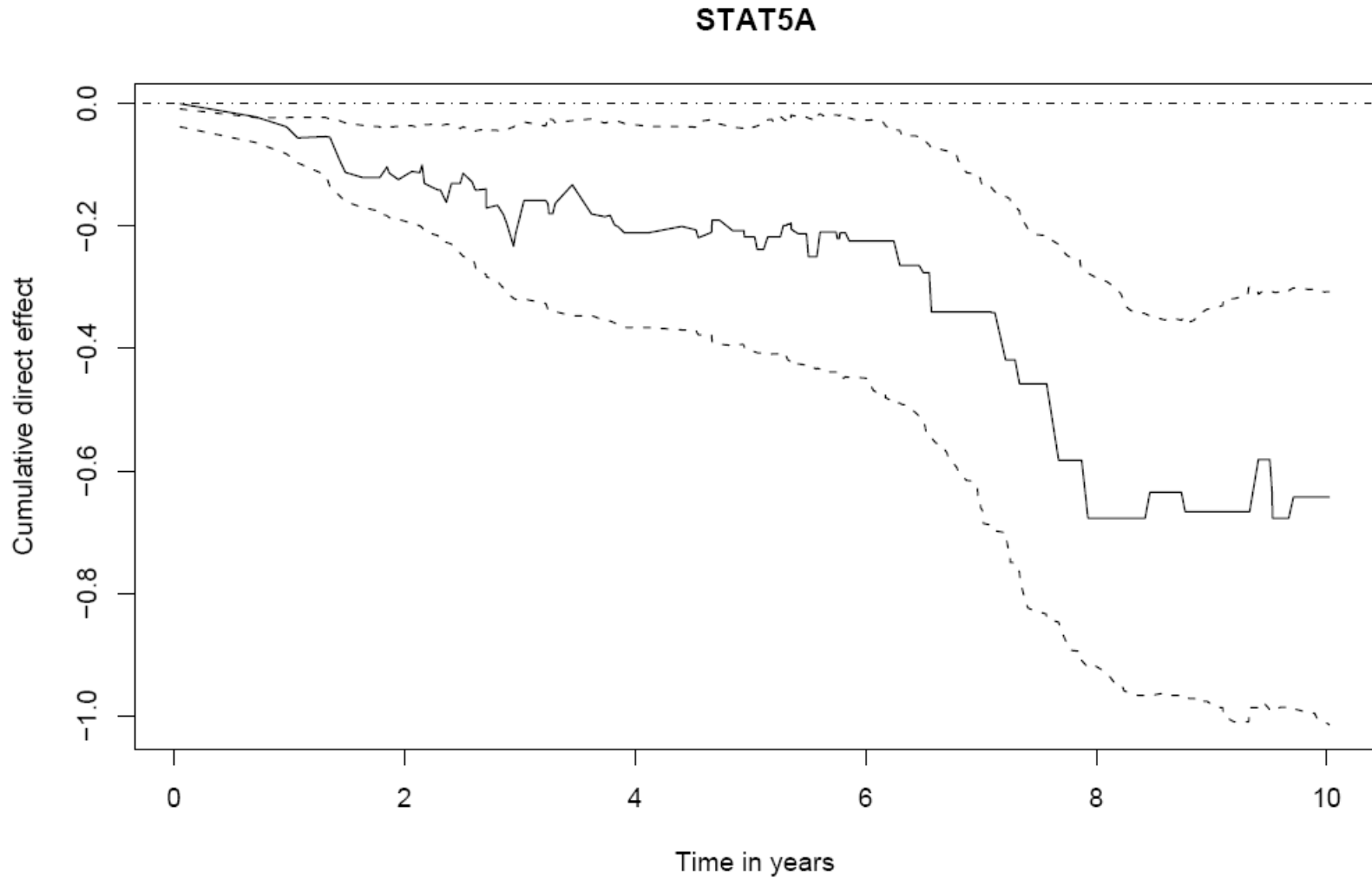
$$\gamma_{12}(t)\beta_2(t) + \underline{\gamma_{13}(t)\gamma_{34}(t)\beta_4(t)}$$

## Dynamic path models for the Dutch breast cancer data set

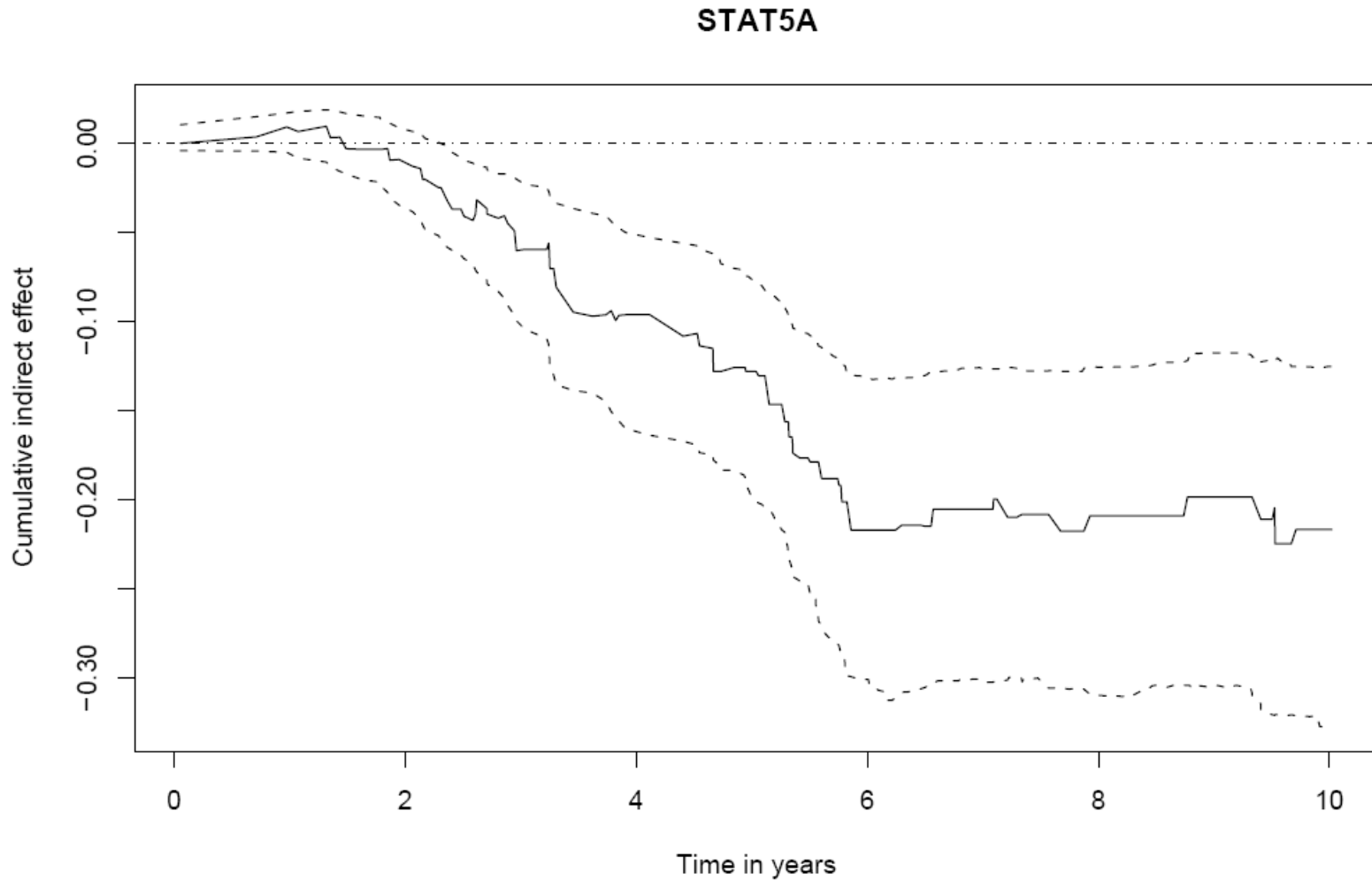


Here STAT5A has both a direct effect and an indirect effect through RAD51. About 37% of the total effect is indirect.

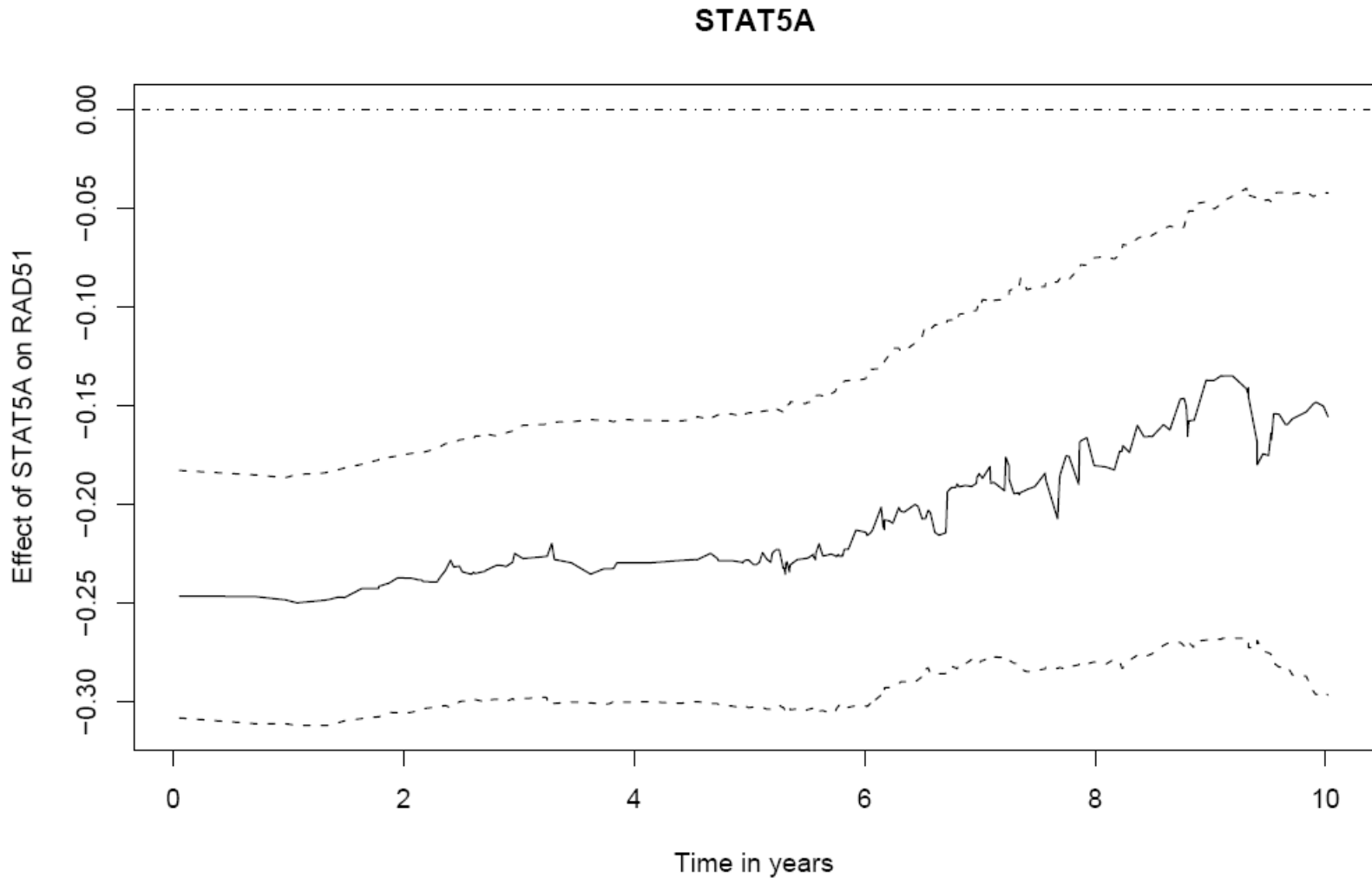
# Direct effect of STAT5A on death:



# Indirect effect of STAT5A on death:



# Effect of STAT5A on RAD51:



(All regressions are repeated at every time point an event happens!)

# HUNTING FOR INDIRECT EFFECTS

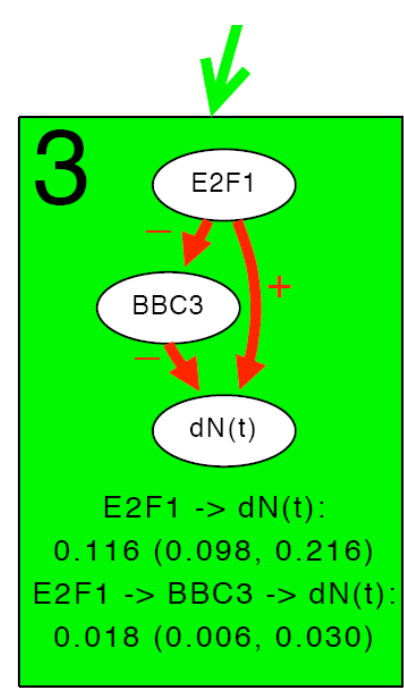
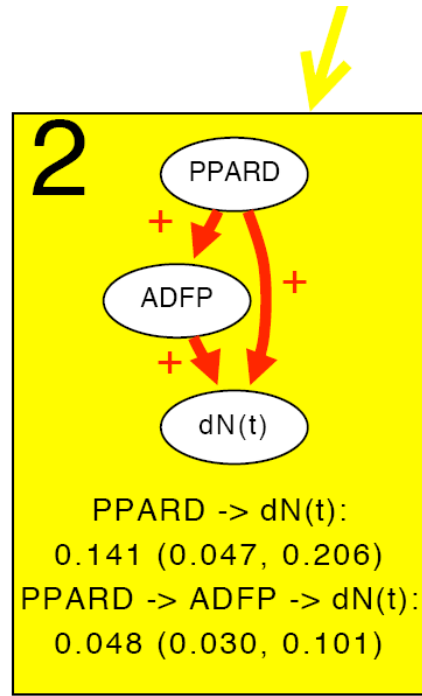
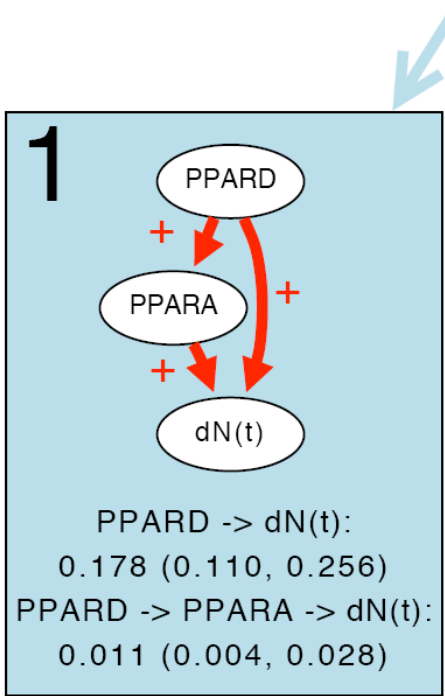
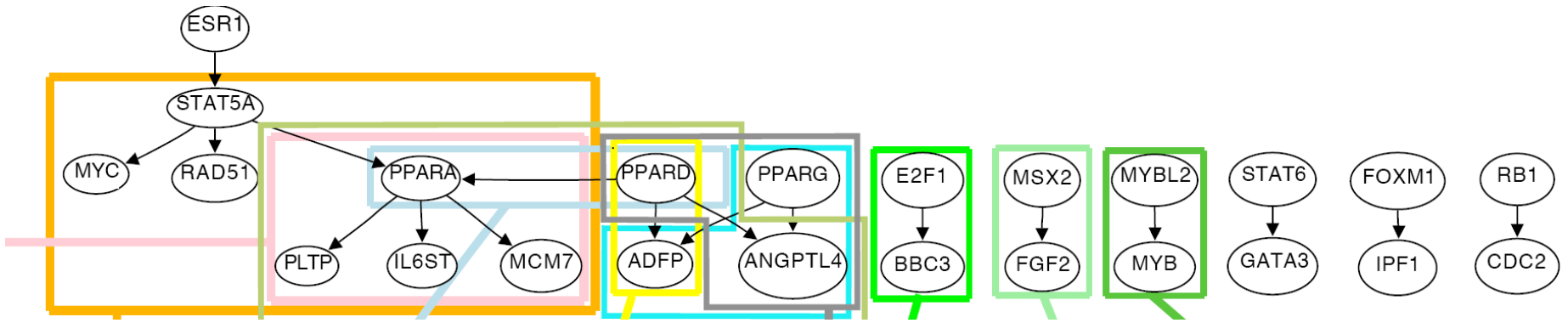
1. Find the genes that are mostly associated with survival (top 1000) (for example: univariate )
2. Determine known pathways of these 1000 genes and their nearest neighbours.

PATHWAY STUDIO®

Software for Visualization and Analysis of Biological Pathways

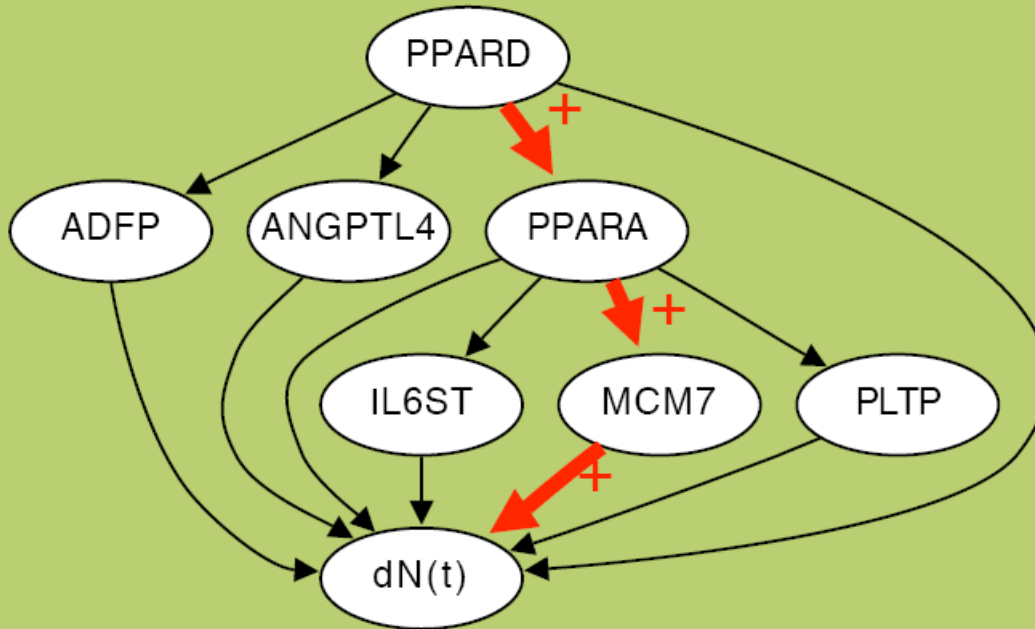
Many rather small pathways are selected in this way.

3. We drop parts of these pathways, where it is unlikely to find indirect effects:  
Regress each pair of genes in the pathways against survival.  
We keep interactions where both genes have significant effect on survival.
4. We run dynamic path analysis on each pathway and all its subgraphs.



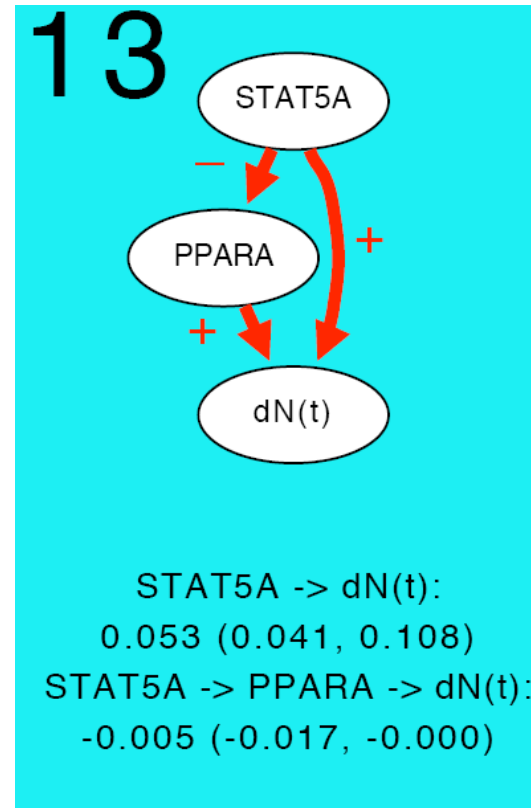
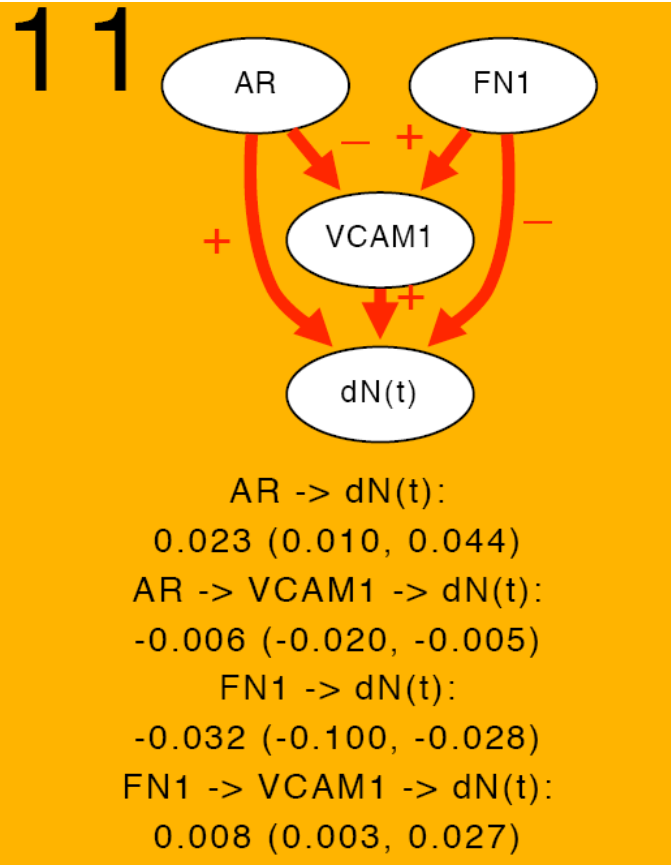


7

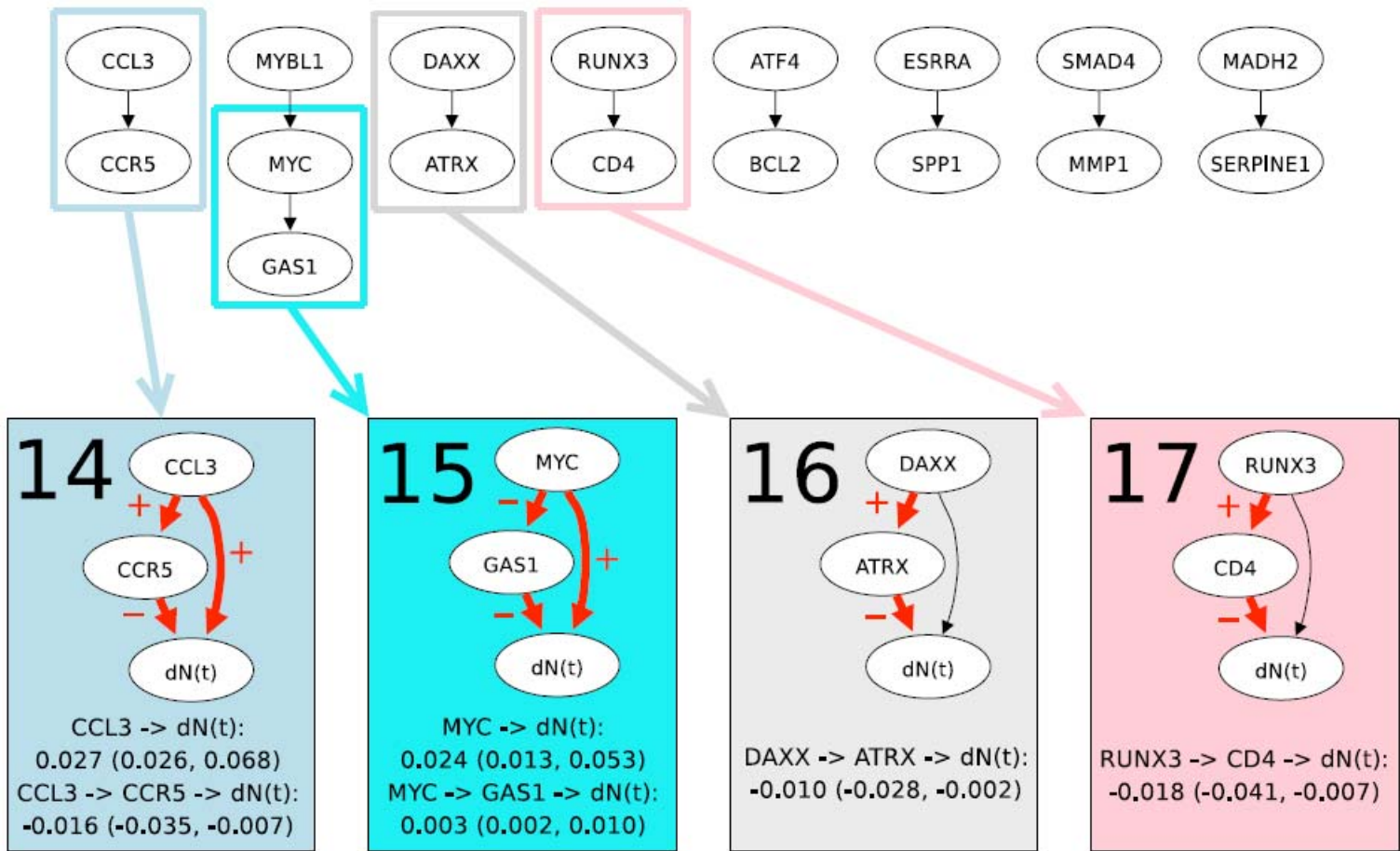


PPARD -> PPARA -> MCM7 -> dN(t):  
0.004 (0.002, 0.018)

# Dynamic path models for the Uppsala breast cancer data set



# Dynamic path models for the diffuse large B-cell lymphoma (DLBCL)



## Tables

**Table 1 — Permutation test**

# interactions	1	2	3	4	5	6	7	8	9	10	11	Our finding
Dutch data	0.844	0.087	0.036	0.017	0.009	0.003	0.003	0.001	0.000	0.000	0.000	19
Uppsala data	0.698	0.150	0.081	0.032	0.018	0.006	0.006	0.002	0.003	0.002	0.002	7
DLBCL data	0.829	0.095	0.043	0.022	0.004	0.005	0.002	0.000	0.000	0.000	0.000	9

This table shows the probabilities of finding the number of interactions listed in the first line, if survival and gene expression were associated at random.

## Summing up

- We detect and quantify (direct and) indirect significant effects on survival of genes interacting in pathways, using dynamic path analysis.
- We detect indirect effects through several target genes of transcription factors like PPAR proteins, E2F1, and MYC in cancer microarray data.
- This points to specific transcription factor - target interactions that play a significant role in the development of aggressive tumor phenotypes.
- Some indirect effects act opposite to the corresponding direct effect on survival.



Method

## Searching for differentially expressed gene combinations

Marcel Dettling<sup>\*</sup>, Edward Gabrielson<sup>\*†</sup> and Giovanni Parmigiani<sup>\*†‡</sup>

Addresses: <sup>\*</sup>Department of Oncology, Johns Hopkins Medical Institutions, Baltimore, MD 21205, USA. <sup>†</sup>Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21205, USA. <sup>‡</sup>Department of Biostatistics, Johns Hopkins Medical Institutions, Baltimore, MD 21205, USA.

Correspondence: Marcel Dettling. E-mail: [dettling@jhu.edu](mailto:dettling@jhu.edu)

Published: 19 September 2005

*Genome Biology* 2005, **6**:R88 (doi:10.1186/gb-2005-6-10-r88)

Received: 4 April 2005

Revised: 23 June 2005

Accepted: 8 August 2005

Comparison of 2 sets of samples from different phenotypes (eg. case/control)  
Which genes show differential expressions?

- How do we do this?

One-gene-at-the-time analysis: test statistics score (t) – p-values –  
multiple comparisons adjustment (not based on  
estimated dependency)

- Do we sometimes use many genes simultaneously?

Yes!      – Multiple testing (but not explicitly)  
             – Classification  
             – Find set of differentially expressed genes for prediction.



Example: (Dettling & Bühlmann, JMA 2004)

Find groups of genes which act together and whose collective expression is strongly associated with an outcome of interest.

*Pelora*, an algorithm based on *penalized logistic regression analysis*, that combines gene selection, gene grouping and sample classification in a Supervised way. “We show that *Pelora* identifies gene groups whose expression centroids have very good **predictive** potential”

Typical result:

“If the centroid of genes A, B and C is high, and the centroid of genes D, E, F and G is low, this is indicative of cancer subtype A”. Such gene groups and their centroids characteristics can be understood as molecular signatures.

Curse of dimensionality: how many pairs/groups of genes can we investigate?

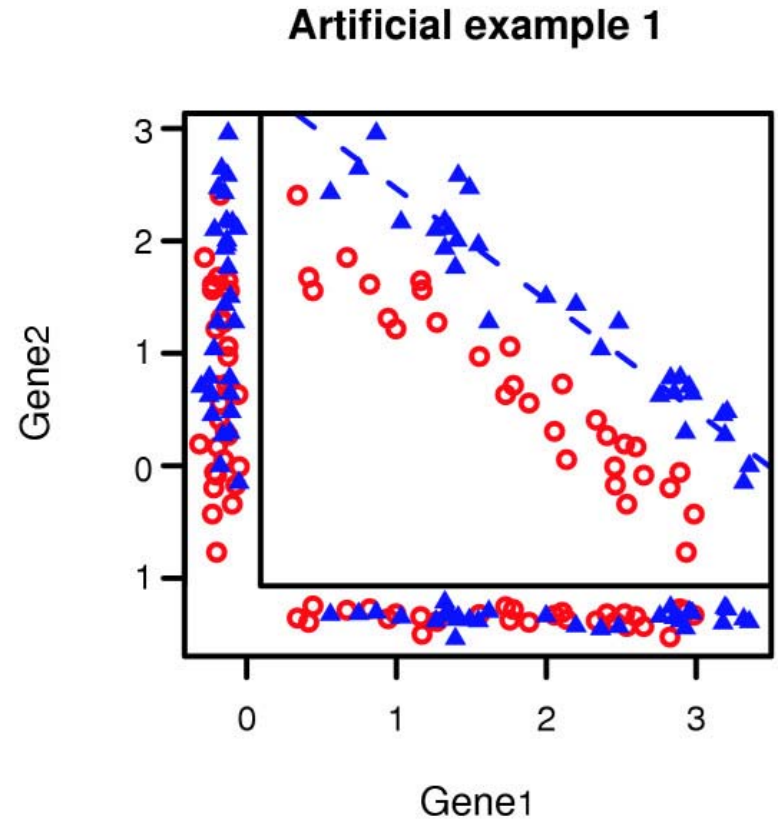
The genes in a group do not need to interact biologically, but have good predictive properties.

Sample cases in blue, controls in red.

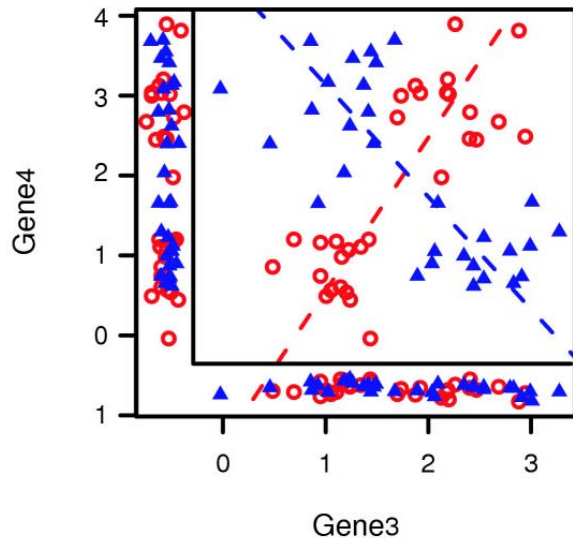
If the sum of the expressions of two genes exceeds 3 units, we find only the blue-triangle phenotype.

A biological mechanism leading to this phenomenon may occur when the two genes are substitutes in a molecular process that is closely linked to the phenotype. ('substitution case').

Neither of the two genes shows a strong association with the phenotype in the univariate marginal distribution, and thus both would not appear in a gene list produced by univariate tests.

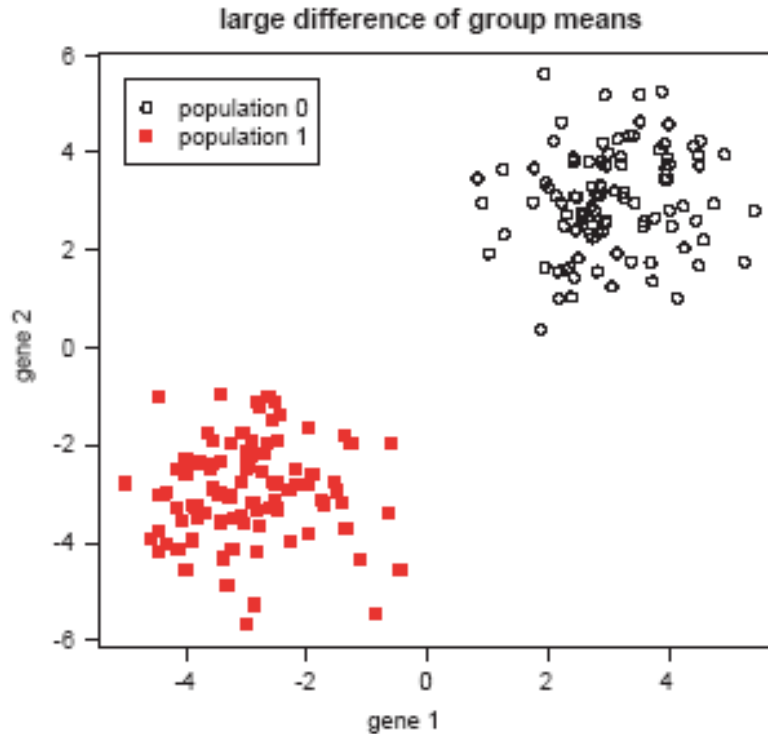


Artificial example 2



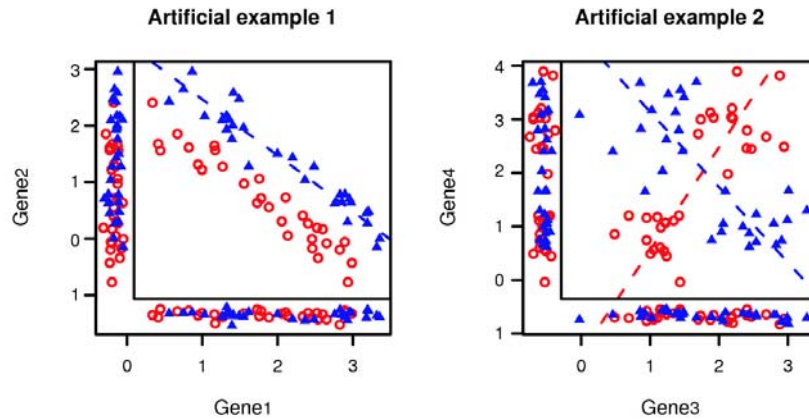
Biologically, this example could reflect an 'on/off situation'. If both genes are off (expressions  $< 1.5$ ), or both genes are on (expressions  $> 1.5$ ), we observe the red-circle phenotype.

In contrast, if only one of the genes is turned on, the blue-triangle phenotype is predominant.



Here the marginal expressions of gene 1 and gene 2 show the difference between the two phenotypes. These genes would be in a list of diff. expressed genes produced by univariate tests.

For these two genes we do not know if they are co-regulated or if they are sitting on two different pathways.



These two genes seem to really interact.

Define **joint differential expression** as good phenotype discrimination by the joint distribution, but not by the univariate marginal distributions of two genes. From a functional genomics perspective, such pairs could represent interesting novel biological interactions, as for example genes that are in the same pathway.

How to find such gene pairs?

Solution:

Plot and test all pairs of genes.

Or compare phenotype prediction based on each pair of gene (a) on its own and (b) in pair together to find those pairs which help in prediction.

Curse of dimensionality:  $p$  genes,  $p(p-1)/2$  gene pairs, usually in the millions.

Prohibitive computational burden.

# Dettling Gabrielson Parmigiani

## CorScor – correlation scoring.

Data:

$n$  samples and  $p$  genes, stored in an  $(n \times p)$  matrix denoted by  $(x_{ig})$ .  
Phenotype information  $y$  is binary.

Application to:

Colon cancer by Alon *et al.* Affymetrix Hum6000 arrays and contains the expressions values of the 2,000 genes in 62 colon tissues, 40 of which were tumors and 22 of which were normal.

Breast cancer dataset from Hedenfalk *et al.*

cDNA microarrays, monitoring 2,654 genes across 22 breast cancer samples, 7 of which were found to carry germline *BRCA1* mutations.

Take genes  $g$  and  $g'$ .

Determine three measures of pairwise dependence among their expression vectors  $(x_{.g})$  and  $(x_{.g'})$ .

- $\rho(g, g')$  using all samples, cases and controls
- $\rho_0(g, g')$  using only samples in class 0 (controls)
- $\rho_1(g, g')$  using only samples in class 1 (cases)

*Here using (Pearson's and Spearman's) correlations, but the general idea can be extended straightforwardly to any easily computed measure of pairwise association among gene expression levels.*

To find gene pairs that jointly discriminate the 2 phenotypes according to the substitution case use the scoring function

$$S(\rho, \rho_0, \rho_1) = | \rho_0 + \rho_1 - \alpha \rho | \quad (1)$$

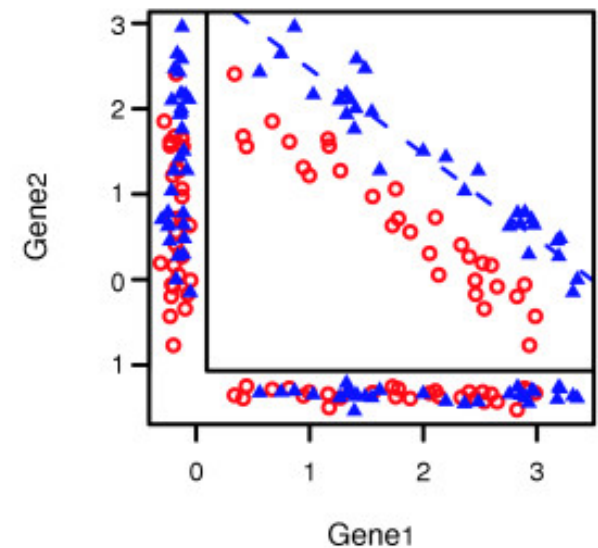
using the Pearson correlation coefficient as dependence measure.

(1) is done on  $(p \text{ times } p)$  correlation matrices.

The higher the score, the more the 2 genes are dependent in each of the two phenotype classes, and the less dependent they are when classes are merged.



$$S(\rho, \rho_0, \rho_1) = | \rho_0 + \rho_1 - \alpha \rho |$$



High correlations within each phenotype arise if the data points within each such class are aligned along a straight line. So we need data sitting on two lines.

Good joint differential expression requires such tight clustering and close-to-parallel line alignment.

Hence, high conditional correlations with concordant sign, and also a shift between the lines, are necessary.

The bigger this shift, and thus the clearer the joint separation, the lower the unconditional correlation  $\rho$  gets. Hence, we diminish the sum by  $\alpha\rho$ .

$\alpha$  governs the balance between separation and parallel alignment.

Empirically good results with  $\alpha \in [1, 2]$ , and use  $\alpha = 1.5$  throughout the paper.

---

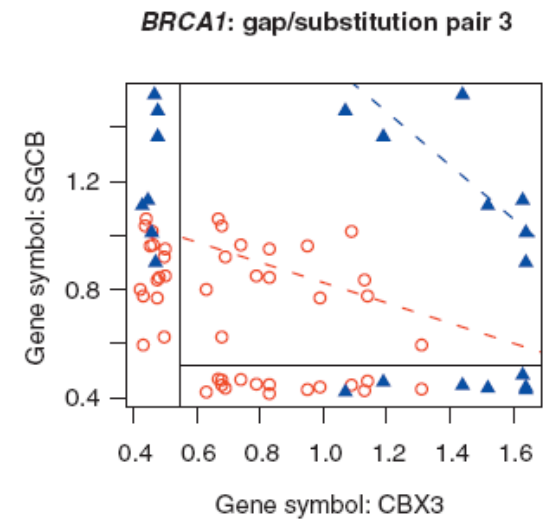
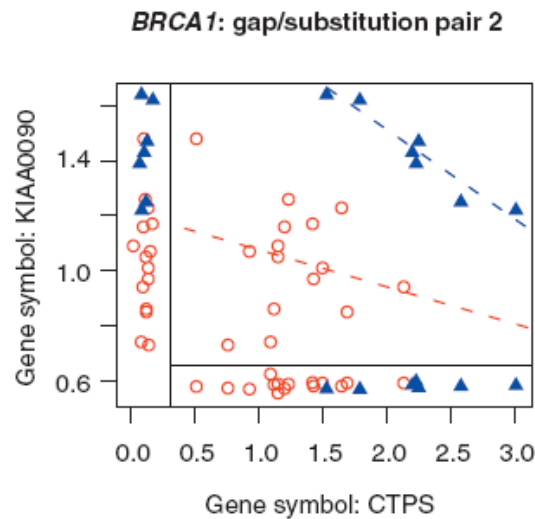
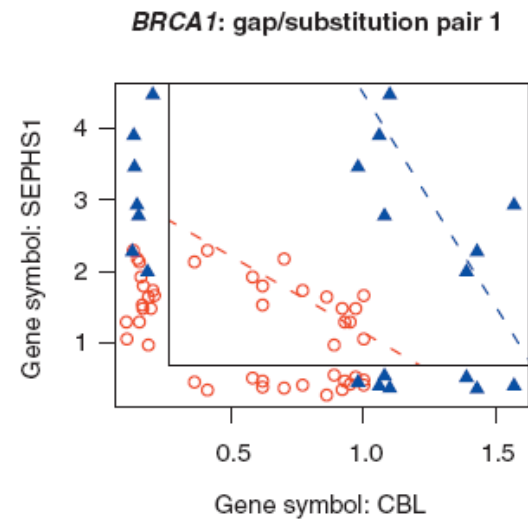
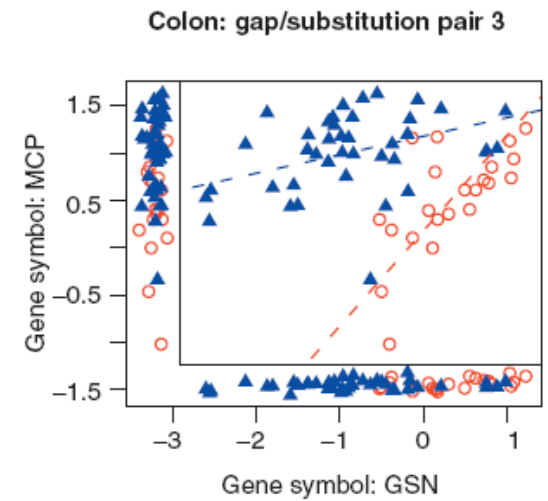
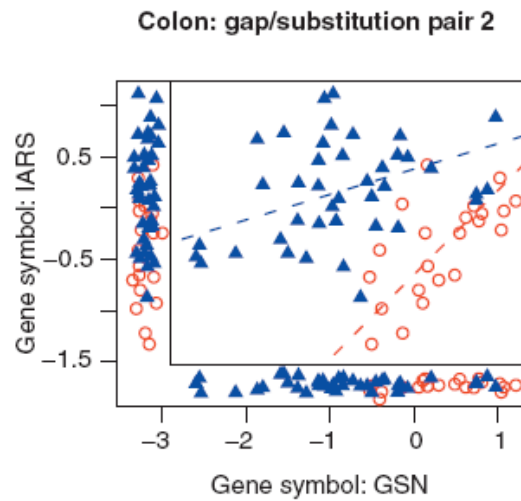
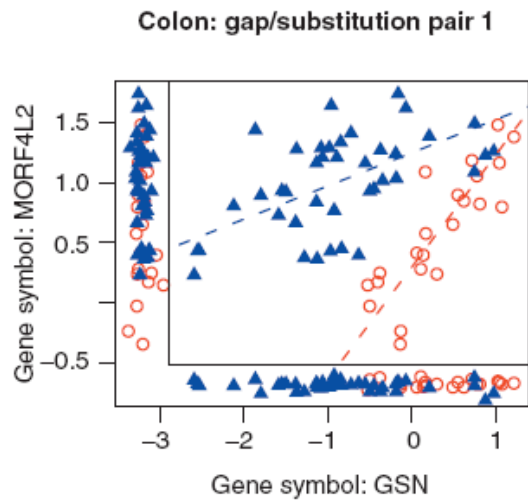
**Correlation coefficients and CorScor values for the gap/substitution scenario**

---

	Colon			BRCA1		
	Pair 1	Pair 2	Pair 3	Pair 1	Pair 2	Pair 3
$\rho$	0.19	-0.01	0.02	0.27	0.32	0.31
$\rho_0$	0.84	0.65	0.67	-0.79	-0.20	-0.38
$\rho_1$	0.53	0.33	0.34	-0.63	-0.96	-0.78
$S(\rho, \rho_0, \rho_1)$	1.09	0.99	0.98	1.82	1.64	1.62

---

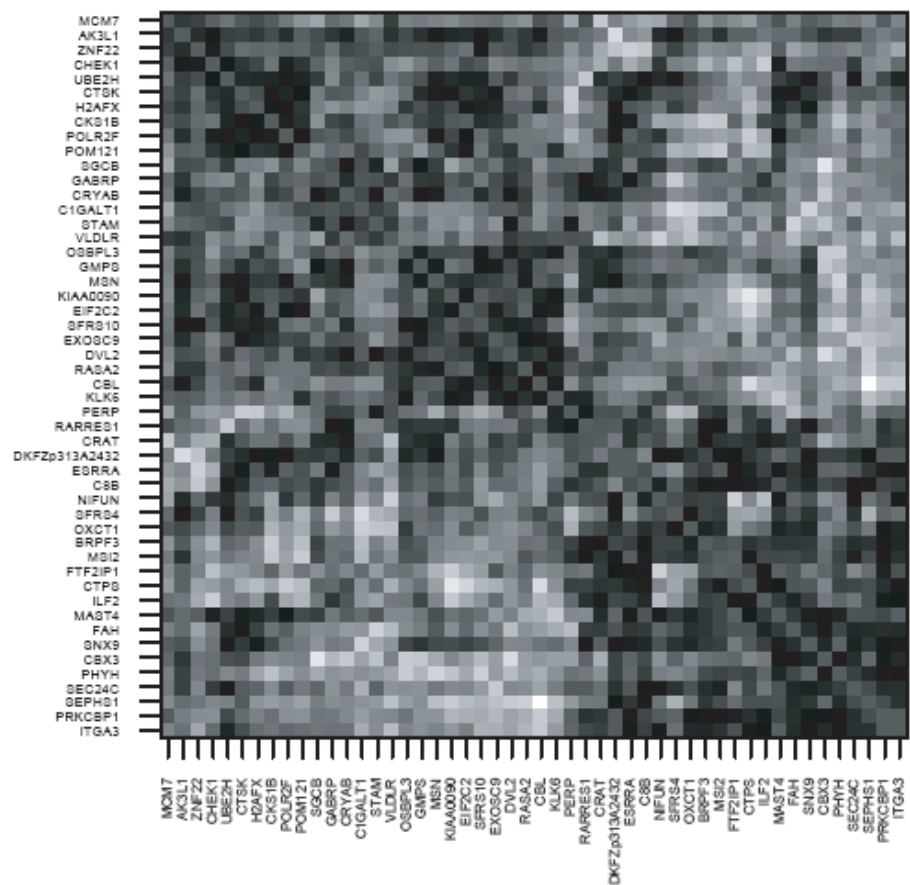
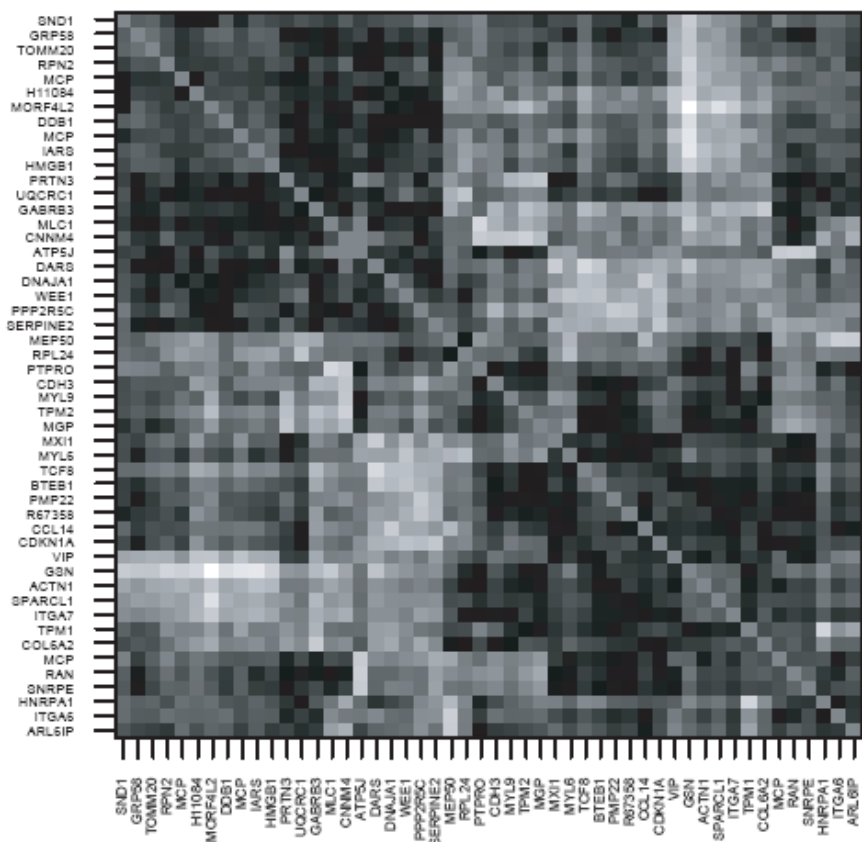
Three highest-scoring gene pairs according to the scoring function (1). As expected, the class-conditional correlations  $\rho_0$  and  $\rho_1$  tend to be high in absolute value and concordant in their signs, whereas the overall correlation is low, and sometimes even has a discordant sign.



Some of the gene pairs are correlated in one group but not in the other. This loss of coregulation can be a biologically relevant feature.

### Colon: gap/substitution scores

### BRCA1: gap/substitution scores



**Figure 3**  
 Symmetric heat map of CorScor values from Equation (1), for the colon and BRCA1 data. Columns and rows are rearranged according to a hierarchical clustering. Displayed are the 50 genes that are involved in the pairs with the highest scores. Black stands for low, grey for intermediate, and white for high score.

Heat map analysis:

Colon data : the most prominent feature is a group of genes (39 to 45 of the matrix). Genes *GSN*, *ACTN1*, *SPARCL1*, *ITGA7*, *TPM1*, and *COL6A2*.

(*GSN*, *ACTN1*, and *SPARCL1*) share a common annotation in the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG). They are all involved in the 'regulation of actin cytoskeleton'.

*ITGA7*, *TPM1*, and *COL6A2* lack pathway annotation in KEGG.

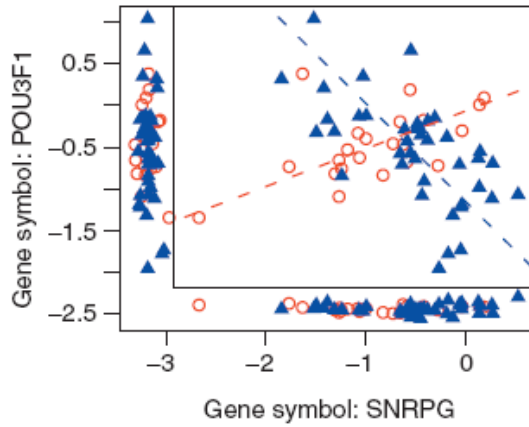
Gene Ontology: *TPM1* has the GO terms 'actin binding' and 'cytoskeleton'. *SPARCL1* is involved in 'calcium ion binding', a term it shares with *GSN* and *ACTN1*.

Here we see how now the authors try to go beyond pairs, and look to groups.

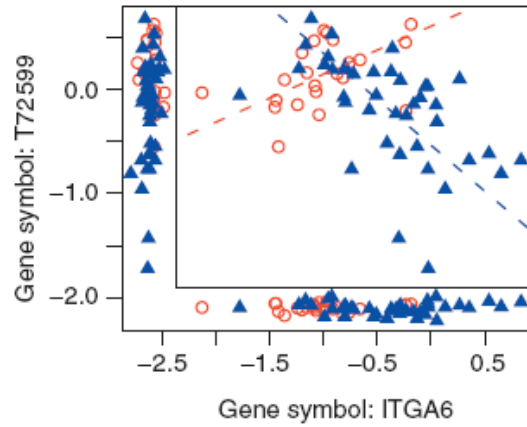
CorScor, compared with established clustering techniques based on the expression values of single genes, is able to capture genes without strong marginal effects. The genes involved in detected pairs do not show pronounced fold changes across the phenotypes, but nevertheless seem to be key in molecular processes closely linked to the phenotype.

On-off gene pairs:  $S(\rho, \rho_0, \rho_1) = |\rho_1 - \rho_0|$  (2)

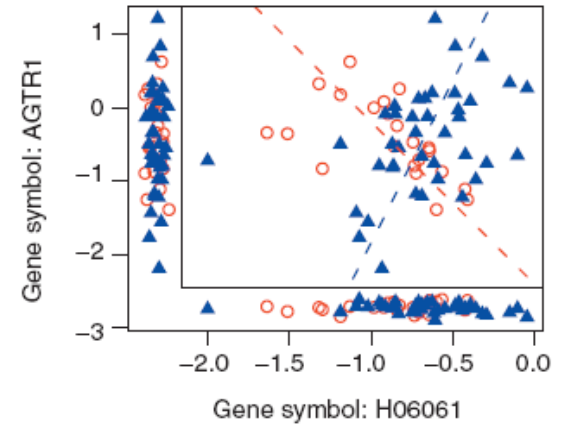
Colon: on/off pair 1



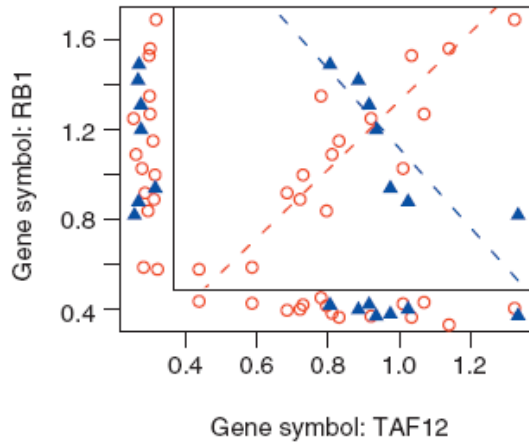
Colon: on/off pair 2



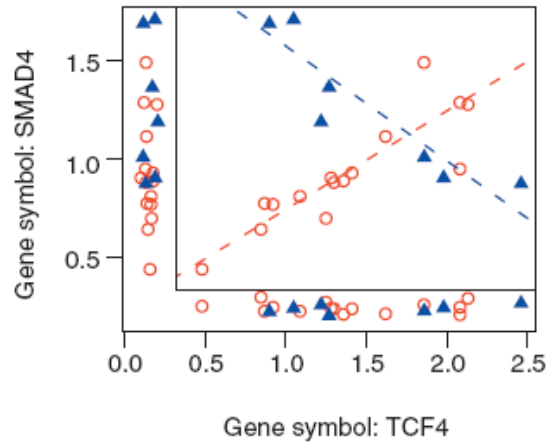
Colon: on/off pair 3



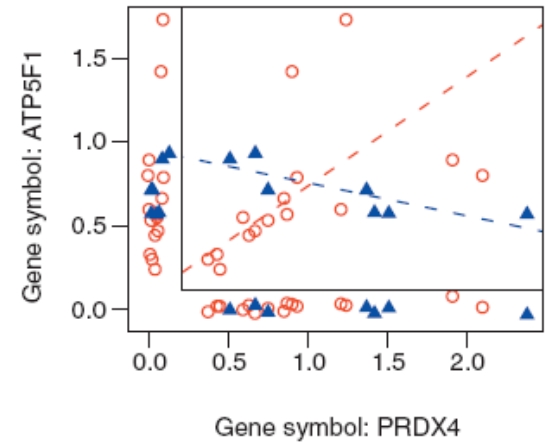
BRCA1: on/off pair 1



BRCA1: on/off pair 2

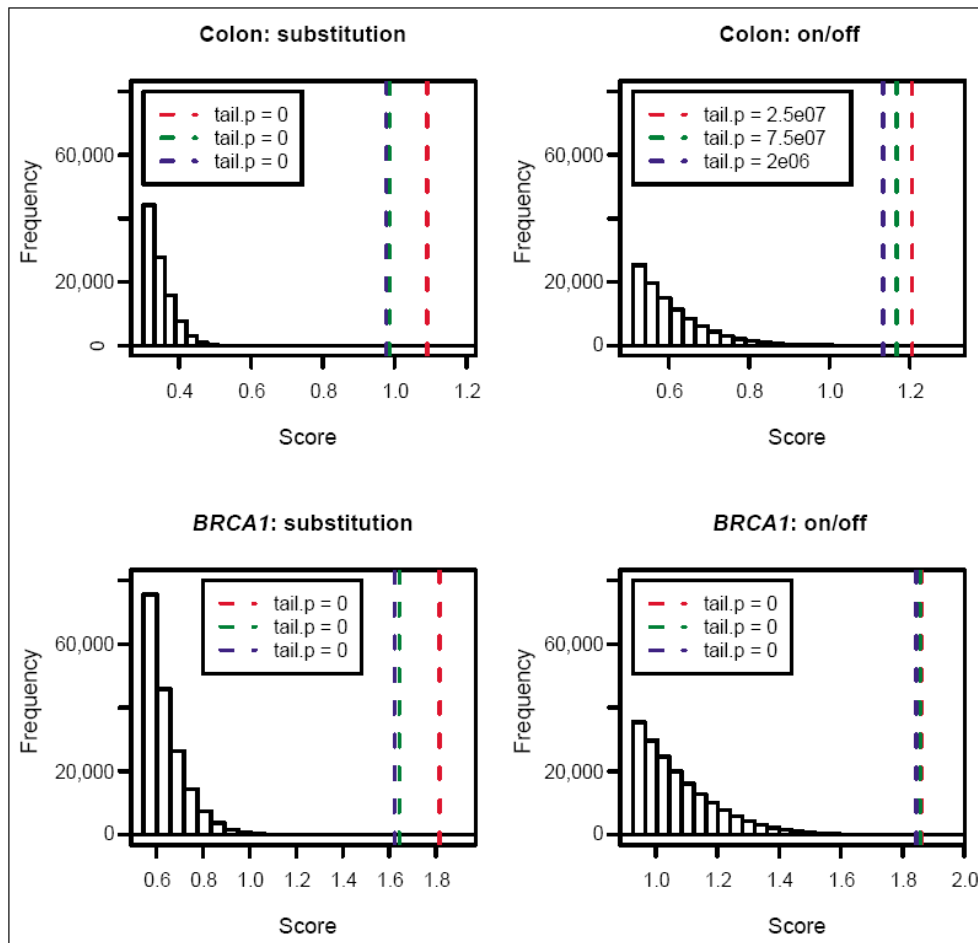


BRCA1: on/off pair 3



Permutation analysis: how many gene pairs achieve promising score values by chance alone?

Author generated 100 noise gene-expression datasets by randomly permuting the phenotype labels. They then run CorScor on each of these 100 false datasets, to obtain an estimated null distribution of CorScor values.



Right tail of the null distribution to the right of the 95% quantile. Vertical lines mark the score value of the top 3 gene pairs.

Plus the empirical false-discovery rate.