

# **Differentially Expressed Genes: notions on multiple testing and p-value adjustments.**

**Dudoit et al., 2002** (symbols for this lecture are slightly different).

One of the experiments described (original data from Callow et al., 2000)

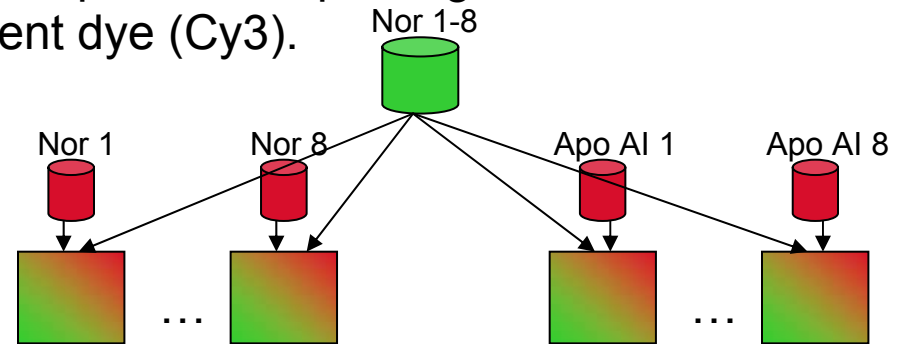
Identify genes with differential expression between:

- “normal” inbred mice (C57Bl/6) livers.
- Apolipoprotein AI (apo AI) knock-out mice livers (apo AI gene plays a role in HDL metabolism; knock-outs have very low HDL cholesterol levels.)

Small number of (biological) replicates available for each condition;  $n_1 = 8$  “normal” mice livers, and  $n_2 = 8$  apo AI knock-out mice livers.

mRNA  $\rightarrow$  (reverse transcription) cDNA samples extracted from each, and labeled with red-fluorescent dye (Cy5). Reference sample formed pooling the 8 “normal” samples, and labeled with a green-fluorescent dye (Cy3).

Hybridization to 16 two-color spotted mice arrays with  $m=6,384$  spots.



Resulting images processed, and signals normalized using lowess-based method within each print-tip group (sector) – lecture on normalization; Yang et al.

$\longrightarrow x^{(k)}_{ji} = \text{normalized red-to-green log ratios } j=1\dots m, i=1\dots n_k, k=1,2$

(but these number could have been obtained from preprocessed affy data...)

gene	Expression →						Stat
	g1	$x(1)_{1,1}$	...	$x(1)_{1,n1}$	$x(2)_{1,1}$	...	
g2	$x(1)_{2,1}$	...	$x(1)_{2,n1}$	$x(2)_{2,1}$	...	$x(2)_{2,n2}$	$t_2$
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
g(m)	$x(1)_{m,1}$	...	$x(1)_{m,n1}$	$x(2)_{m,1}$	...	$x(2)_{m,n2}$	$t_m$
cond	<b>1</b>	...	<b>1</b>	<b>2</b>	...	<b>2</b>	

How do we go about deciding which genes show significant differential (up or down) expression? i.e. for which  $j=1\dots m$  can we reject

$H_j$ : no association of expression with conditions, e.g.  $\mu(2)=\mu(1)$  vs

$H_{j,ALT}$ : association of expression with conditions e.g.  $\mu(2)\neq\mu(1)$

Consider a t-type statistic comparing the two replicates sets for each gene:

$$t_j = \frac{\bar{x}(2) - \bar{x}(1)}{\sqrt{\frac{s^2(2)}{n_2} + \frac{s^2(1)}{n_1}}}$$

**Two issues:**

1. For each  $j$ , it is unlikely that  $t_j$  has a T null distribution...
2. Performing thousands of tests, how do we control for false positives? (i.e. false rejections; type-I errors).

for  $j$   $p_j = \Pr(|T| \geq |t_j|)$

using a T-distribution

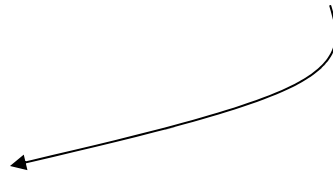
for  $j$   $p_j^* = \frac{1}{B} \#(|t_j(b)| \geq |t_j|)$

using simulated null distrib  $t_j(1) \dots t_j(B)$

**OR RATHER...**  
Random permutations of cond labels

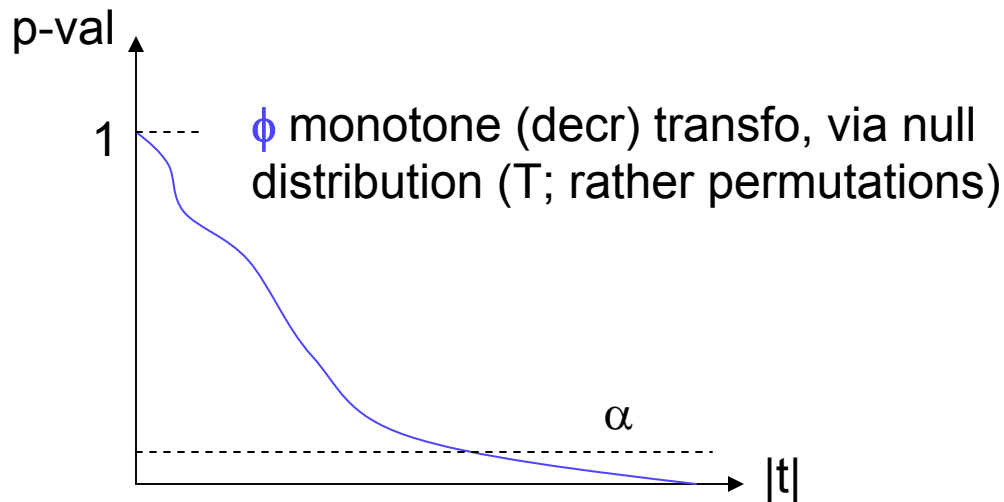
g1		$t_1$	$p_1$
g2		$t_2$	$p_2$
		.	.
		.	.
		.	.
g(m)		$t_m$	$p_m$
	<b>Cond labels</b>	<b>Stat</b>	<b>T p-vals</b>

	$t(1)_1$	...	$t(B)_1$	$p^*_1$
	$t(1)_2$		$t(B)_2$	$p^*_2$
	.		.	.
	.		.	.
	.		.	.
	$t(1)_m$		$t(B)_m$	$p^*_m$
<b>perm labels</b>	<b>Stat</b>			<b>perm p-vals</b>

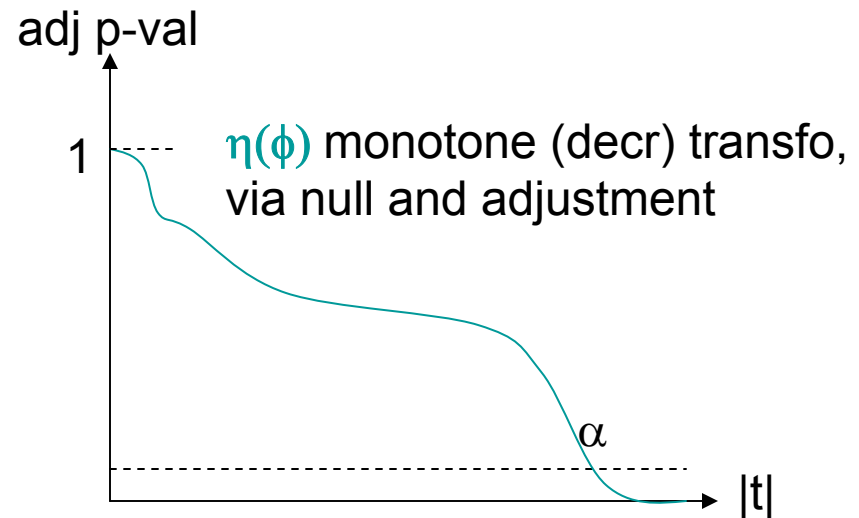
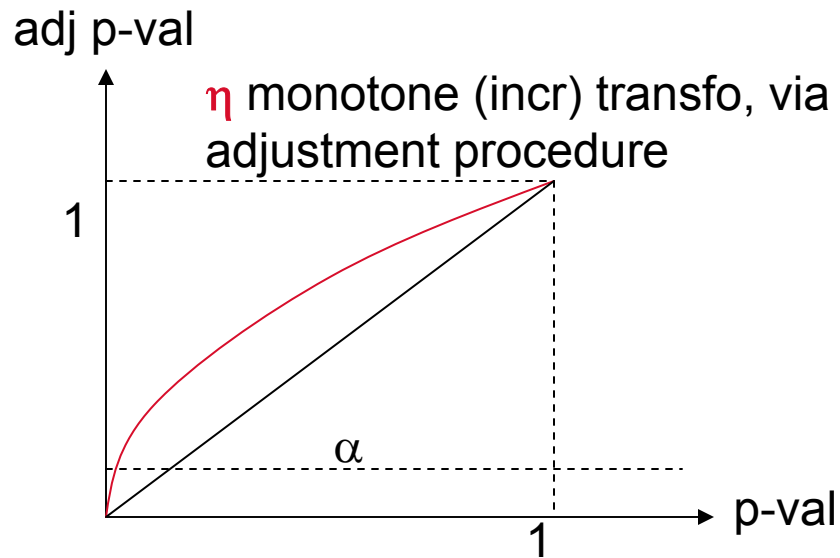


Note: with one set of  $B$  random perm's of the labels, we obtain simulated null distrib's and thus empirical p-val's, for all j's.

Note also this type of MA data arrangement is transposed wrt standard stat notation (rows = "variables", clmns = "observations".)



Say we declare diff expressed all genes with  $p\text{-val} \leq \alpha$ . This way we control the probability of type-I error (false positive) for each test individually, but not for the whole family of  $m$  tests.



Want an adjustment procedure such that, when we threshold on adjusted p-values, we “control” the probability of type-I error for the whole family of tests.

## Statistics: some notions on multiple testing

Complete null hyp:  $\mathbf{H}_c = \cap \mathbf{H}_j$  (no assoc btw any of the genes and the conditions)

FWER (family-wise type-I error rate) =  $\Pr(\text{at least one rejection} \mid \mathbf{H}_c \text{ true})$   
is what we would like to control; note this is a very “stringent” aim!

$P_j$  random vbl of which  $p_j$  is a realization (obtained with ref or simul null distrib).

Upper bound

Exact, but more assumptions

$$\Pr\left(\bigcup_{j=i\dots m} P_j \leq p \mid H_c\right) \leq \sum_{j=i\dots m} \Pr(P_j \leq p \mid H_c) \stackrel{(A)}{=} mp$$

$$(A): P_j \mid H_c \sim Un[0,1], \forall j$$

BONFERRONI CORRECTION:

$$\tilde{p} = \eta(p) = \min(mp, 1)$$

$$\text{better known: } \tilde{p}_j \leq \alpha \Leftrightarrow p_j \leq \frac{\alpha}{m}$$

$$\begin{aligned} \Pr\left(\bigcup_{j=i\dots m} P_j \leq p \mid H_c\right) &= 1 - \Pr\left(\bigcap_{j=i\dots m} P_j > p \mid H_c\right) \\ &\stackrel{(B)}{=} 1 - \prod_{j=i\dots m} \Pr(P_j > p \mid H_c) \\ &\stackrel{(A)}{=} 1 - (1 - p)^m \end{aligned}$$

$$(B): (P_j \dots P_m) \mid H_c \text{ independent}$$

SIDAK CORRECTION:

$$\tilde{p} = \eta(p) = 1 - (1 - p)^m$$

Both are single-step (same adjustment performed on all p-values, regardless of their size i.e. strength of evidence). Very simple to compute, but very conservative: that is, in order to control FWER through such simple corrections, we are likely to generate a large number of false negatives (lose family-wise power).

Make adjustments less conservative (I): Step-down adjustments; adapt the correction to the ordering of the p-values.

$p_{r_1} \leq p_{r_2} \dots \leq p_{r_m}$  p-val's in increasing order  
 $(|t_{r_1}| \geq |t_{r_2}| \dots \geq |t_{r_m}|$  stats in decreasing order)  
**HOLM (1) STEP-DOWN CORRECTION:**  
 $\tilde{p}_{r_j} = \eta(p_{r_j}) = \max_{k=1 \dots j} \left\{ \min\{(m - k + 1)p_{r_k}, 1\} \right\}$   
 $\min\{(m - k + 1)p_{r_k}, 1\}$  Bonferroni for  $(m - k + 1)$   
 or HOLM (2):  
 $\tilde{p}_{r_j} = \eta(p_{r_j}) = \max_{k=1 \dots j} \left\{ 1 - (1 - p_{r_k})^{m-k+1} \right\}$   
 $1 - (1 - p_{r_k})^{m-k+1}$  Sidak for  $(m - k + 1)$   
 $\max_{k=1 \dots j}$  enforce *j*th rank for  $\tilde{p}_{r_j}$  ( $\eta$  monotone)

rnk	P-val	Bonferroni below	Enforce rnk
1	$p_{r_1}$	$\min\{m p_{r_1}, 1\}$ strongest	itself
j	$p_{r_j}$	$\min\{(m-j+1)p_{r_1}, 1\}$	max above
m	$p_{r_m}$	$\min\{p_{r_m}, 1\}$ No correction	max all

The diagram shows a table with four rows. The first row (rank 1) has a p-value  $p_{r_1}$  and a Bonferroni adjustment  $\min\{m p_{r_1}, 1\}$  labeled 'strongest'. A blue bracket labeled 'itself' spans from the p-value to the adjustment. A red arrow points from the adjustment down to the adjustment for rank  $m$ . The second row (rank  $j$ ) has a p-value  $p_{r_j}$  and a Bonferroni adjustment  $\min\{(m-j+1)p_{r_1}, 1\}$ . A red bracket labeled 'max above' spans from the p-value to the adjustment. A red arrow points from the adjustment down to the adjustment for rank  $m$ . The third row (rank  $m$ ) has a p-value  $p_{r_m}$  and a Bonferroni adjustment  $\min\{p_{r_m}, 1\}$  labeled 'No correction'. A blue bracket labeled 'max all' spans from the p-value to the adjustment. A blue arrow points from the adjustment up to the adjustment for rank  $j$ . A blue arrow points from the adjustment up to the adjustment for rank 1.

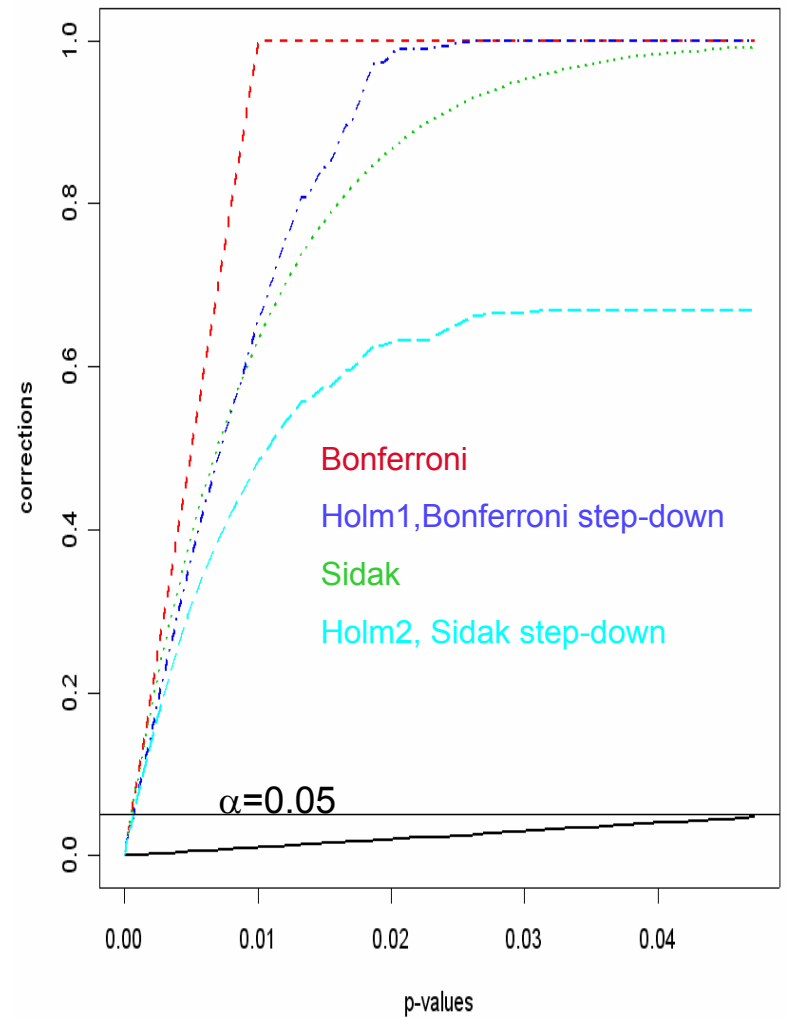
Adjustment is stronger earlier in the ranking (for smaller p-val's)

Monotonicity: If a number coming before *j*th is larger, "bump" *j*th up to max

(sorted, p-values generated randomly, m=100)

...Look at some numbers

	p_vals	bonf_p_vals	sidak_p_vals	holm1_p_vals	holm2_p_vals
[1,]	2.796785e-06	0.0002796785	0.0002796398	0.0002796785	0.0002796398
[2,]	4.593479e-05	0.0045934790	0.0045830501	0.0045475442	0.0045373237
[3,]	5.506744e-05	0.0055067440	0.0054917604	0.0053966091	0.0053822213
[4,]	1.005774e-04	0.0100577358	0.0100078267	0.0097560037	0.0097090542
[5,]	1.403472e-04	0.0140347186	0.0139376623	0.0134733299	0.0133839037
[6,]	1.570989e-04	0.0157098880	0.0155883462	0.0149243936	0.0148147318
[7,]	1.865423e-04	0.0186542266	0.0184830213	0.0175349730	0.0173837373
[8,]	1.877971e-04	0.0187797148	0.0186062054	0.0175349730	0.0173837373
[9,]	2.004890e-04	0.0200488987	0.0198512261	0.0184449868	0.0182777346
[10,]	2.107069e-04	0.0210706938	0.0208524316	0.0191743313	0.0189936551
[11,]	2.188548e-04	0.0218854794	0.0216500732	0.0196969315	0.0195063279
[12,]	2.647850e-04	0.0264784988	0.0261344315	0.0235658639	0.0232934052
[13,]	2.763125e-04	0.0276312476	0.0272567106	0.0243154978	0.0240255369
[14,]	3.055362e-04	0.0305536180	0.0300961020	0.0265816477	0.0262354204
[15,]	3.435432e-04	0.0343543165	0.0337766100	0.0295447122	0.0291174621
[16,]	3.542952e-04	0.0354295189	0.0348152997	0.0301150911	0.0296713274
[17,]	3.551637e-04	0.0355163656	0.0348991489	0.0301150911	0.0296713274
[18,]	3.575531e-04	0.0357553053	0.0351298044	0.0301150911	0.0296713274
[19,]	3.616006e-04	0.0361600596	0.0355204012	0.0301150911	0.0296713274
[20,]	3.902363e-04	0.0390236322	0.0382793436	0.0316091421	0.0311207726
...					
[90,]	4.229125e-02	1.0000000000	0.9867154929	1.0000000000	0.6691709475
[91,]	4.249260e-02	1.0000000000	0.9869918991	1.0000000000	0.6691709475
[92,]	4.383928e-02	1.0000000000	0.9886997047	1.0000000000	0.6691709475
[93,]	4.391367e-02	1.0000000000	0.9887872880	1.0000000000	0.6691709475
[94,]	4.428082e-02	1.0000000000	0.9892097825	1.0000000000	0.6691709475
[95,]	4.446889e-02	1.0000000000	0.9894200670	1.0000000000	0.6691709475
[96,]	4.582042e-02	1.0000000000	0.9908164239	1.0000000000	0.6691709475
[97,]	4.608705e-02	1.0000000000	0.9910695240	1.0000000000	0.6691709475
[98,]	4.690354e-02	1.0000000000	0.9918024232	1.0000000000	0.6691709475
[99,]	4.722406e-02	1.0000000000	0.9920735641	1.0000000000	0.6691709475
[100]	4.72805e-02	1.0000000000	0.9921204224	1.0000000000	0.6691709475



Make adjustments less conservative (II): Step-down adjustments that account for/exploit dependencies among p-values (very useful for MA data because genes' expression levels are highly interdependent).

$$\Pr\left(\bigcup_{j=i\dots m} P_j \leq p \mid H_c\right) = \Pr\left(\min_{j=i\dots m} P_j \leq p \mid H_c\right) \text{ with no assumptions}$$

$p_{r_1} \leq p_{r_2} \dots \leq p_{r_m}$  p-vals in increasing order

( $|t_{r_1}| \geq |t_{r_2}| \dots \geq |t_{r_m}|$  stats in decreasing order)

**WESTFALL YOUNG STEP-DOWN minP CORRECTION:**

$$\tilde{p}_{r_j} = \eta(p_{r_j}) = \max_{k=1\dots j} \left\{ \Pr\left(\min_{l=k\dots m} P_{r_l} \leq p_{r_k} \mid H_c\right) \right\}$$

$\Pr\left(\min_{l=k\dots m} P_{r_l} \leq p_{r_k} \mid H_c\right)$  need joint distrib of  $(P_1 \dots P_m) \mid H_c$

**WESTFALL YOUNG STEP-DOWN maxT CORRECTION:**

$$\tilde{p}_{r_j} = \eta(p_{r_j}) = \max_{k=1\dots j} \left\{ \Pr\left(\max_{l=k\dots m} |T_{r_l}| \geq |t_{r_k}| \mid H_c\right) \right\}$$

$\Pr\left(\max_{l=k\dots m} |T_{r_l}| \geq |t_{r_k}| \mid H_c\right)$  need joint distrib of  $(T_1 \dots T_m) \mid H_c$

$\max_{k=1\dots j}$  enforce  $j$ th rank for  $\tilde{p}_{r_j}$  ( $\eta$  monotone)

maxT is a parallel construction to minP; corrections are exactly equivalent only if  $T_j$ 's all have the same marginal null distribution, however maxT is convenient computationally.

When we use random permutations to compute (unadjusted) p-val's corresponding to each observed  $t_j$ , we are actually simulating at once the joint null distribution for  $(T_1 \dots T_m) | H_c$ . So it is convenient to compute directly the Westfall Young step-down maxT adjusted p-val's:

	incr order	...any		...any		...any		...any		decr order	
Rnk (obs)	Stat obs	Stat perm 1	Max below	...	Stat perm b	Max below	...	Stat perm B	Max below	ratios	Enforce rnk
1	$ t_{r1} $	$ t_{r1}(1) $	$u_{r1}(1)$ =max all	...	$ t_{r1}(b) $	$u_{r1}(b)$ =max all	...	$ t_{r1}(B) $	$u_{r1}(B)$ =max all	$\#(u_{r1}(b) \geq  t_{r1} ) / B$	itself
.				...			...				
j	$ t_{rj} $	$ t_{rj}(1) $	$u_{rj}(1)$ =max below	...	$ t_{rj}(b) $	$u_{rj}(b)$ =max below	...	$ t_{rj}(B) $	$u_{rj}(B)$ =max below	$\#(u_{rj}(b) \geq  t_{rj} ) / B$	max above
.				...			...				
m	$ t_{rm} $	$ t_{rm}(1) $	$u_{rm}(1)$ = $ t_{rm}(1) $ itself	...	$ t_{rm}(b) $	$u_{rm}(b)$ = $ t_{rm}(1) $ itself	...	$ t_{rm}(B) $	$u_{rm}(B)$ = $ t_{rm}(1) $ itself	$\#(u_{rm}(b) \geq  t_{rm} ) / B$	max all

Count, and thus adj p-val, for  $r_j$  increases (evidence for diff expression erodes) when  $r_j$  itself, or any gene that ranked below it in the observed stat, has a permuted stat  $|t_{(.)}(b)| \geq |t_{rj}|$ .

This mechanism exploits all genes as a group (thus interdependencies among their  $T_j$ 's).

## Results for apo AI experiment from Dudoit et al. (2002)

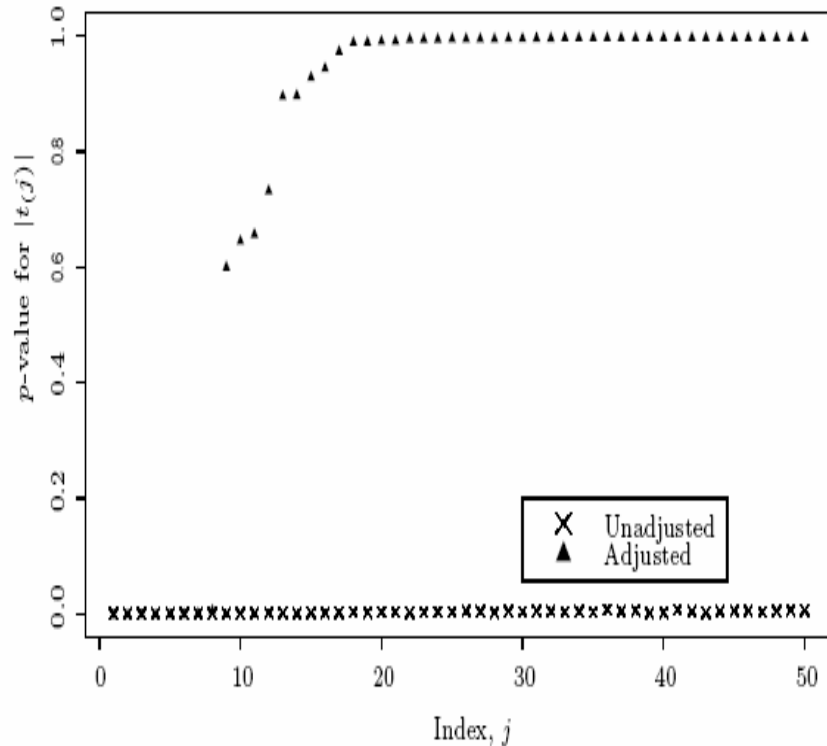


Figure 5. *Apo AI*. Westfall and Young maxT adjusted  $p$ -values (filled triangles) and unadjusted  $p$ -values (crosses) for the 50 genes with the largest absolute  $t$ -statistic.

For the first 8 genes, triangle (adjusted) and cross (unadjusted)  $p$ -vals are both down below 0.01. Without adjustment, we would have declared these 50, and probably many more differentially expressed. On the other hand, with a more conservative adjustment (e.g. Bonferroni) we would have missed some of them. Note all 8 are downregulated (have repressed expression in apo AI knock-out compared to normal mice liver).

Table 1. *Apo AI*. Genes with maxT adjusted  $p$ -values  $\leq 0.01$ . For each gene, the table lists the gene name, the permutation adjusted  $p$ -value ( $\tilde{p}^*$ ), the two-sample  $t$ -statistic ( $t$ ), the numerator (Num) and denominator (Den) of the  $t$ -statistic.

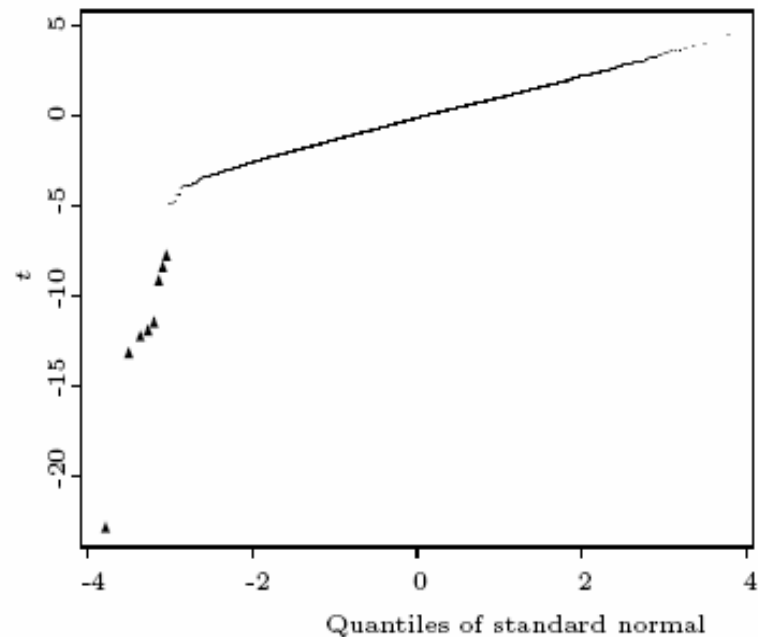
Gene name	$\tilde{p}^*$	$t$	Num	Den
Apo AI	0.00	-22.85	-3.19	0.14
Sterol C5 desaturase	0.00	-13.14	-1.06	0.08
Apo AI	0.00	-12.21	-1.90	0.16
Apo CIII	0.00	-11.88	-1.02	0.09
Apo AI	0.00	-11.44	-3.09	0.27
EST AA080005	0.00	-9.11	-1.02	0.11
Apo CIII	0.00	-8.36	-1.04	0.12
Sterol C5 desaturase	0.01	-7.72	-1.04	0.13

## Using q-q plots:

quantiles of the observed  $t$  statistics against quantiles of a  $N(0,1)$ . Without any claim that they should approximate a normal, this plot will give us an idea of which genes present unusual  $t$  statistic relative to the others. In a sense, exploiting behavior of the whole group of genes, without computing nor adjusting  $p$ -values.

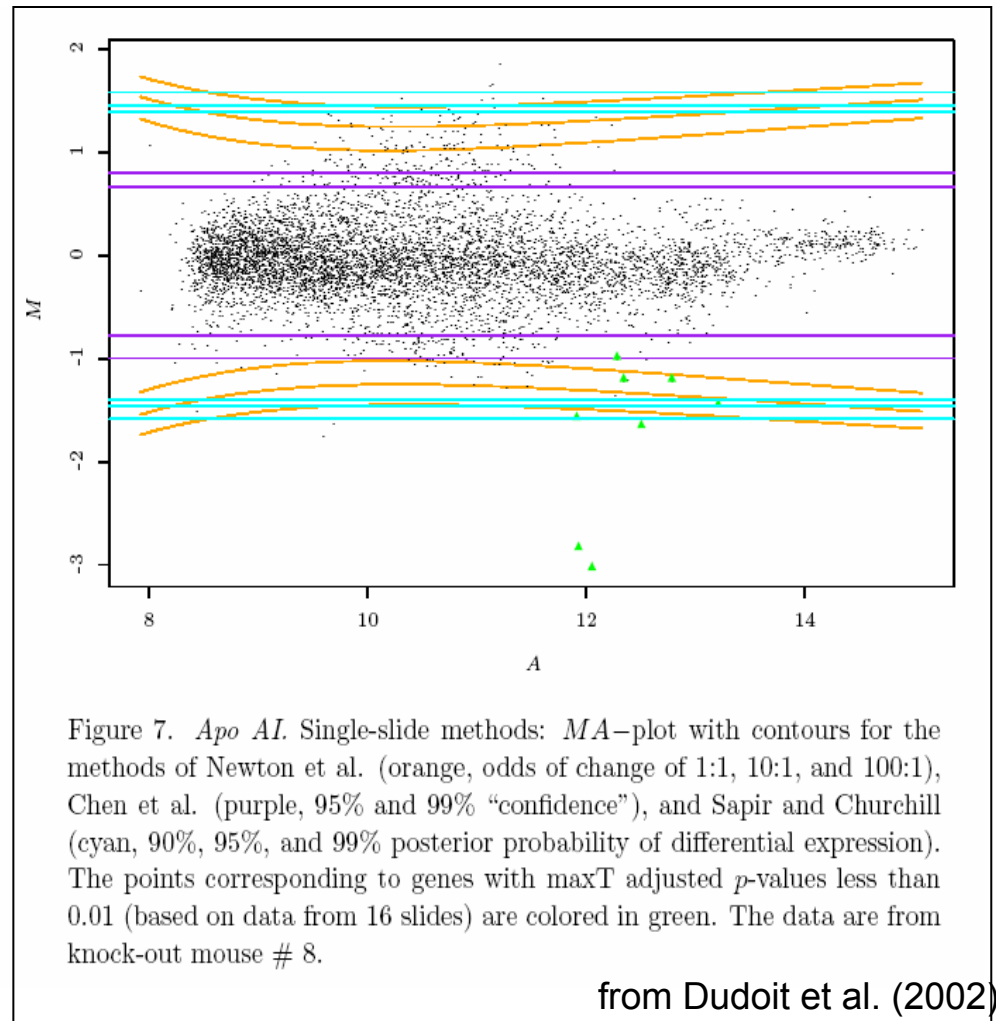
Note: in principle, could replace quantiles of the  $N(0,1)$  with quantiles for the  $t$ 's as obtained through random permutations.

Figure 4. *Apo AI*. Histogram and normal Q-Q plot for two-sample  $t$ -statistics. The points corresponding to genes with  $\max T$  adjusted  $p$ -values less than 0.01 are indicated by filled triangles.



from Dudoit et al. (2002)

**Remark 1:** Methods to identify differentially expressed genes btw two sets of expression numbers (two conditions), without replicates (e.g. R and G on one spotted array). They exist (see references in Dudoit et al.), but require pretty heavy modeling of the process generating the data, as to be able to eventually consider genes as “replicate observations” from this process – and thus have information for inference.



**Remark 2:** identifying genes that are differentially expressed across several (more than two) conditions for each of which replicates are available. The approach described in Dudoit et al. works; what we need to do is to change statistic . For instance, for each gene we could compute an F statistic from a one-factor anova, (where factor “levels” are the conditions).

**Remark 3:** even with step-down corrections exploiting interdependencies, trying to control FWER (probability of at least one false positive i.e. false rejection, in the family of tests) may simply be too conservative all together. Maybe we are willing to live with some false positive, and control instead

$$\text{FDR} = \text{false discovery rate} = \text{Expected ratio} \frac{\# \text{ false positives}}{\# \text{ positives}}$$

This is discussed both in Efron et al. (2001) and in Tusher et al. (2001; SAM – ***Significance Analysis of Microarrays***).

Finally, more detail on multiple testing and p-values adjustments can be found in Dudoit et al. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, Vol. 18, No. 1, p. 71-103.