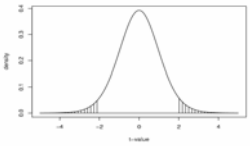
Computational Genomics

10-810/02-710, Spring 2009

Differential Analysis of Microarray Gene Expression Data

Eric Xing



Lecture 16, March 16, 2009

Reading: class assignment

Outline

- Motivation & examples
- Univariate hypothesis testing
- Multiple hypothesis testing
- Results for the two examples
- Discussion

Introduction

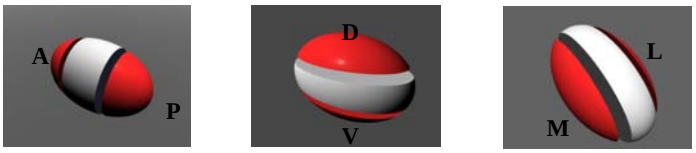
- Many microarray experiments are carried out to find genes which are differentially expressed between two (or more) samples of cells:
  - cells (from the liver, say), in a mouse with a gene knocked out, compared with liver cells in a normal mouse of the same strain
  - cells in one region of the brain (say cerebellum), compared with cells in a different region (say the anterior cingulate region)
  - tumor cells in some organ (say the liver), compared with normal cells from the same organ
  - cells from an organism (say yeast) after a treatment (say by heat, or cold, or a drug) compared with cells of the same kind in the untreated state
  - cells from some part of a developing organ or organism at one time, compare with cells of the same kind at a later time, and so on
  - …

Motivation

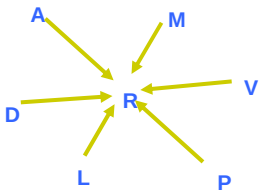
- SCIENTIFIC: To determine which genes are differentially expressed between two sources of mRNA (trt, ctl).
- STATISTICAL: To assign appropriately adjusted *p*-values to thousands of genes, and/or make statements about false discovery rates.
- We will discuss the issues in the context of two experiments, one which fits the aims above, and one which doesn't, but helps make a number of points.

Preliminary: experimental design

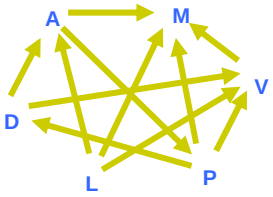
- Comparisons



- Two Ways to Do the Comparisons



Compare all samples to a common reference sample



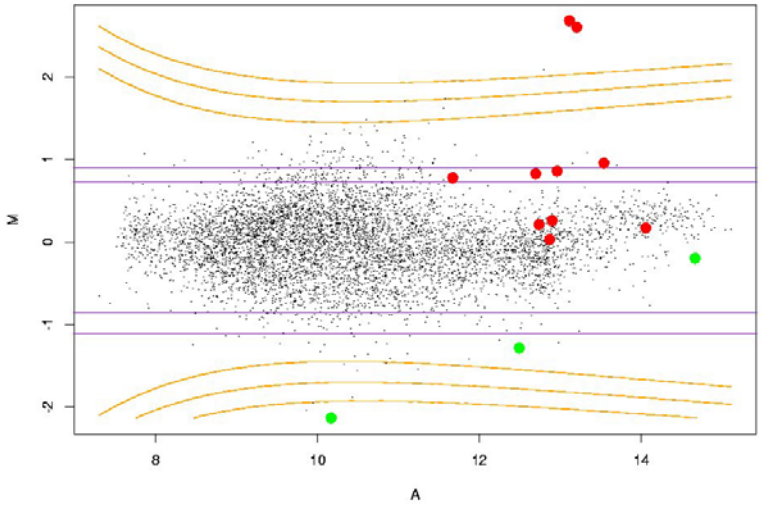
Multiple direct comparisons between different samples

Preliminary: differential analysis with one slide

- The simplest cDNA microarray data analysis problem is identifying differentially expressed genes using **one slide**
  - This is a common enough hope
  - Efforts are frequently successful
  - It is not hard to do by eye
  - The problem is probably beyond formal statistical inference (valid p-values, etc) for the foreseeable future….why?
- In the next two panels, genes found to be up- or down-regulated in an 8 treatment (Srb1 over-expression) versus 8 control comparison are indicated in red and green, respectively, on plots of the data from single hybridizations.
- Also depicted are “confidence lines” determined by different methods and/or different “confidence” levels, which claim to be able to delineate differentially expressed genes using just one hybridization (slide).

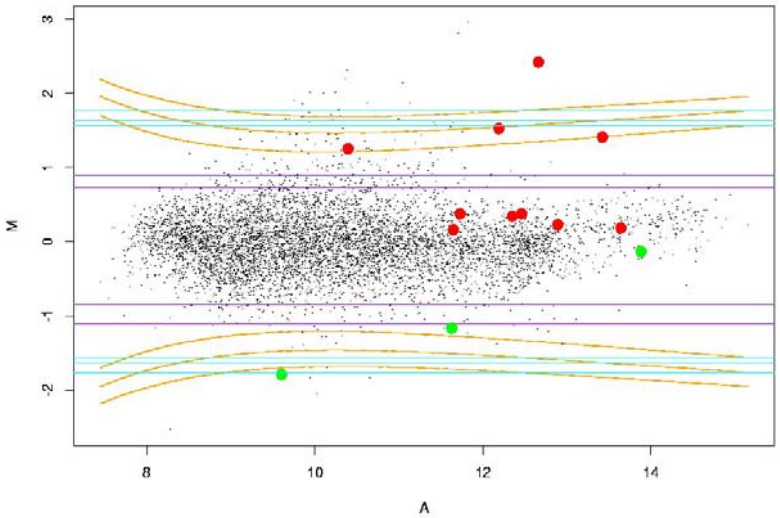
Matt Callow's Srb1 dataset (#5).

Newton's and Chen's single slide method



Matt Callow's Srb1 dataset (#8).

Newton's, Sapir & Churchill's and Chen's single slide method

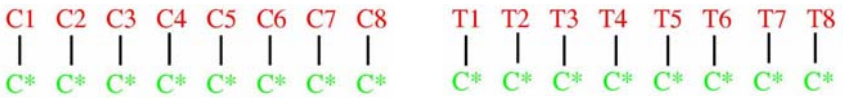


Differential analysis with replicated hybridizations

- The second simplest cDNA microarray data analysis problem is identifying differentially expressed genes using replicated hybridizations
  - There are a number of different aspects:
  - First, between-slide normalization;
  - Then, what should we look at: averages, SDs, t-statistics, other summaries?
  - How should we look at them?
  - Can we make valid probability statements?
- We will discuss the issues in the context of two experiments, one which fits the aims above, and one which doesn't, but helps make a number of points.

Apo AI experiment: (Matt Callow)

- **Goal:** To identify genes with altered expression in the livers of Apo AI knock-out mice (T) compared to inbred C57Bl/6 control mice (C).
  - 8 treatment mice and 8 control mice
  - 16 hybridizations: liver mRNA from each of the 16 mice (Ti , Ci) is 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.



Golub *et al* (1999) experiments

- **Goal.** To identify genes which are differentially expressed in acute lymphoblastic leukemia (ALL) tumors in comparison with acute myeloid leukemia (AML) tumors.
  - 38 tumor 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 × 38 array of expression values.
- Comment: this wasn't really the goal of Golub et al.

Data

- The gene expression data can be summarized as follows

$$X = \begin{bmatrix} \overbrace{x_{1,1} & \cdots & x_{1,n_1}}^{\text{treatment}} & \overbrace{x_{1,n_1+1} & \cdots & x_{1,n}}^{\text{control}} \\ x_{2,1} & \cdots & x_{2,n_1} & x_{2,n_1+1} & \cdots & x_{2,n} \\ \vdots & & & \vdots & & \\ x_{i,1} & \cdots & x_{i,n_1} & x_{i,n_1+1} & \cdots & x_{i,n} \\ \vdots & & & \vdots & & \\ x_{m,1} & \cdots & x_{m,n_1} & x_{m,n_1+1} & \cdots & x_{m,n} \end{bmatrix}$$

- Here $x_{i,j}$ is the (relative) expression value of gene i in sample j. The first n_1 columns are from the treatment (T); the remaining $n_2 = n - n_1$ columns are from the control (C).

Test strategy

- Which genes have changed? When permutation testing possible.

1. For each gene and each hybridization (8 ko + 8 ctl), use $M=\log_2(R/G)$.
2. For each gene form the *t*-statistic:

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

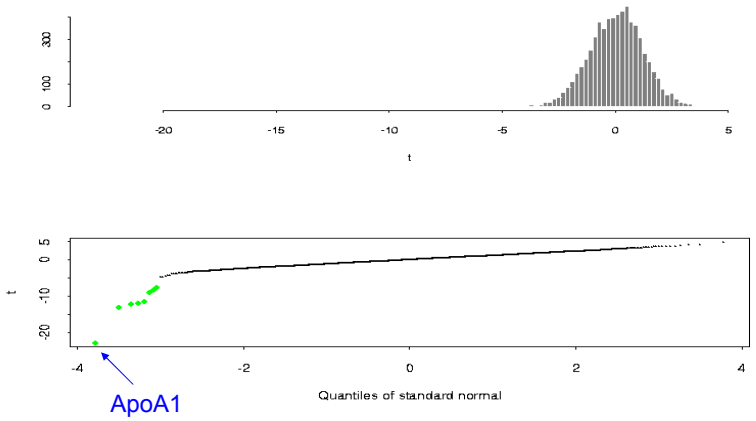
3. Form a histogram of 6,000 *t*-values.
4. Do a normal qq-plot; look for values "off the line".
5. Compute the raw p-values for each gene by permutation procedures or from distribution models.
6. Adjust for multiple testing.

Univariate hypothesis testing

- Initially, focus on one gene only.
- We wish to test the null hypothesis *H* that the gene is not differentially expressed.
- In order to do so, we use a two sample *t*-statistic:

$$t = \frac{averof\, n_1 trtx - averof\, n_2 ctlx}{\sqrt{[\frac{1}{n_1}(SDof\, n_1 trtx)^2 + \frac{1}{n_1}(SDof\, n_1 ctlx)^2]}}$$

Histogram & normal *qq*-plot of *t*-statistics



What is a normal qq-plot?

We have a random sample, say t_i, $i=1, \ldots, n$, which we believe might come from a normal distribution. If it did, then for suitable μ and σ, $\Phi((t_i-\mu)/\sigma)$, $i=1,\ldots n$ would be uniformly distributed on [0,1](why?), where Φ is the standard normal c.d.f.. Denoting the order statistics of the t-sample by $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$ we can then see that $\Phi((t_{(i)}-\mu)/\sigma)$ should be approximately i/n (why?). With this in mind, we'd expect $t_{(i)}$ to be about $\sigma\Phi^{-1}(i/n) + \mu$ (why?).

Thus if we plot $t_{(i)}$ against $\Phi^{-1}((i+1/2/(n+1))$, we might expect to see a straight line of slope about σ with intercept about μ. (The $1/2$ and 1 in numerator and denominator of the i/n are to avoid problems at the extremes.)
This is our normal quantile-quantile plot, the i/n being a quantile of the uniform, and the Φ^{-1} being that of the normal.

Why do a normal q-q plot?

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

It is natural to rank them, but how extreme is extreme? Since the sample sizes here are not too small (two samples of 8 each gives 16 terms in the difference of the means), approximate normality is not an unreasonable expectation for the null marginal distribution.

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 more extreme random variation, which is a big problem in this context. See later in the class for a discussion of these issues.

gene index	t statistic
2139	-22
4117	-13
5330	-12
1731	-11
538	-11
1489	-9.1
2526	-8.3
4916	-7.7
941	-4.7
2000	+3.1
5867	-4.2
4608	+4.8
948	-4.7
5577	-4.5

Gene annotation

Apo AI

EST, weakly sim. to STEROL DESATURASE

CATECHOL O-METHYLTRANSFERASE

Apo CIII

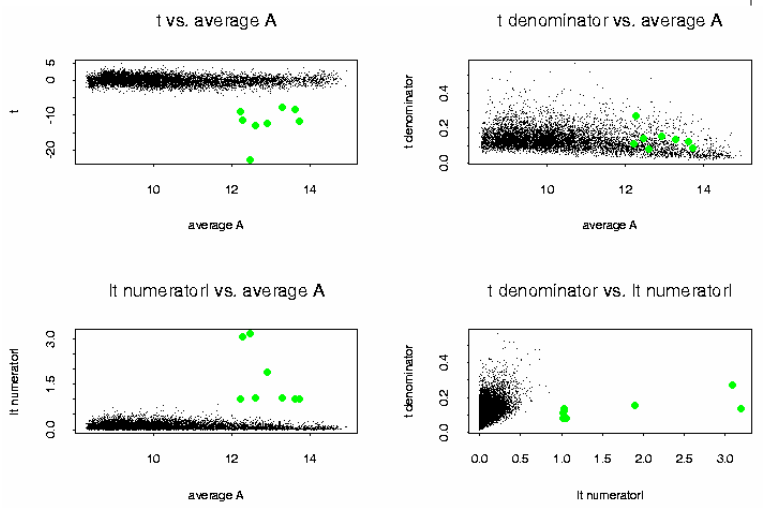
EST, highly sim. to Apo AI

EST

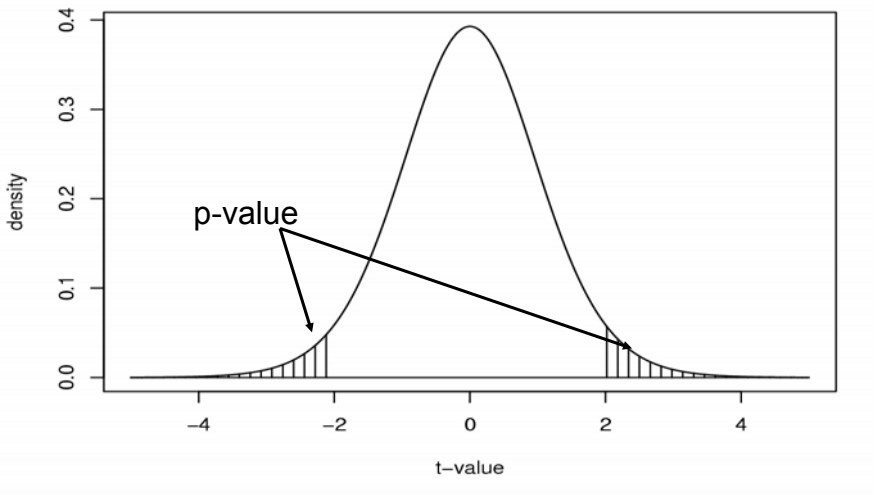
Highly sim. to Apo CIII precursor

similar to yeast sterol desaturase

Useful plots of *t*-statistics



The *p*-values for two sample *t*-stat



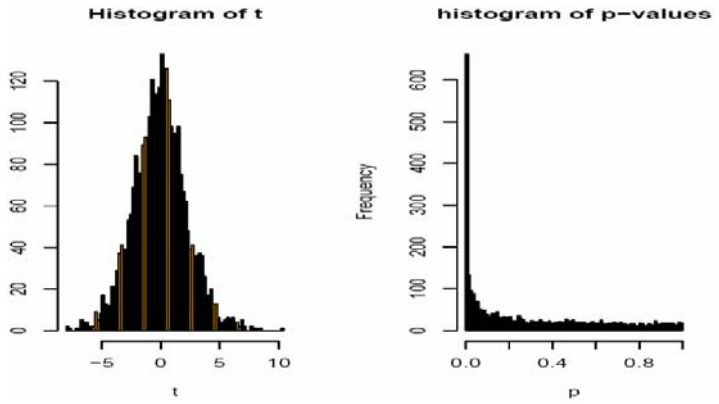
p-values

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

- Although the previous test statistic is denoted by *t*, it would be unwise to assume that its null distribution is that of *Student's t*. We have another way to assign *p*-values which is more or less valid: using permutations.

Example

- Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



t-test: 1045 genes with p < 0.05.

Computing *p*-values by permutations

- We focus on one gene only. For the *b*th iteration, $b = 1, \cdots, B$;
- Permute the n data points for the gene (x). The first n_1 are referred to as "treatments", the second n_2 as "controls".
- For each gene, calculate the corresponding two sample t-statistic, t_b.
- After all the B permutations are done;
- Put $p = \#\{b: |t_b| \geq |t|\}/B$ (plower if we use >).

With **all** permutations in the Apo AI data, $B = n!/n_1!\,n_2! = 12{,}870$;

for the leukemia data, $B = 1.2 \times 10^9$.

Many tests: a simulation study

- **Simulation of this process for 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 now present the 10 smallest raw (unadjusted) permutation *p*-values.**

Unadjusted *p*-values

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

Clearly we can't just use standard p-value thresholds of 05 or .01.

Discussion

- What assumptions on the null distributions of the gene expression values $x_i = (x_{i,1}, x_{i,2}, \ldots x_{i,n})$ are necessary or sufficient for the permutation-based *p*-values just described to be *valid*? And, are they applicable in our examples?
  - First, *p*-values are *valid* if their distribution is *uniform(0,1)* under the null hypothesis.
  - Secondly, if the null distribution of x_i is exchangeable, i.e. invariant under permutations of *1,…,n,* then, we could reasonably hope (and actually prove) that the distribution of the permutation-based *p*-values is indeed uniform on *1,…,n.*
  - We also noted that having the joint distribution i.i.d. would be sufficient, as this implied exchangeability.
- Considered the ApoAI experiment.
  - Because the 16 log-ratios for each gene involved a term from the pooled control mRNA, called *C** above, it seems clear that an i.i.d. assumption is unreasonable.
  - Had the experiment been carried out by using pooled control mRNA from mice other than the controls in the experiment, an exchangeability assumption under the null hypothesis would have been quite reasonable.
  - Unfortunately, *C** did come from the same mice as the C_i, so exchangeability is violated, and the assumption is at best an approximation.

Multiple testing: the problem

Multiplicity problem: thousands of hypotheses are tested simultaneously.

- Increased chance of false positives.
- E.g. suppose you have 10,000 genes on a chip and not a single one is differentially expressed. You would expect 10000x0.01 = 100 of them to have a p-value < 0.01.
- Individual p–values of e.g. 0.01 no longer correspond to significant findings.

Need to **adjust for multiple testing** when assessing the statistical significance of findings.

Multiple testing: Counting errors

Assume we are testing $H^1, H^2, \cdots, H^m$.

m_0 = # of true hypotheses R = # of rejected hypotheses

	# true null hypo.	# false null hypo.	
# accepted	U	T	$m - R$
# rejected	V	S	R
	m_0	$m-m_0$	

V = # Type I errors [false positives]

T = # Type II errors [false negatives]

Type I error rates

- **Per comparison error rate** (PCER): the expected value of the number of Type I errors over the number of hypotheses,

  $$PCER = E(V)/m.$$

- **Per-family error rate** (PFER): the expected number of Type I errors,

  $$PFER = E(V).$$

- **Family-wise error rate**: the probability of at least one type I error

  $$FEWR = pr(V \geq 1)$$

- **False discovery rate** (FDR) is the expected proportion of Type I errors among the rejected hypotheses

  $$FDR = E(V/R;\ R>0) = E(V/R \mid R>0)pr(R>0).$$

- **Positive false discovery** rate (pFDR): the rate that discoveries are false

  $$pFDR = E(V/R \mid R>0).$$

Multiple testing: Controlling a type I error rate

- Aim:

  For a given type I error rate, use a procedure to select a set of "significant" genes that guarantees a type I error rate $\leq \alpha$.

Multiple testing
Family-wise error rates

- Definition:

  FWER = Pr(# of false discoveries >0)

  = Pr(V>0)

  Bonferroni (1936)
  Tukey (1949)
  Westfall and Young (1993) discussed resampling
  ……

- FWER and microarrays
  - maxT step-down procedure
    - Dudoit et al (2002)
    - Westfall et al (2001)
  - minP step-down procedure
    - Ge et al (2003), a novel fast algorithm

Multiple testing
False discovery rates

- Definition:

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

  - Q is set to be 0 when R=0
  - FDR = expectation of Q = E(V/R; R>0)
    - Seeger (1968)
    - Benjamini and Hochberg (1995)
- Caution with FDR
  - Cheating:
    - Adding known diff. expressed genes reduces FDR
  - Interpreting:
    - FDR applies to a set of genes in a global sense, not to individual gene

Types of control of Type I error

- **strong control:** control of the Type I error whatever the true and false null hypotheses. For FWER, strong control means controlling

$$\max_{M_0 \subset H_0^C} pr(V \geq 1 | M_0)$$

  where M_0 = the set of true hypotheses (note $|M_0| = m_0$);

- **exact control**: under M_0, even though this is usually unknown.
- **weak control:** control of the Type I error only under the ***complete null hypothesis*** $H_0^C = \cap_i H_i$, For FWER, this is control of pr(V ≥ 1 | H_0^C).

Adjustments to *p*-values

- **For strong control of the FWER at some level α, there are procedures which will take *m* unadjusted *p*-values and modify them separately, so-called *single step* procedures, the *Bonferroni* adjustment or correction being the simplest and most well known. Another is due to Sidák.**
- **Other, more powerful procedures, adjust sequentially, from the smallest to the largest, or vice versa. These are the *step-up* and *step-down* methods, and we'll meet a number of these, usually variations on single-step procedures.**
- **In all cases, we'll denote adjusted *p*-values by π, usually with subscripts, and let the context define what type of adjustment has been made. Unadjusted *p*-values are denoted by *p***

p-value adjustments: single-step

- Suppose we conduct a hypothesis test for each gene g = 1, . . . ,m, producing
  - an observed test statistic: T_i
  - an unadjusted p–value: p_i.
- Define adjusted *p*-values π, such that the FWER is controlled at level α where H_i is rejected when $\pi_i \leq \alpha$.

  Bonferroni: $\pi_i = min\ (mp_i, 1)$
- Bonferroni always gives strong control.

Proof for Bonferroni (single-step adjustment)

pr (reject at least one H_i at level α | H_0^C)

= pr (at least one $\Pi_i \leq \alpha$ | H_0^C)

$\leq \sum_1^m$ pr ($\Pi_i \leq \alpha$ | H_0^C) , by Boole's inequality

= $\sum_1^m$ pr ($P_i \leq \alpha/m$ | H_0^C), by definiton of Π_i

= $m \times \alpha/m$, assuming $P_i \sim U[0,1]$)

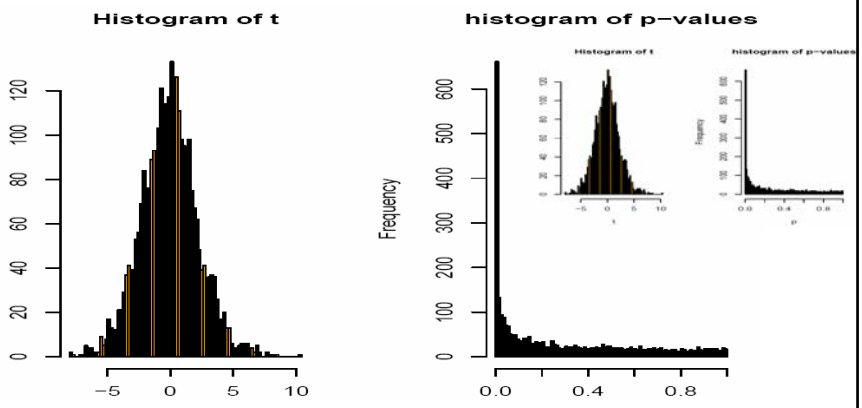
= α.

Notes:

1. We are testing *m* genes, H_0^C is the complete null hypothesis, that no gene is differentially expressed.
2. P_i is the unadjusted *p*-value for gene i , while Π_i here is the Bonferroni adjusted *p*-value.
3. We use lower case letters for observed *p*-values, and upper case for the corresponding random variables.

Example

- Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



98 genes with Bonferroni-adjusted $\pi_i < 0.05 \Leftrightarrow p_i < 0.000016$ (t-test)

More is not always better

- Suppose you produce a small array with 500 genes you are particularly interested in.
- If a gene on this array has an unadjusted *p*-value of 0.0001, the Bonferroni-adjusted *p*-value is still 0.05.
- If instead you use a genome-wide array with, say, 50,000 genes, this gene would be much harder to detect, because roughly 5 genes can be expected to have such a low p-value by chance.

p-value adjustments: single-step

- Suppose we conduct a hypothesis test for each gene g = 1, . . . ,m, producing
  - an observed test statistic: T_i
  - an unadjusted p–value: p_i.
- Define adjusted *p*-values π, such that the FWER is controlled at level α where H_i is rejected when $\pi_i \leq \alpha$.

  Sidák: $\pi_i = 1 - (1 - p_i)m$
- Sidák is less conservative than Bonferroni. When the genes are independent, it gives strong control exactly (FWER= α), proof later. It controls FWER in many other cases, but is still conservative.

Proof for Sidák's method (single-step adjustment)

pr(reject at least one $H_i \mid H_0^C$)
= pr(at least one $\Pi_i \leq \alpha \mid H_0^C$)
= 1 - pr(all $\Pi_i > \alpha \mid H_0^C$)
= $1-\prod_{i=1}^{m}$ pr($\Pi_i > \alpha \mid H_0^C$) assuming independence

Here Π_i is the Sidák adjusted *p*-value, and so $\Pi_i > \alpha$ if and only if $P_i > 1-(1- \alpha)^{1/m}$ (check), giving

$1-\prod_{i=1}^{m}$ pr($\Pi_i > \alpha \mid H_0^C$)
= $1-\prod_{i=1}^{m}$ pr($P_i > 1-(1- \alpha)^{1/m} \mid H_0^C$)
= 1- { $(1- \alpha)^{1/m}$ }^{m} since all $P_i \sim U[0,1]$,
= $\tilde{\alpha}$

Single-step adjustments (ctd)

FWER: Improvements to Bonferroni

- The *minP method of Westfall and Young:*

$$\Pi_i = pr(\min_{1 \le l \le m} P_l \le p_i \mid H)$$

- Based on the joint distribution of the *p*-values $\{P_l\}$. This is the most powerful of the three single-step adjustments.
- If $P_i \sim U[0,1]$, it gives a FWER exactly = α (see next page).
- It always confers weak control, and gives strong control under subset pivotality (definition next but one slide).

Proof for (single-step) minP adjustment

Given level α, let c_α be such that

$$pr(\min_{1 \le i \le m} P_i \le c_\alpha \mid H_0^C) = \alpha .$$

Note that $\{\Pi_i \le \alpha\} \equiv \{P_i \le c_\alpha\}$ for any i.

pr(*reject at least one* H_i *at level* $\alpha \mid H_0^C$)

= pr (*at least one* $\Pi_i \le \alpha \mid H_0^C$)

= pr ($\min_{1 \le i \le m} \Pi_i \le \alpha \mid H_0^C$)

= pr ($\min_{1 \le i \le m} P_i \le \alpha \mid H_0^C$)

= α.

Strong control and subset pivotality

- The above proofs are under H_0^C, which is what we term weak control
- In order to get strong control, we need the condition of *subset pivotality.*

- The distribution of the unadjusted *p*-values (P_1, P_2, …P_m) is said to have the *subset pivotality* property if for all subsets $L \subseteq \{1,\ldots,m\}$ the distribution of the subvector $\{P_i: i \in L\}$ is identical under the restrictions $\cap\{H_i: i \in L\}$ and H_0^C .

  - Using the property, we can prove that for each adjustment under their conditions, we have

    pr (reject at least one H_i at level α, $i \in M_0 \mid H_{M0}\}$

    = pr (reject at least one H_i at level α, $i \in M_0 \mid H_0^C\}$

    $\leq$ pr (reject at least one H_i at level α, for all $i \mid H_0^C\}$

    $\leq \alpha$.
  - Therefore, we have proved strong control for the previous three adjustments, assuming subset pivotality.

Permutation-based single-step minP adjustment of *p*-values

- For the *b*th iteration, $b = 1, \cdots, B$;
- Permute the *n* columns of the data matrix *X*, obtaining a matrix X_b. The first n_1 columns are referred to as "treatments", the second n_2 columns as "controls".
- For each gene, calculate the corresponding unadjusted *p*-values, $p_{i,b}$, $i= 1,2, \cdots m$, (e.g. by further permutations) based on the permuted matrix X_b.
- After all the *B* permutations are done.
- Compute the adjusted *p*-values $\pi_i = \#\{b: min_l\, p_{l,b} \leq p_i\}/B$.

Example

- Suppose p_{min} = 0.0003 (the minimal unadjusted $_p$-value).
- Among the randomized data sets (permuted sample labels), count how often the minimal *p*-value is smaller than 0.0003. If this appears e.g. in 4% of all cases, π_{min} = 0.04.

The computing challenge: iterated permutations

- The procedure is quite computationally intensive if *B* is very large (typically at least 10,000) **and** we estimate all unadjusted *p*-values by further permutations.
- Typical numbers:

  To compute one unadjusted *p*-value B = 10,000
  # unadjusted *p*-values needed B = 10,000
  # of genes m = 6,000. In general run time is $O(mB^2)$.
- How to avoid computational difficulty?

Single-step minP adjustment

- *maxT* method: (Chapter 4 of Westfall and Young)

$$\pi_i = \Pr\left(\max_{1 \le l \le m} |T_l| \ge |t_i| \mid H_0^C \right)$$

- needs B = 10,000 permutations only.
- However, if the distributions of the test statistics are not identical, it will give more weight to genes with heavy tailed distributions (which tend to have larger *t*-values)
- There is a fast algorithm which does the minP adjustment in O(mBlogB+mlogm) time.

Proof for the single-step maxT adjustment

Given level α, let c_α such that $\Pr(\max_{1 \le i \le m} |T_i| \le c_\alpha \mid H_0^C) = \alpha$.
Note the $\{ P_i \le \alpha \} \equiv \{ |T_i| \le c_\alpha \}$ for any i. Then we have (cf. min P)

$\Pr$(reject at least one H_i at level $\alpha \mid H_0^C$)
$= \Pr$ (at least one $P_i \le \alpha \mid H_0^C$)
$= \Pr (\min_{1 \le i \le m} P_i \le \alpha \mid H_0^C)$
$= \Pr (\max_{1 \le i \le m} |T_i| \le c_a \mid H_0^C)$
$= \alpha$.

To simplify the notation we assumed a two sided test by using the statistic T_i .We also assume $P_i \sim U[0,1]$.

More powerful methods: step-down adjustments

The idea: S Holm's modification of Bonferroni.

Also applies to Sid*ák*, maxT, and minP.

S Holm's modification of Bonferroni

- Order the unadjusted *p*-values such that $p_{r1} \leq p_{r2} \leq \cdots \leq p_{rm}$.

  The indices $r_1, r_2, r_3, ..$ are fixed for given data.

- For control of the FWER at level, the step-down Holm adjusted p-values are

$$\pi_{rj} = max_{k \in \{1,\ldots,j\}} \{min((m-k+1)p_{rk}, 1)\}.$$

  - The point here is that we don't multiply every p_{rk} by the same factor m, but only the smallest. The others are multiplied by successively smaller factors: m-1, m-2, ..,. down to multiplying p_{rm} by 1.
  - By taking successive maxima of the first terms in the brackets, we can get monotonicity of these adjusted *p*-values.
  - Holm's adjusted *p*-values deliver strong control.

Step-down adjustment of minP

- Order the unadjusted *p*-values such that $p_{r1} \le p_{r2} \le \cdots \le p_{rm}$.
- Step-down adjustment: it has a complicated formula, see below, but in effect is

  1.Compare min{$P_{r1}, \cdots, P_{rm}$} with p_{r1} ;

  2.Compare min{$P_{r2}, \cdots, P_{rm}$} with p_{r2} ;

  3Compare min{$P_{r3} \cdots, P_{rm}$} with p_{ri3} …….

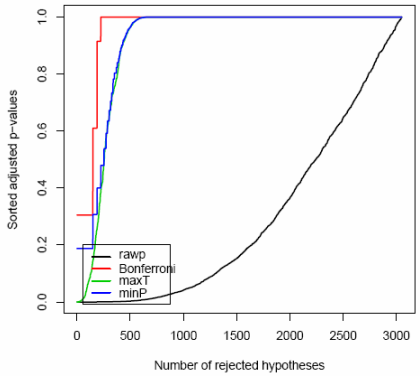
  m.Compare P_{rm} with p_{rm} .

- Enforce **monotonicity** on the adjusted p_{ri} . The formula is

  $\pi_{rj} = max_{k \in \{1,,\ldots,j\}} \{pr(min_{l \in \{rk,\ldots rm\}} P_l \le p_{rk} \mid H_0^C)\}$.

FWER: Comparison of different methods

- Golub data, 27 ALL vs. 11 AML samples, 3,051 genes.



The FWER is a conservative criterion: many interesting genes may be missed.

False Discovery Rate

- FDR: = E(V/R)
  = E(F/S|S>0)P(S>0)

In case R=0, define F/R=0 if R=0.

- Alternatively, define pFDR=E(V/R|R>0).
  - When m is large, P(S>=0) is approx. 1 and FDR is approx. equal to pFDR.
- FDR is a measure of the overall accuracy of a set of significant features.

False discovery rate
(Benjamini and Hochberg 1995)

Steps:

- Select desired limit α on E(FDR)
- Rank the p-values $p_{r1} \leq p_{r2} \leq \ldots \leq p_{rm}$.
- The adjusted *p*-values are to control FDR when P_i are independently distributed are given by the step-up formula:

$$\pi_{ri} = \min_{k \in \{i \ldots m\}} \{ \min (m p_{rk}/k, 1) \}.$$

- We use this as follows: reject $H_{r1}, H_{r2}, \ldots, H_{rk^*}$ where k^* is the largest k such that $p_{rk} \leq (k/m)\alpha$. This keeps the FDR $\leq \alpha$ under independence
  - Thus the FDR Adjusted p-value = lowest level of FDR for which the hypothesis is first included in the set of rejected hypothesis
  - Compare the above with Holm's adjustment to control FWE, the step-down version of Bonferroni, which is $\pi_i = \max_{k \in \{1,\ldots i\}} \{ \min (k p_{rk}, 1) \}$.

Positive false discovery rate
(Storey, 2001, independent case)

- A new definition of FDR, called positive false discovery rate (pFDR)

$$pFDR = E(V/R \mid R > 0)$$

- The logic behind this is that in practice, at lease one gene should be expected to be differentially expressed.
- The adjusted *p*-value (called *q*-value in Storey's paper) are to control pFDR.

$$\Pi_i = \min_{k \in \{1..i\}} \{(m p_k / k)\, \pi_0\}$$

- Note $\pi_0 = m_0 / m$ can be estimated by the following formula for suitable β

$$\pi_0 = \#\{p_i > \beta\} / \{(1-\beta)\, m\}.$$

Estimation of the FDR

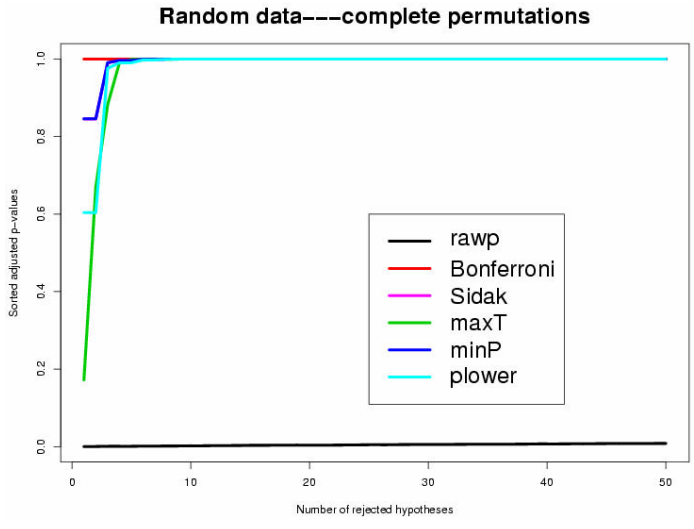
- Idea: Depending on the chosen cutoff-value(s) for the test statistic T_i, estimate the expected proportion of false positives in the resulting gene list through a permutation scheme.
  - Estimate the number m_0 of non-diff. genes: $m_0 = \#\{p_i > \beta\}/(1-\beta)$.
  - Compute the average number of significant genes under permutations of the sample labels.
    - For b = 1, . . . ,B, (randomly) permute the sample labels – this corresponds to the complete null hypothesis. Compute test statistics T_{ib} for each gene.
    - For any threshold t_0 of the test statistic, compute the numbers V_b of genes with $T_{ib} > t_0$ (numbers of false positives).
    - compute the mean of the V_b.
  - Estimate the FDR

$$E(V/R] = \hat{V}\frac{m_0}{m}/R$$

FWER or FDR?

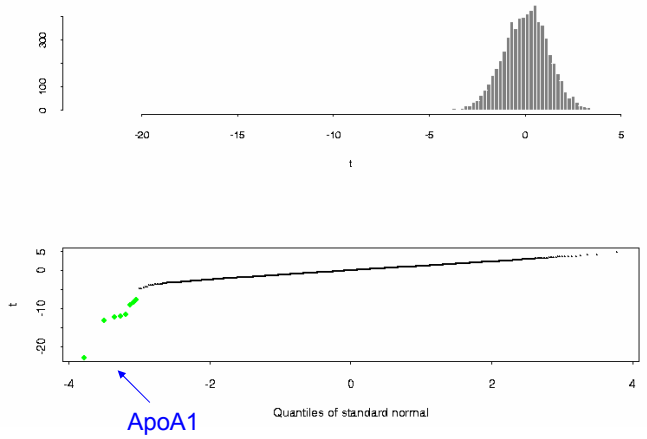
- Chose control of the FWER if high confidence in all selected genes is desired. Loss of power due to large number of tests: many differentially expressed genes may not appear as significant.
- If a certain proportion of false positives is tolerable: Procedures based on FDR are more flexible; the researcher can decide how many genes to select, based on practical considerations.

Results: random data

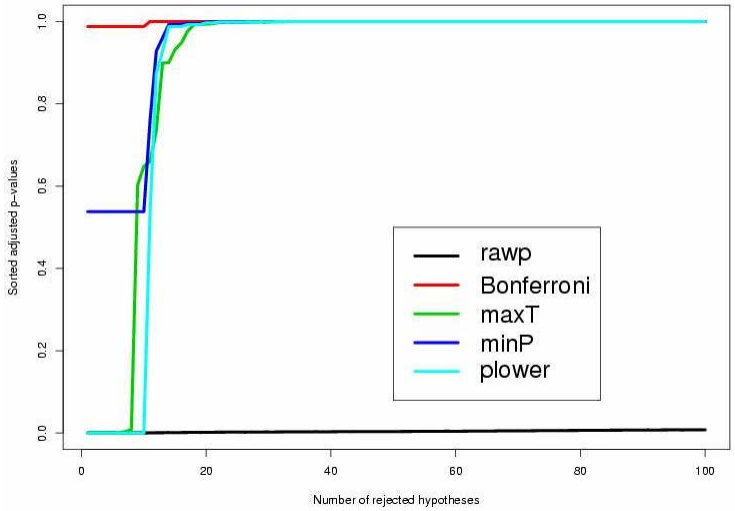


Results: Apo AI data

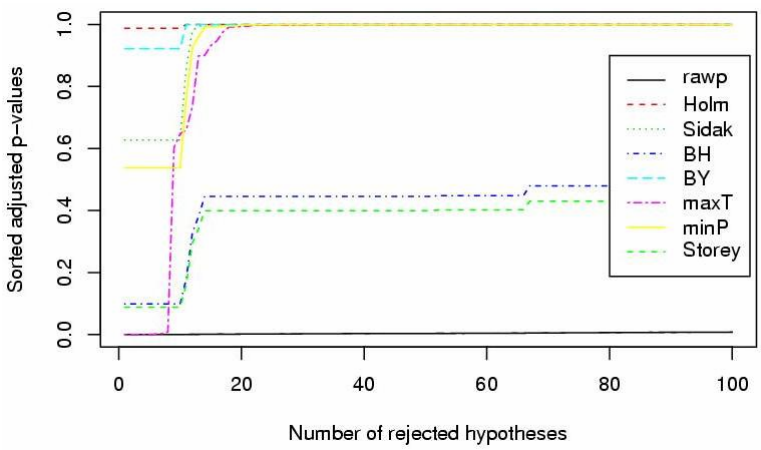
- Histogram & normal q-q plot of t-statistics



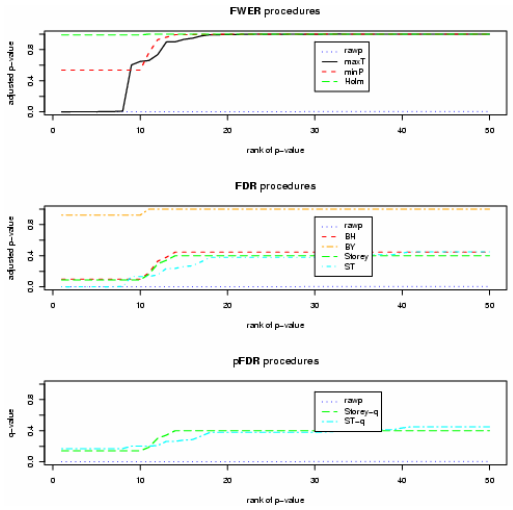
Callow's AI ko data – complete permutation



Callow data with some FDR values included



Comparison



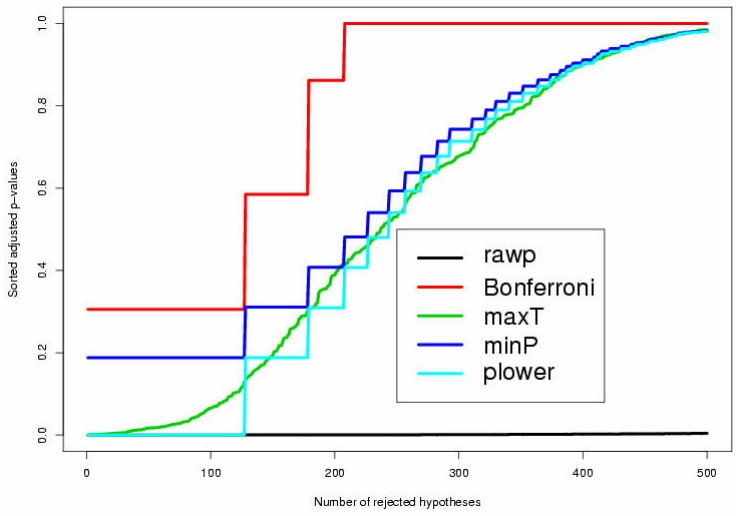
Comparison

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

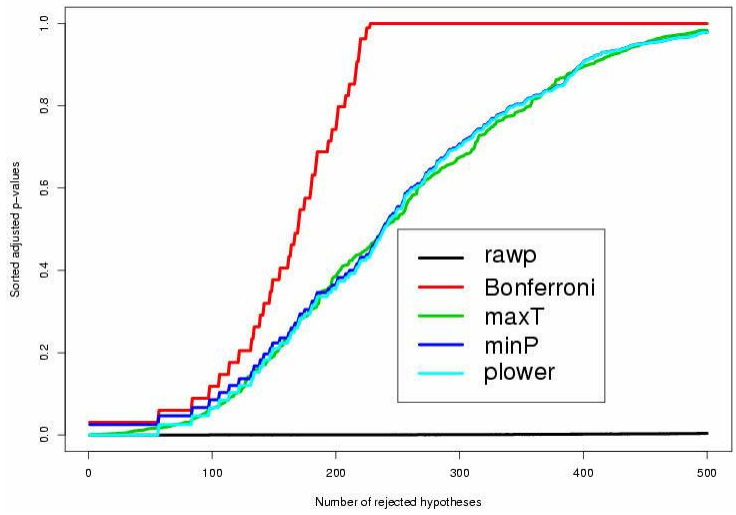
The gene names

Index	Name
2139	Apo AI
4117	EST, weakly sim. to STEROL DESATURASE
5330	CATECHOL O-METHYLTRANSFERASE
1731	Apo CIII
538	EST, highly sim. to Apo AI
1489	EST
2526	Highly sim. to Apo CIII precursor
4916	similar to yeast sterol desaturase

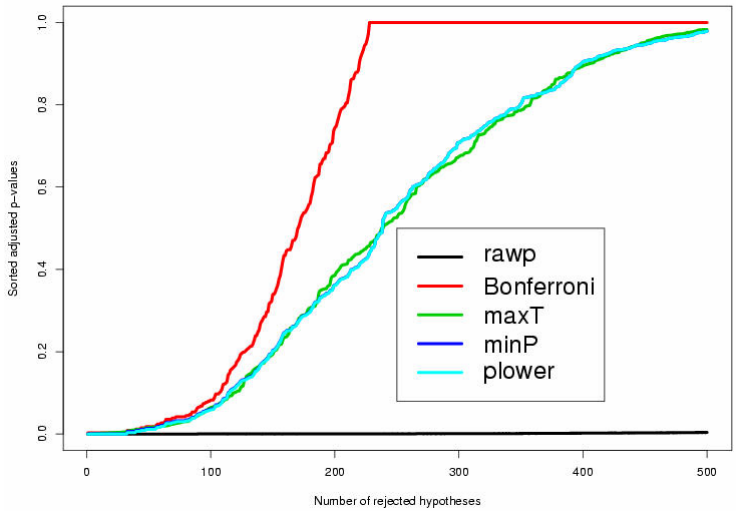
Golub's data --- 10K simulations



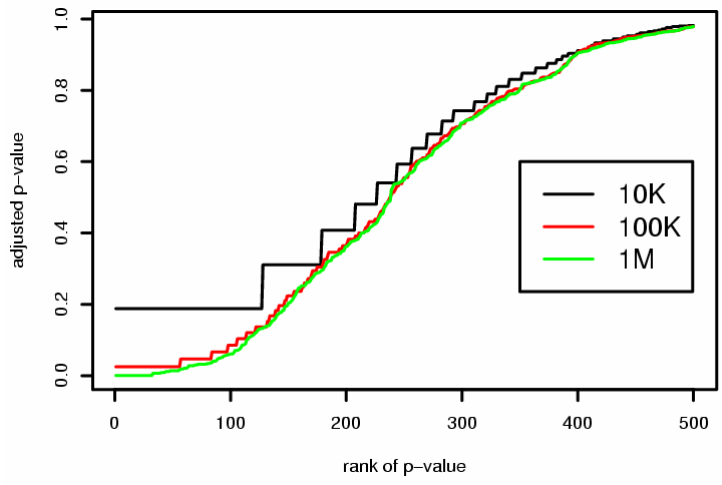
Golub's data --- 100K simulations



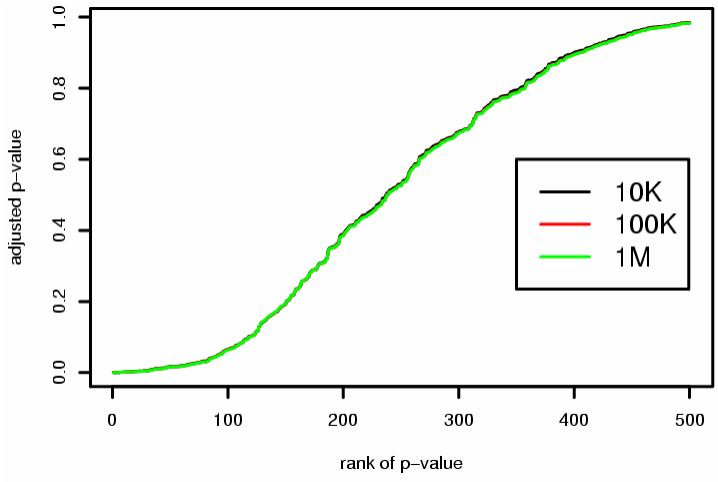
Golub's data --- 1M simulations



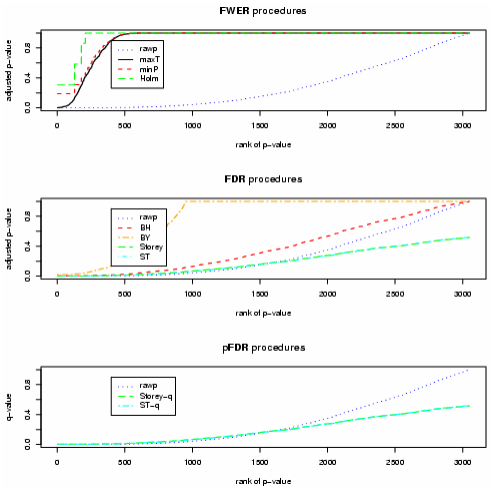
Golub data with minP



Golub data with maxT



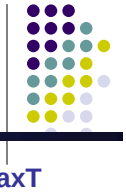
Comparisons



What should one look for in a multiple testing procedure?

- There is a bewildering variety of multiple testing procedures. How can we choose which to use? There is no simple answer here, but each can be judged according to a number of criteria:
  - **Interpretation**: does the procedure answer a relevant question for you?
  - **Type of control**: strong, exact or weak?
  - **Validity**: are the assumptions under which the procedure applies clear and definitely or plausibly true, or are they unclear and most probably not true?
  - **Computability**: are the procedure's calculations straightforward to calculate accurately, or is there possibly numerical or simulation uncertainty, or discreteness?

Discussion



- **The minP adjustment seems more conservative than the maxT adjustment, but is essentially model-free.**
- **With the Callow data, we see that the adjusted minP values are very discrete; it seems that 12,870 permutations are not enough for 6,000 tests.**
- **With the Golub data, we see that the number of permutations matters. Discreteness is a real issue here to, but we do have enough permutations.**
- **The same ideas extend to other statistics: Wilcoxon, paired t, F, blocked F.**
- **Same speed-up works with the bootstrap.**

Selected references

- Westfall, PH and SS Young (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment*, John Wiley & Sons, Inc
- Benjamini, Y & Y Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing *JRSS B* 57: 289-300
- J Storey (2001): 3 papers (some with other authors), www-stat.stanford.edu/~jstorey/
  - The positive false discovery rate: a Bayesian interpretation and the q-value.
  - A direct approach to false discovery rates
  - Estimating false discovery rates under dependence, with applications to microarrays
- Y Ge et al (2001) Resampling-based multiple testing for microarray data analysis, *Test* (to appear), see #633 in http://www.stat.Berkeley.EDU/tech-reports/index.html
- Software
  - C and R code available for different tests: multtest in http://www.bioconductor.org

Acknowledgements

Slides are adapted from Lecture notes of

Terry Speed

