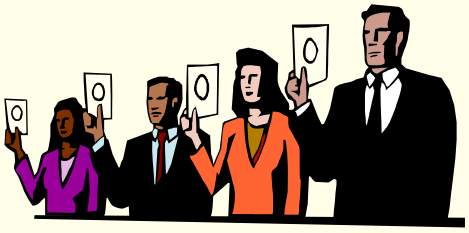
THE PROBLEM OF MULTIPLE TESTING



HYPOTHESIS TESTING

(Statistical Errors)

Significance level

	H_0 is not rejected	H_0 is rejected
H_0 is true	No error (1-α)	Type I error (α)
H_0 is false	Type II error (β)	No error (1-β)

Power

➡ Standard approach:

① Specify an acceptable type I error rate (α)

② Seek tests that minimize the type II error rate (β), i.e., maximize power (1 - β)

THE PROBLEM OF MULTIPLE TESTING

Suppose you carry out 10 hypothesis tests at the 5% level
(assume independent tests)

The probability of declaring a particular test significant under its null hypothesis is 0.05

But the probability of declaring at least 1 of the 10 tests significant is 0.401

$1 - 0.95^{10}$

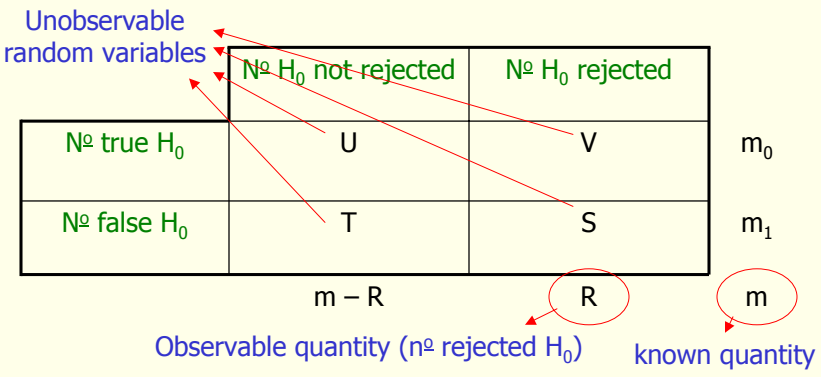
If you perform 20 hypothesis tests, this probability increases to 0.642…

- Typically thousands of genes simultaneously
- Type I error rate (α)

Suppose: Self-self hybridization with g = 1000 genes; and α = 5% (for each test)

- Expected $1000 \times 0.05 = 50$ false positives…

- Adjusting p-values:
  - ① Controlling family-wise type I error rates (Westfall and Young, 1993)
  - ② False discovery rate (Benjamini and Hochberg, 1995; Storey et al., 2002)
- Set-up: Testing m null hypothesis H_j (j = 1,…,m) (m_o true and m_1 false null hypothesis; R: nº H_0 rejected (false positives)



➡ Some definitions (main rates):

- Per-comparison error rate (PCER): $PCER = \frac{E[V]}{m}$
- Family-wise error rate (FWER):

$$FWER = Pr(V \geq 1) = 1 - Pr(V = 0)$$

- False discovery rate (FDR):

$$FDR = E[V/R \mid R > 0] = \underbrace{E[V/R \mid R > 0]}\, Pr(R > 0)$$

Positive FDR (pFDR); Storey (2002)

➡ Strong and weak control of type I error rate:

- Strong control: control type I error rate under any combination of true and false hypothesis (m_0 and m_1)
- Weak control: control type I error rate only when $m_1 = 0$ (H_0^C) (very unlikely scenario in the microarray setting)

➡ Adjusted p-values:

- Single-step procedures: Equivalent adjustments are performed for all hypothesis
- Stepwise procedures: Adjustments based not only on m, but also on outcome of the tests
  - Step-down methods: Order unadjusted p-values and start with the most significant
  - Step-down methods: Order unadjusted p-values and start with the least significant

➡ Strong control of FWER at level α:

① Bonferroni: Rejects any hypothesis H_j with p-value less than or equal to α/m, i.e.:

$$\tilde{p}_j = \min[mp_j, 1]$$

adjusted p-value — unadjusted p-value

② Sidák: Rejects any hypothesis H_j with p-value less than or equal to $1-(1-\alpha)^{1/g}$, i.e.:

$$\tilde{p}_j = \min[1-(1-p_j)^g, 1]$$

- Very similar to Bonferroni adjustment.
- Both are too conservative...

③ Holm step-down approach: $p_{r_1} \le p_{r_2} \le \ldots \le p_{r_m}$

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j}\{\min[(m-k+1)p_{r_k}, 1]\}$$

- Less conservative than Bonferroni.

None of these methods, however, take into account dependence between tests (co-regulated genes)

④ Westfall and Young (1993)

Step-down minP adjusted p-values:

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j}\left\{\Pr\left[\min_{l\in\{r_k,\ldots,r_m\}} P_l \le p_{r_k} \,\middle|\, H_o^C\right]\right\}$$

Step-down maxT adjusted p-values:

$$\tilde{p}_{s_j} = \max_{k=1,\ldots,j}\left\{\Pr\left[\max_{l\in\{s_k,\ldots,s_m\}} |T_l| \ge |t_{s_k}| \,\middle|\, H_o^C\right]\right\}$$

$|t_{s_1}| \ge |t_{s_2}| \ge \ldots \ge |t_{s_m}|$: ordered test statistics

- These procedures guarantee weak control of FWER in all cases, and strong under the additional assumption of subset pivotality (Dudoit et al., 2002).

➡ Strong control of FDR:

① Benjamini and Hochberg (1995):

$$\tilde{p}_{r_j} = \min_{k=i,\ldots,m} \left\{ \min\left(\frac{m p_{r_k}}{k}, 1 \right) \right\}$$

- Strong assumption of independence between tests…
- More general approach multiply p_{r_k} by mlog(m) (reasonable approximation for a particular form of dependence, when m is large)

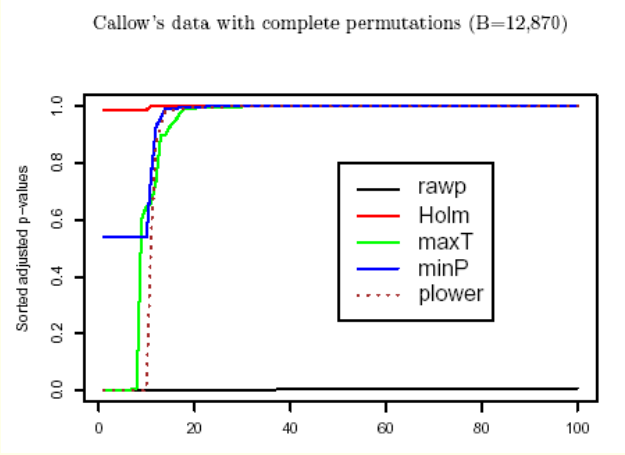
② q-value (Storey, 2001):

- q-value: minimum pFDR that can occur when rejecting a statistic equal to the observed one for a nested set of rejection regions.

➡ Comparison of Multiple Testing Procedures:

Callows's APO AI knock-out experiment

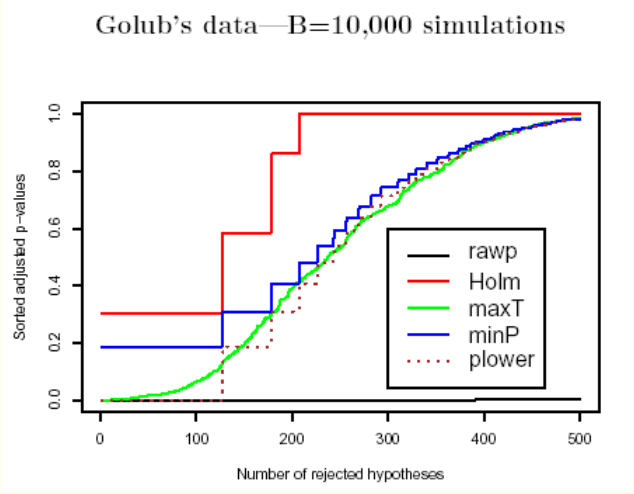
- Two samples of 8 mice each (12870 possible permutations)
- Probes m = 6,356 cDNA



Source: Terry Speed's web page

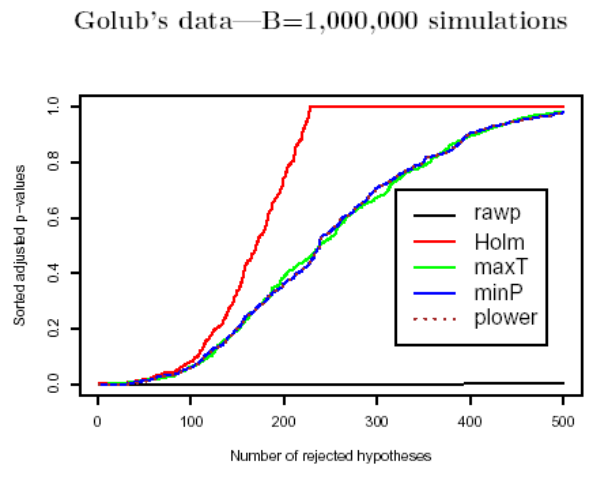
Golub's leukemia study

- Two types of leukemia: ALL (n = 27) and AML (n = 11)
- Oligonucleotide arrays with 6,817 genes
- Final dataset: m = 3051



Source: Terry Speed's web page

Golub's leukemia study



Source: Terry Speed's web page

➡ Final Remarks:

- FWER: too conservative for microarray experiments
- Permutation based adjustments generally not satisfactory because of small sample sizes (discreteness)
- FDR (or pFDR) appear to be most promising alternative

EXAMPLE: PROC MULTTEST

```
DATA clones;
INPUT group $ clone1-clone10;
CARDS;
Control 6.23 6.69 4.45 5.63 6.55 4.98 5.94 6.00 6.05 4.68
Control 6.69 6.32 3.65 6.81 5.98 4.58 7.37 5.45 7.47 6.67
Control 6.12 5.87 5.09 4.98 3.85 4.87 4.79 6.54 5.32 6.26
Control 4.60 5.18 3.92 7.62 5.67 5.67 6.44 6.48 5.35 5.96
Control 6.34 5.61 6.48 5.39 7.71 5.28 5.32 7.56 7.20 5.74
trt     6.29 5.36 6.38 3.96 7.76 7.73 5.30 4.74 7.40 5.68
trt     7.03 6.24 4.84 4.81 4.26 9.05 7.62 5.23 5.31 6.60
trt     6.25 6.62 6.72 5.50 6.47 6.15 5.04 4.27 5.97 4.84
trt     5.70 4.34 4.80 6.54 7.90 7.17 5.16 5.37 5.68 7.09
trt     5.89 5.92 6.44 8.15 6.38 6.88 8.31 4.96 6.16 7.51
trt     6.52 3.59 8.25 5.54 4.58 7.95 4.82 6.45 5.99 7.23

;

PROC ANOVA DATA=clones;
CLASS group;
MODEL clone1-clone10 = group;
RUN;

PROC MULTTEST DATA=clones BONFERRONI SIDAK FDR PERMUTATION;
CLASS group;
CONTRAST 'Diff. Expression' 1 -1;
TEST MEAN(clone1-clone10);
RUN;
```

RESULTS

Continuous Variable Tabulations

Variable	group	NumObs	Mean	Standard Deviation
clone1	Control	5	5.9960	0.8092
clone1	trt	6	6.2800	0.4711
clone2	Control	5	5.9340	0.5912
clone2	trt	6	5.3450	1.1703
clone3	Control	5	4.7180	1.1283
clone3	trt	6	6.2383	1.2933
clone4	Control	5	6.0860	1.0946
clone4	trt	6	5.7500	1.4545
clone5	Control	5	5.9520	1.4095
clone5	trt	6	6.2250	1.5371
clone6	Control	5	5.0760	0.4160
clone6	trt	6	7.4883	0.9973
clone7	Control	5	5.9720	0.9996
clone7	trt	6	6.0417	1.5139
clone8	Control	5	6.4060	0.7801
clone8	trt	6	5.1700	0.7377
clone9	Control	5	6.2780	1.0127
clone9	trt	6	6.0850	0.7102
clone10	Control	5	5.8620	0.7471
clone10	trt	6	6.4917	1.0333

p-Values

Variable	Contrast	Raw	Bonferroni	Sidak	False Discovery Rate	Permutation
clone1	Diff. Expression	0.4847	1.0000	0.9987	0.8079	1.0000
clone2	Diff. Expression	0.3361	1.0000	0.9834	0.6721	0.9818
clone3	Diff. Expression	0.0702	0.7023	0.5172	0.2341	0.5399
clone4	Diff. Expression	0.6811	1.0000	1.0000	0.8532	1.0000
clone5	Diff. Expression	0.7678	1.0000	1.0000	0.8532	1.0000
clone6	Diff. Expression	0.0007	0.0072	0.0072	0.0072	0.0113
clone7	Diff. Expression	0.9320	1.0000	1.0000	0.9320	1.0000
clone8	Diff. Expression	0.0245	0.2451	0.2198	0.1226	0.2402
clone9	Diff. Expression	0.7188	1.0000	1.0000	0.8532	1.0000
clone10	Diff. Expression	0.2862	1.0000	0.9657	0.6721	0.9682

REFERENCES

Westfall P. H. and S. S. Young (1993) Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment. John Wiley & Sons, New York.

Benjamini, V. and V. Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B. 57, 289-300.

Storey, J. D., J. E. Taylor and D. Siegmund (2002) A unified estimation approach to false discovery rates. Technical Report 623, Department of Statistics, University of California, Berkeley.

Dudoit, S., J. P. Shaffer and J. C. Boldrick (2002) Multiple hypothesis testing in microarray experiments. Technical Report 110, Division of Biostatistics, University of California, Berkeley.

Storey, J. D. (2001) The positive false discovery rate: a Bayesian interpretation and the q-value. Technical Report 12, Department of Statistics, Stanford University.

Efron, B., Storey, J. D., Tibshirani, R. (2001) Microarrays, empirical Bayes methods, and false discovery rates. Stanford Technical Report # 216.

Tusher, V. G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarray applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* USA 98: 5116-5121.