Multiple testing in DNA microarray experiments

Sandrine Dudoit

Bioconductor short course
Summer 2002

Differential gene expression

- Identify genes whose expression levels are **associated** with a response or covariate of interest
  - clinical outcome such as survival, response to treatment, tumor class;
  - covariate such as treatment, dose, time.
- **Estimation**: estimate effects of interest (e.g. difference in means, slope, interaction) and **variability** of these estimates.
- **Testing**: assess the statistical **significance** of the observed associations.

Hypothesis testing

- Test for each gene the **null hypothesis** of no differential expression, e.g. using t– or F–statistic.

H_g : the expression level of gene g
is not associated with the covariate or response

Two types of errors can be committed

- **Type I error** or **false positive**
  say that a gene is differentially expressed when it is not, i.e., reject a *true null* hypothesis.
- **Type II error** or **false negative**
  fail to identify a truly differentially expressed gene, i.e., fail to reject a *false null* hypothesis.

Multiple hypothesis testing

- Large multiplicity problem: thousands of hypotheses are tested simultaneously!
  - Increased chance of false positives.
  - E.g. chance of at least one p–value $< \alpha$ for G independent tests is $1-(1-\alpha)^G$ and converges to one as G increases. For $G = 1,000$ and $\alpha = 0.01$, this chance is 0.9999568!
  - Individual p–values of 0.01 no longer correspond to significant findings.
- Need to **adjust for multiple testing** when assessing the statistical significance of the observed associations.

Multiple hypothesis testing

- Define an appropriate **Type I error** or **false positive rate**.
- Develop multiple testing procedures that
  - provide **strong control** of this error rate,
  - are **powerful** (few false negatives),
  - take into account the **joint distribution** of the test statistics.
- Report **adjusted** ***p*–values** for each gene which reflect the **overall** Type I error rate for the experiment.
- **Resampling** methods are useful tools to deal with the unknown joint distribution of the test statistics.

Multiple hypothesis testing

	# non–rejected hypotheses	# rejected hypotheses	
# true nullhypotheses	U	**V** **Type I error**	G_0
# false null hypotheses	**T** **Type II error**	S	G_1
	$G - R$	R	G

From Benjamini & Hochberg (1995).

Type I error rates

1. **Per–family error rate (PFER)**. The PFER is defined as the expected number of Type I errors, i.e.,

$$PFER = E(V).$$

2. **Per–comparison error rate (PCER)**. The PCER is defined as the expected value of (number of Type I errors/number of hypotheses), i.e.,

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

Type I error rates

3. **Family–wise error rate (FWER)**. The FWER is defined as the probability of at least one Type I error, i.e.,

$$FWER = p(V > 0).$$

4. **False discovery rate (FDR)**. The FDR of Benjamini & Hochberg (1995) is the expected proportion of Type I errors among the rejected hypotheses, i.e.,

$$FDR = E(Q),$$

where by definition

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

Strong vs. weak control

N.B. All probabilities are **conditional** on which hypotheses are true.

Strong control refers to control of the Type I error rate under *any combination* of true and false hypotheses, i.e., under $\cap_{g \in K} \mathrm{H}_g$ for any $K \subseteq \{1, \dots, G\}$.

Weak control refers to control of the Type I error rate only when *all* the null hypotheses are true, i.e., under the **complete null hypothesis** $\mathrm{H}_0^C = \cap_{g=1}^{G} \mathrm{H}_g$ with $G_0 = G$.

In general, weak control without any other safeguards is unsatisfactory.

Comparison of Type I error rates

In general, for a given multiple testing procedure,

$$PCER \leq FWER \leq PFER,$$

and

$$FDR \leq FWER,$$

with $FDR = FWER$ under the complete null.

Thus, for a fixed criterion α for controlling the Type I error rates, the order reverses for the number of rejected hypotheses R: procedures controlling the FWER are generally more conservative than those controlling either the FDR or PCER.

p–value adjustment

If interest is in controlling the FWER, the **adjusted p–value** for hypothesis H_g is:

$$\tilde{p}_g = \inf \{\alpha : \mathrm{H}_g \text{ is rejected at FWER } \alpha\}.$$

Hypothesis H_g is rejected at FWER α if $\tilde{p}_g \leq \alpha$.

Adjusted p–values for other Type I error rates are defined similarly.

p–value adjustment

- The level of the test does not need to be specified in advance.
- Some multiple testing procedures are most conveniently described in terms of their adjusted p–values.
- Adjusted p–values can usually be easily estimated using resampling.
- For any given procedure, adjusted p–values provide a convenient way of relating the Type I error rate to the number of rejected hypotheses.
- Different multiple testing procedures can be readily compared based on their respective adjusted p–values.

Notation

For hypothesis H_g, $g = 1, \ldots, G$

observed test statistic: t_g

observed unadjusted p–value: p_g.

Ordering of the observed absolute test statistics: $\{r_g\}_{g=1,\ldots,G}$

such that $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_G}|$.

Ordering of the observed unadjusted p–values: $\{r_g\}_{g=1,\ldots,G}$

such that $p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_G}$.

The corresponding random variables are denoted by upper case letters.

Control of the FWER

Bonferroni **single–step** adjusted p–values

$$\tilde{p}_g = \min(Gp_g, 1).$$

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

$$\tilde{p}_{r_g} = \max_{k=1,\ldots,g} \left\{ \min\left((G-k+1)\, p_{r_k}, 1 \right) \right\}.$$

Hochberg (1988) **step–up** adjusted p–values (Simes inequality)

$$\tilde{p}_{r_g} = \min_{k=g,\ldots,G} \left\{ \min\left((G-k+1)\, p_{r_k}, 1 \right) \right\}.$$

Control of the FWER

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

$$\tilde{p}_{r_g} = \max_{k=1,\ldots,g} \left\{ p\left(\min_{l \in \{r_k,\ldots,r_G\}} P_l \leq p_{r_k} \mid H_0^C \right) \right\}.$$

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

$$\tilde{p}_{r_g} = \max_{k=1,\ldots,g} \left\{ p\left(\max_{l \in \{r_k,\ldots,r_G\}} |T_l| \geq |t_{r_k}| \mid H_0^C \right) \right\}.$$

maxT and minP adjusted p–values

- Step–down procedures: successively smaler 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.
- Fast permutation algorithm for minP adjusted p–values implemented in R `multtest` package (Ge & Dudoit, 2002).

maxT and minP adjusted p–values

- The maxT and minP adjusted p–values are the same when the test statistics are identically distributed.
- When the test statistics are not identically distributed, procedures based on maxT adjusted p–values can lead to unbalanced adjustments.
- maxT adjusted p–values are more tractable computationally than minP p–values.
- Procedures based on maxT adjusted p–values can be more powerful in "small n, large G" situations.

Control of the FDR

Benjamini & Hochberg (1995): step–up procedure which controls the FDR under some dependency structures

$$\tilde{p}_{r_g} = \min_{k=g,\dots,G} \left\{ \min\left(\frac{G}{k}\, p_{r_k}, 1 \right) \right\}.$$

Benjamini & Yekutieli (2001): conservative step–up procedure which controls the FDR under general dependency structures

$$\tilde{p}_{r_g} = \min_{k=g,\dots,G} \left\{ \min\left(a_G \frac{G}{k}\, p_{r_k}, 1 \right) \right\}.$$

where $a_G = \sum_{g=1}^{G} 1/g \approx \log G$ for large G.

Yekutieli & Benjamini (1999): resampling based adjusted p–values controlling the FDR under certain dependency structures.

Significance Analysis of Microarrays, SAM

Order statistics: $T_{(1)} \geq \cdots \geq T_{(G)}$.
Permutation estimates of the expected values of the order statistics under the complete null: $\bar{t}_{(g)}$, $g = 1, \cdots, G$.

1. Efron et al. (2000). Reject $\mathrm{H}_{(g)}$ if

$$|t_{(g)} - \bar{t}_{(g)}| \geq \Delta,$$

where Δ is chosen based on a permutation estimate of the PFER under the complete null.

Adjusted p–values (for PCER):

$\tilde{p}_{(g)} = \sum_{l=1}^{G} p\left(|T_{(l)} - \bar{t}_{(l)}| \geq |t_{(g)} - \bar{t}_{(g)}| \mid \mathrm{H}_0^C\right)/G.$

Only weak control of the PFER.

The adjusted p–values are not monotone in g, i.e., in the test statistics.

Significance Analysis of Microarrays, SAM

2. Tusher et al. (2001). Reject H_g if $t_g \geq cut_{up}(\Delta)$ or $t_g \leq cut_{low}(\Delta)$, where $cut_{low}(\Delta)$ and $cut_{up}(\Delta)$ are chosen from the Quantile–Quantile plot of $t_{(g)}$ vs. $\bar{t}_{(g)}$ and based on a permutation estimate of the PFER under the complete null.

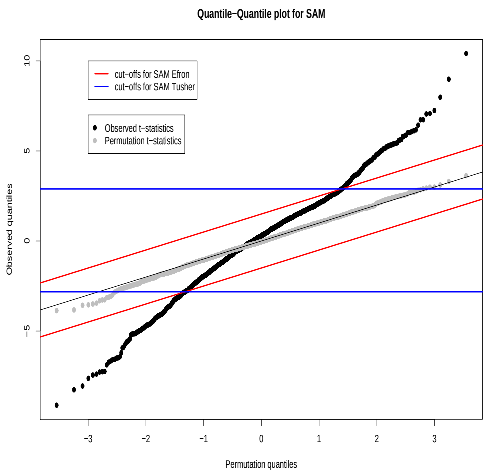
Order statistics are not used in estimating the PFER. The PFER is thus controlled in the strong sense.

For binary outcomes, SAM is similar to a t–test for each gene using asymmetric cut–offs.

The SAM estimate of the FDR is $E_0(V)/R$ – can be greater than one.

Dudoit et al. (2002)

Efron et al. vs. Tusher et al. SAM versions



Neighborhood analysis

Golub et al. (1999)

Reject H_g if $|t_g| \geq c$ andet

$$r(c) \quad = \quad \sum_{g=1}^{G} I(|t_g| \geq c) = \text{observed number of rejected hypotheses}$$

$$R(c) \quad = \quad \sum_{g=1}^{G} I(|T_g| \geq c) = \text{r.v. for number of rejected hypotheses.}$$

Chose a critical value c such that

$$G(c) = p\big(R(c) \geq r(c) \mid \mathrm{H}_0^C\big) = \alpha,$$

where $G(c)$ is estimated by permutation.

Neighborhood analysis

Difficulties:

- G(c) is a left–continuous function, with discontinuities at $|t_g|$.
- G(c) is not monotone in $c \Rightarrow$ which c to choose?
- G(c) is a random variable.
- What type of error rate control is really achieved?

Adjusted p–values for neighborhood analysis

Order statistics $|T|_{(1)} \geq \cdots \geq |T|_{(G)}$.
Step–down adjusted p–values

$$\tilde{p}_{(g)} = \max_{k=1,\dots,g} \Big\{ p\big(|T|_{(k)} \geq |t|_{(k)} \mid \mathrm{H}_0^C\big) \Big\}.$$

Step–up adjusted p–values

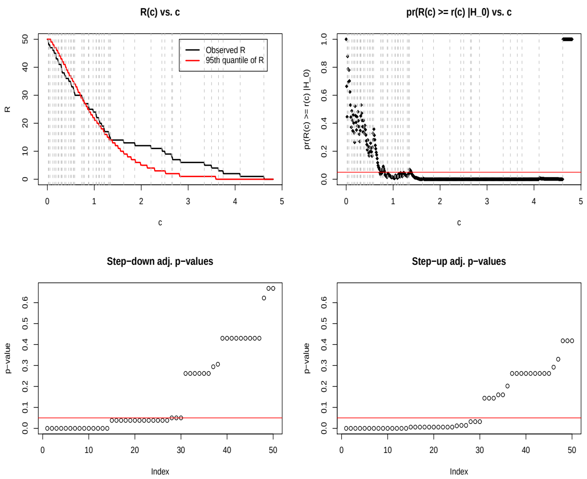
$$\tilde{p}_{(g)} = \min_{k=g,\dots,G} \Big\{ p\big(|T|_{(k)} \geq |t|_{(k)} \mid \mathrm{H}_0^C\big) \Big\}.$$

The procedure is based on the distribution of the order statistics under the complete null hypothesis
$\Longrightarrow$ in general **only weak control** of the Type I error rate.

Step–down procedure controls FWER weakly, step–up procedure does not.

Dudoit et al. (2002)

Adjusted p–values for neighborhood analysis



Host genomic responses to pathogenic bacteria
Jen Boldrick, Stanford

In vitro study of the gene expression response of human peripheral blood mononuclear cells (PBMCs) to infection by pathogenic bacteria.

Monitor the effect of **three factors** on the expression response

- Bacteria: Gram–negative, *B. pertussis*,
  Gram–positive, *S. aureus*;
- Dose: 1X, 10X, 100X, and 1000X;
- Time: 0.5, 2, 4, 6, and 12 hours,
  and also 1 and 24 hours for dose 100X.

Microarray data

- Lymphochip: 18,432 cDNA probes.
- 44 hybridizations
  (2 × 4 × 5 plus 1 and 24 hour measurements for dose 100X)
  - Cy5: mRNA from PBMCs t hours after infection by bacteria b at dose d;
  - Cy3: reference pool of mRNA from 6 immune cell lines.

$\Longrightarrow$ Expression response of gene g at time t in PBMCs infected by bacteria b at dose d (after normalization):

$$x_{gbdt} = \log_2 R/G.$$

Differentially expressed genes

Question. Identify the genes that have a different expression response to infection by the Gram + and Gram − bacteria.

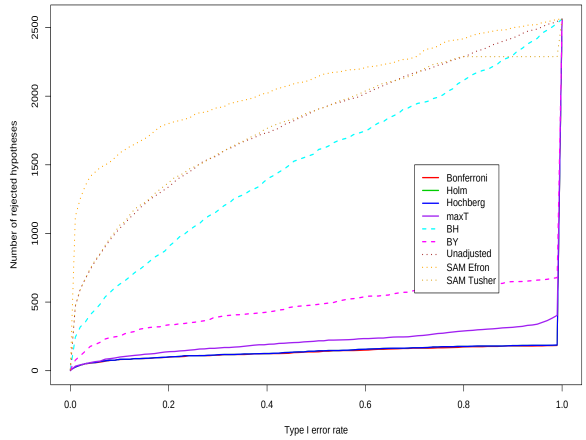
Approach. Simultaneously test G null hypotheses, one for each gene g

H_g : no bacteria effect on the expression response of gene g.

- Compute a paired t–statistic for each gene.
- Compute permutation p–values from the distribution of the test statistics for the 2^{22} permutations of the responses *within* the 22 dose × time blocks.
- Adjust for multiple hypothesis testing.

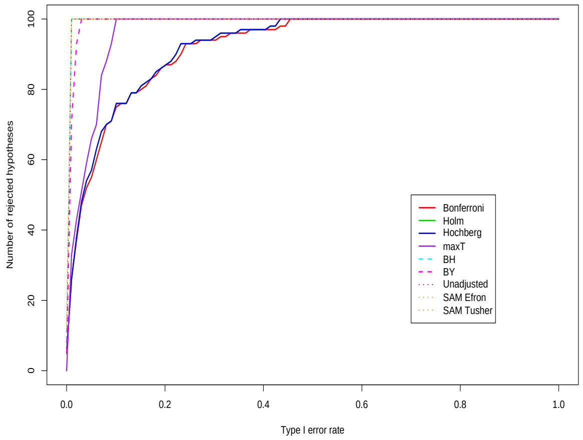
Plot of number of rejections vs. Type I error rate

$r(\alpha) = \sum_{g=1}^{G} I(\tilde{p}_g \leq \alpha)$ **vs.** α



Plot of number of rejections vs. Type I error rate

$r(\alpha) = \sum_{g=1}^{G} I(\tilde{p}_g \leq \alpha)$ **vs.** α



Results

- 66 spotted DNA sequences had maxT adjusted *p*-values less than 0.05.
- Several of these sequences actually corresponded to the same genes: CD64 (3 copies), Iκ B alpha (5 copies), SHP–1 (2 copies), plasma gelsolin (2 copies).
- The nature of the differential response varied among genes, as they exhibited different dose responses to infection by the pathogens.

A FAQ

Q: What about pre–screening to reduce the number of tests with the aim of increasing power?

A: The Type I error rate is controled at the claimed level in situations where

- we only focus on a subset of genes that are of interest – selected *before* looking at the data;
- the statistic used for screening is independent of the test statistic under the null.

Other situations still need to be better understood.

Discussion

- In multiple testing situations, there are several possible definitions for the Type I error rate (FWER, PCER, PFER, or FDR).
- FDR controlling procedures are promising alternatives to more conservative FWER controlling procedures.
- Strong control of the Type I error rate is essential in the microarray context.
- Adjusted p–values provide flexible summaries of the results from a multiple testing procedure.

Discussion

- Substantial gains in power can be obtained by taking into account the joint distribution of the test statistics (e.g. Westfall & Young (1993)).
- More work is needed to develop procedures that take into account the joint distribution of the test statistics.
- Resampling methods are needed to estimate adjusted p–values for complex multivariate datasets.
- 2D–multiple testing problems: thousands of genes, several hypotheses for each gene.

Discussion

Rather than choosing a specific error rate to control:

1. Choose a number r of hypotheses to reject with which the researcher feels comfortable. Evaluate the adjusted p–values $\tilde{p}_{(r)}$ necessary to reach this number under various procedures and types of error control.
2. For a given level, find the number of hypotheses that would be rejected under one method, and give the level required to achieve that number under other methods.
3. Find the number of hypotheses that would be rejected using a procedure controlling FWER at a fixed level, and find how many others would be rejected using procedures controlling FDR and PCER at that level.

Discussion

- Microarray experiments have renewed the interest in multiple testing
  - → lots of papers;
  - → old methods with new names;
  - → new methods with inadequate or unknown control properties;
  - → a lot of confusion!
- New proposals should be formulated precisely, within the standard statistical framework, to alow a clear assessment of the properties of different procedures.
- The same applies to other problems, such as clustering and classification.

R multiple testing software

- Bioconductor R `multtest` package.
- Multiple testing procedures for controlling
  - **FWER**: Bonferroni, Holm (1979), Hochberg (1986), Westfall & Young (1993) maxT and minP.
  - **FDR**: Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001).
- Tests based on t– or F–statistics for one– and two–factor designs.
- Permutation procedures for estimating adjusted p–values.
- Fast permutation algorithm for minP adjusted p–values.
- Documentation: tutorial on multiple testing.

References

- Benjamini & Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*. Vol. 57, p. 289–300.
- Dudoit et al. (2002). Multiple hypothesis testing in microarray experiments.
- Efron et al. (2000). Microarrays and their use in a comparative experiment. Tech. Report, Department of Statistics, Stanford University.
- Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. Vol. 286, p. 531–537.
- Shaffer (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.*. Vol. 46, p. 561–584.
- Tusher et al. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci.*. Vol. 98, p. 5116–5121.
- Westfall & Young (1993). *Resampling-based multiple testing: examples and methods for p–value adjustment.*

Acknowledgments

Jen Boldrick, Microbiology & Immunology, Stanford.

Pat Brown, Biochemistry, Stanford.

Yongchao Ge, Statistics, UC Berkeley.

David Relman, Microbiology & Immunology, Stanford.

Juliet Shaffer, Statistics, UC Berkeley.