

Inference on Gene Expression Changes as Measured with DNA Microarrays

Chiara Sabatti, UCLA

**DNA Microarray: Principles and Biotechnological Applications
UCLA September 10-13, 2001**

Agenda

- Assume preprocessing (see lectures of prof Liao).
- Sources of error.
- Role of statistical models in extracting information from data.
- One goal: estimate the change of expression
- Curses and blessing of high dimensional data.

Bibliography

(in order of appearance)

Kerr and Churchill (2001), Biostatistics, 2:183-201.

M.A. Newton et al. (2001), Journal of Computational Biology, 8:37-52.

G. Tzeng et al. (2001), Nucleic Acids Research, 29:2549-2557

S. Dudoit et al. (2000), Berkeley Stat. Tech Rep 578.

Tusher, Tibshirani and Chu (2001), PNAS 2001 98: 5116-5121.

Scientific question

We consider two different biological conditions and we want to evaluate the change in expression for a set of genes between these two conditions.

Notation:

genes are indicated by $i = 1, \dots, N$;

θ_i indicates for every gene, the change in expression between the two conditions.

Question: What are the values θ_i ? Which of them are different from 0?

Data-Notation

Basic Measurement: on one array, for each spot, we have:

r is the log-intensity of the
(background corrected and normalized) red-channel and
 g the log-intensity of the the green channel;

$y = r - g$ the log-ratio of the intensities;

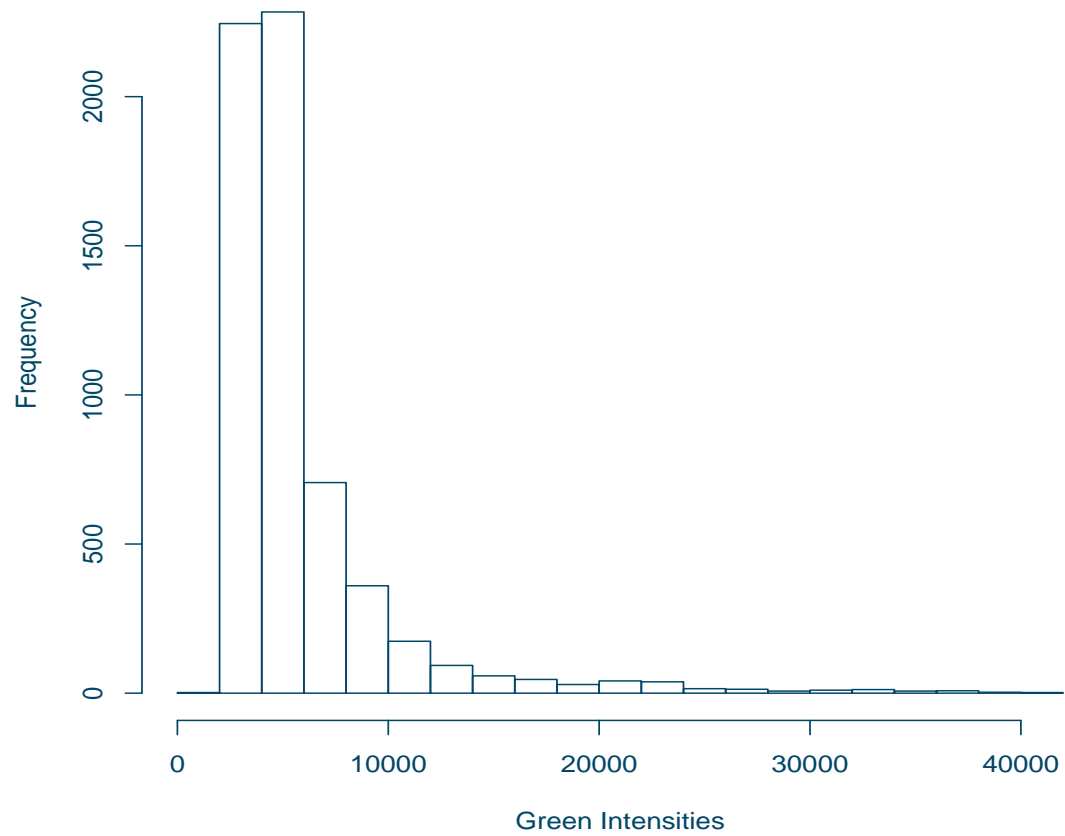
Number of measurements:

- $i = 1, \dots, N$ genes
- $l = 1, \dots, E$ experiments
- $k = 1, \dots, K$ slides for each experiment;
- $j = 1, \dots, S$ spots where the gene i is printed on each slide;

Data : y_{ikjl} is log ratio of intensities for gene i on slide k , spot j and experiment l .

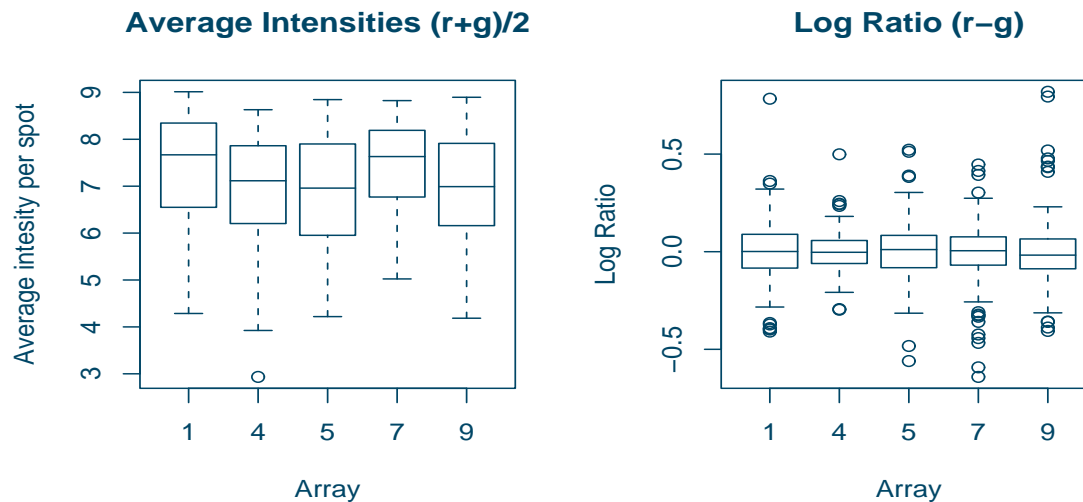
Why log?

Histogram of Intensities



Why differences of log-intensities

- In statistical lingo “paired data”
- There is a lot of spot specific variability (printing + hybridization) that can be eliminated by looking at $y = r - g$
- Possibly work only with r and g (Newton, Anova model)



The role of replication

⇒ identify the sources of variability;

⇒ obtain robust estimates;

When we re-do the experiment:

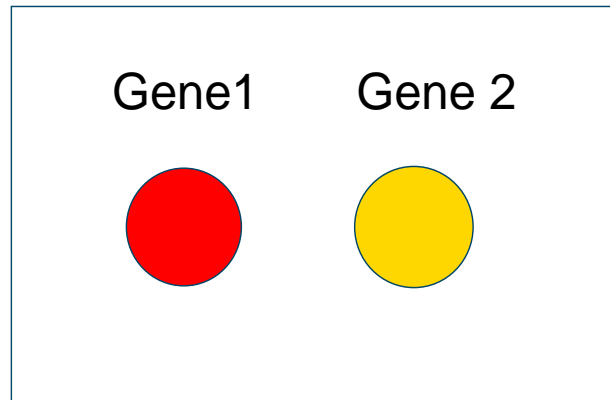
1. different **biological material** (“experiment”, or “subject”)
2. different **sample preparation**: purification, labeling, etc.
3. different **slide** (printing variability)
4. different **hybridization**

Statistical Models

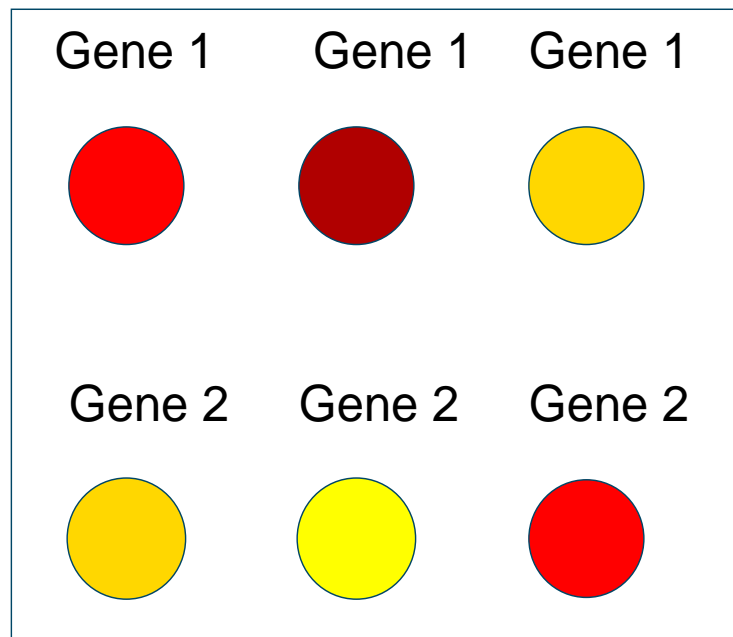
1. Experimental Design and ANOVA
2. One-slide hierarchical Bayes
3. Multiple slides hierarchical Bayes
4. Non-parametric

1. Experimental Design and Microarrays

- Proposed by Fisher. Original problem: how does a fertilizer work? To test the effect of a fertilizer, you have to control for seed type, irrigation, sun exposure, wind, etc..
- Microarray are one experiment we can design. Response: expression level; Effect of interest: interaction of cell condition/type and genes; Factors to control for: gene, cell condition/type, dye, slide, spot, subject, various interactions,...
- To estimate effects, avoid confounding: ex. gene and spot are often confounded.
- To avoid confounding, introduce the appropriate repetitions.



Is the difference due to genes or spots?



Multiple spots for every gene help

Experimental Design and ANOVA

An ideal design:

1. multiple experiments
2. for each experiment multiple arrays
3. dye switching
4. the same gene spotted multiple times on the same array
5. the spots for each gene are selected randomly and change from array to array

Possible to estimate (ANOVA model + Bootstrap):

$$r_{ijkl} = \text{Gene}_i + \text{Condition} + \text{Gene}_i\text{Condition} + \\ + \text{Dye} + \text{Slide}_k + \text{Spot}_j + \text{Experiment}_l + \dots$$

ANOVA for microarrays?

Program (Kerr and Churchill, 2001)

1. design an experiment with many replicates
2. estimate effect with appropriate averages (work with log-intensities)
3. evaluate significance with Bootstrap (normal approximation not valid)

Problems:

- much less replicates available
- dye effect is non-linear

Take Home Message: plan experiments that are as close as possible to the ideal one

Curses and blessing of dimensionality

Curses : many genes \longrightarrow each gene spotted often only once;
to estimate interactions we need a lot of repetitions;

Blessing : many genes \longrightarrow they are similar to each other, we
should be able to pool some information across them

2. A one-slide Bayesian hierarchical model

Newton et al. 2001

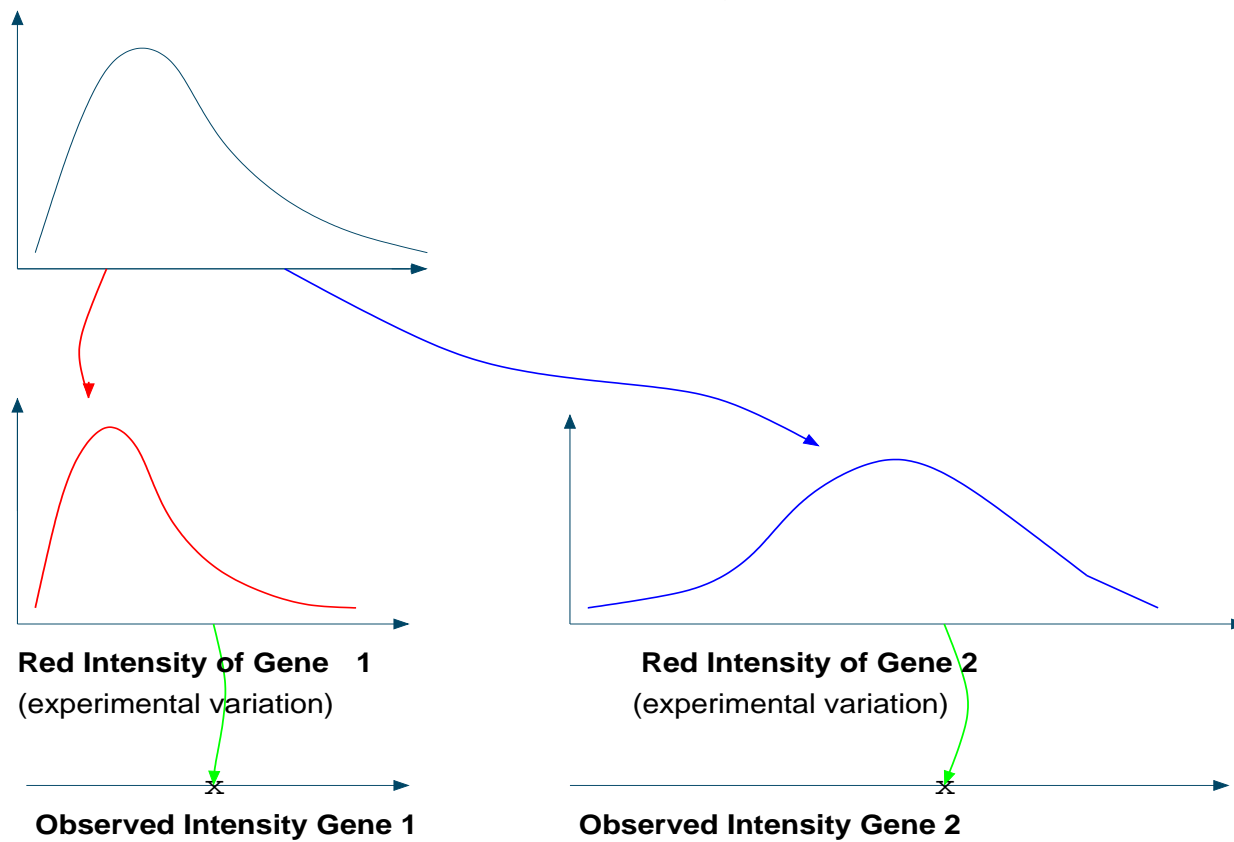
Caution An extreme solution. One slide never satisfactory.

The model

- 1 slide
- one spot per gene
- work with the intensity measurements (10^{r_i} , 10^{g_i} in previous notation): R_i, G_i
- Gamma-gamma model:
 $R_i \sim \text{Gamma}(\theta_i^r, a); \quad G_i \sim \text{Gamma}(\theta_i^g, a)$
 $\theta_i^r, \theta_i^g \sim \text{Gamma}(\alpha, \nu) \quad \forall i.$
- R_i is independent from R_j

Gamma-Gamma Model of Intensities

Red intensities for all genes (differences between genes)



Take home message

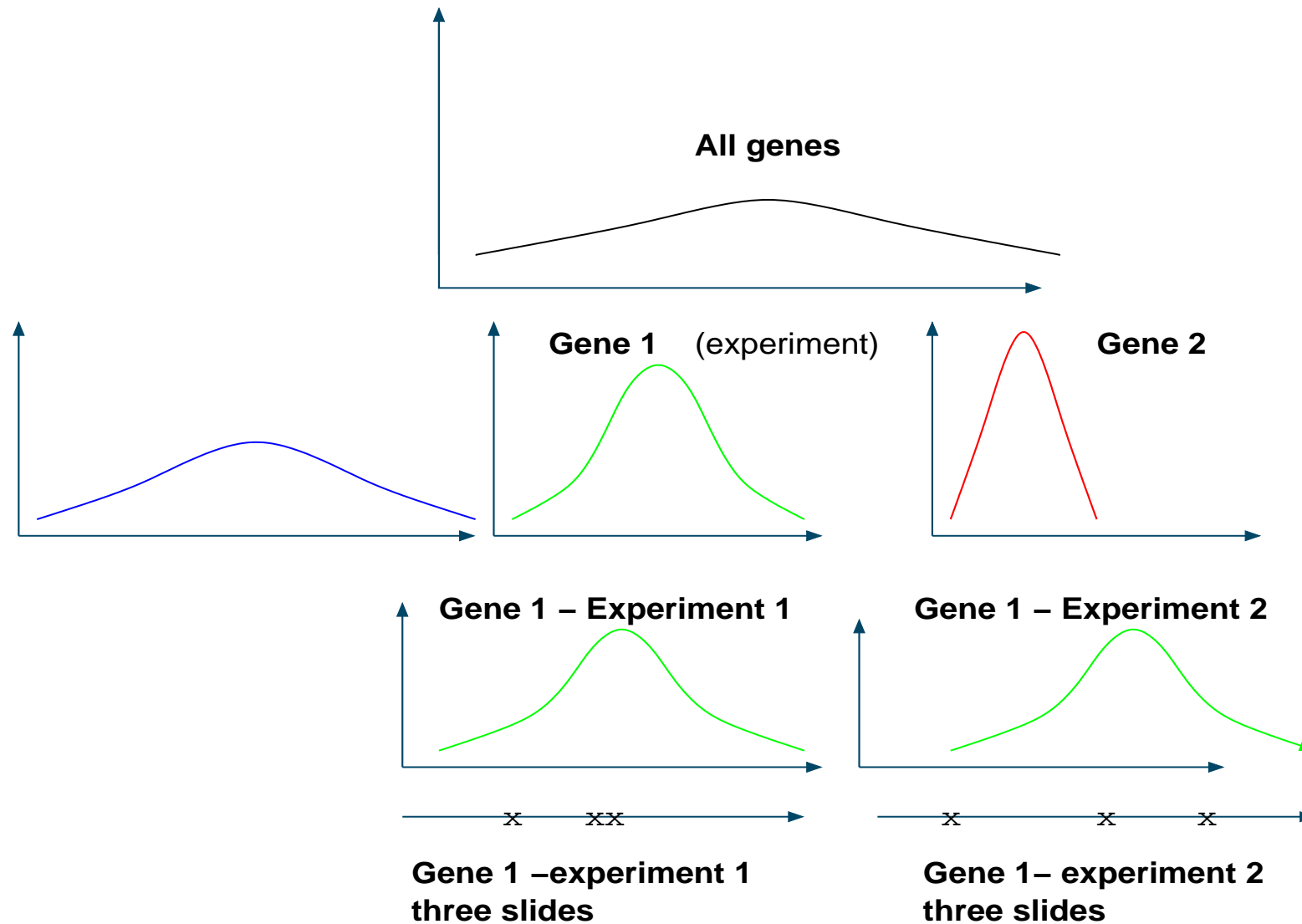
- Common features between genes are emphasized;
- Empirical Bayes technique to fix value of unknown parameters (a, α, ν) .
- We get a posterior distribution for each θ_i^r : we can evaluate the probability with which there is a change in expression.
- BUT: very unrealistic and strong assumptions.

3. A more realistic hierarchical model

Tseng et al. 2001

- a series of experiments with the same conditions;
- multiple slides with the same experiment;
- multiple spot for the same gene on each slide;
- work with log-ratios y ;
- assume normality of log-ratios, with different means and variances for each gene.
- Fix hyperparameters with empirical Bayes method and using calibration slides (baseline condition vs baseline condition)
- evaluate posterior distribution of θ_i s using a Markov Chain Monte Carlo method.

Scheme of Randomness in the model



The model

- Each gene g is described by the following parameters:
 - θ_g : overall mean log ratio for the condition under study across experiments and slides;
 - μ_{ge} : mean log ratio for one particular experiments across cells;
 - σ_g : variability of log ratio from experiment to experiment;
 - τ_g : variability of log ratio for one experiment across slides.
- The distributional assumptions are as follows:
 - The prior on σ_g and τ_g are the same for all genes.
 - The prior on θ_g is flat
 - $\mu_{ge} \sim \mathcal{N}(\theta_g, \sigma_g^2)$
 - $y_{ges} \sim \mathcal{N}(\mu_{ge}, \tau_g^2)$
 - Given σ_g and τ_g , μ_{ge} and y_{ges} are independent.

Results

- To estimate the parameters use MCMC.
- One gets confidence statements on the value of θ_g :

$$\text{Prob}(\theta_g \in (a, b)) = ??$$

- ⇒ Captures some of the **dependency** between measurements for different genes—there may be more, however.
- ⇒ Based on assumption of **normality** –questionable;
- ⇒ Can evaluate posterior probability of expression changes, but ignores **multiple comparison** effects.

Non parametric approaches

Dudoit et al. 2000; Tusher et al. 2001

- Minimize model distribution assumptions.
- Experimental design: n_C slides comparing baseline condition vs baseline condition (C) and n_D slides comparing baseline vs study sample (D).
- Variability between slides should be the same: each slide a different subject or replicate of the same experiment.

- **T statistic:**

$$T_i = \frac{\bar{y}_{iD} - \bar{y}_{iC}}{\sqrt{s_{iD}^2/n_D + s_{iC}^2/n_C}}$$

classical statistical index of how the means are different.

- Use permutation/bootstrap methods to assess significance.
- Multiple comparisons.

Significance of the T statistics

- When is the difference between average expression in groups C and D significant?
- If one has many replicates, normal distribution.
- Otherwise, try to calculate the frequency with which we record a difference as large as the observed one, when there is no real distinction between C and D .
- Permutation of “labels”: if no difference between C and D any of the observed expression values could come from either conditions.

Permutations and T statistics

Original data:

| <i>genes</i> | <i>C</i> | <i>C</i> | <i>C</i> | <i>C</i> | <i>D</i> | <i>D</i> | <i>D</i> | <i>D</i> | <i>T-stat</i> |
|--------------|------------|------------|------------|------------|------------|------------|------------|------------|----------------------|
| | <i>s1</i> | <i>s2</i> | <i>s3</i> | <i>s4</i> | <i>s5</i> | <i>s6</i> | <i>s7</i> | <i>s8</i> | |
| <i>g1</i> | <i>y11</i> | <i>y12</i> | <i>y13</i> | <i>y14</i> | <i>y15</i> | <i>y16</i> | <i>y17</i> | <i>y18</i> | <i>T₁</i> |
| <i>g2</i> | <i>y21</i> | <i>y22</i> | <i>y23</i> | <i>y24</i> | <i>y25</i> | <i>y26</i> | <i>y27</i> | <i>y28</i> | <i>T₂</i> |
| <i>g3</i> | <i>y31</i> | <i>y32</i> | <i>y33</i> | <i>y34</i> | <i>y35</i> | <i>y36</i> | <i>y37</i> | <i>y38</i> | <i>T₃</i> |
| <i>g4</i> | <i>y41</i> | <i>y42</i> | <i>y43</i> | <i>y44</i> | <i>y45</i> | <i>y46</i> | <i>y47</i> | <i>y48</i> | <i>T₄</i> |
| <i>g5</i> | <i>y51</i> | <i>y52</i> | <i>y53</i> | <i>y54</i> | <i>y55</i> | <i>y56</i> | <i>y57</i> | <i>y58</i> | <i>T₅</i> |

Permutations and T statistics

Permuting the labels of the columns:

Permutation 1: C D C D C C D D \implies T-stat

| <i>genes</i> | <i>C</i> | <i>D</i> | <i>C</i> | <i>D</i> | <i>C</i> | <i>C</i> | <i>D</i> | <i>D</i> | <i>T-stat</i> |
|--------------|------------|------------|------------|------------|------------|------------|------------|------------|----------------------|
| | <i>s1</i> | <i>s2</i> | <i>s3</i> | <i>s4</i> | <i>s5</i> | <i>s6</i> | <i>s7</i> | <i>s8</i> | |
| <i>g1</i> | <i>y11</i> | <i>y12</i> | <i>y13</i> | <i>y14</i> | <i>y15</i> | <i>y16</i> | <i>y17</i> | <i>y18</i> | <i>T₁</i> |
| <i>g2</i> | <i>y21</i> | <i>y22</i> | <i>y23</i> | <i>y24</i> | <i>y25</i> | <i>y26</i> | <i>y27</i> | <i>y28</i> | <i>T₂</i> |
| <i>g3</i> | <i>y31</i> | <i>y32</i> | <i>y33</i> | <i>y34</i> | <i>y35</i> | <i>y36</i> | <i>y37</i> | <i>y38</i> | <i>T₃</i> |
| <i>g4</i> | <i>y41</i> | <i>y42</i> | <i>y43</i> | <i>y44</i> | <i>y45</i> | <i>y46</i> | <i>y47</i> | <i>y48</i> | <i>T₄</i> |
| <i>g5</i> | <i>y51</i> | <i>y52</i> | <i>y53</i> | <i>y54</i> | <i>y55</i> | <i>y56</i> | <i>y57</i> | <i>y58</i> | <i>T₅</i> |

Permutation 2: D C D C D C C D \implies T-stat

...

P-value and multiple comparison

$$\text{P-value for gene } i : \frac{\#\text{Permutations : } T_i^p \geq T_i}{\#\text{Permutations}}.$$

⇒ If P-value = 0.05, there is 1/20 chances that I declare that $\theta_i \neq 0$, when really $\theta = 0$. One chance out of 20 is generally considered like acceptable margin of error.

10000 genes: if all the $\theta_i = 0$, I make a mistake $1/20 \times 10000 = 500$ times

Remember the curses of dimensionality?

Error types

Two hypothesis:

H_0 (null): no change of expression (the “standard” hypothesis)

H_1 (alternative): change of expression

Two error types:

(1) wrongly rejecting H_0 ; (2) wrongly accepting H_0

What do we control?

The probability of wrongly rejecting H_0
has to be small (this is what we compare with P-value)

Multiple tests

If N null hypothesis are considered at the same time, count the total errors:

| | Accept H_0 | reject H_0 | |
|-------------|--------------|--------------|-------|
| H_0 true | U | V | N_0 |
| H_0 false | T | S | N_1 |
| | $N - R$ | R | N |

What do we control?

FWER (family-wise error rate): $Pr(V \leq 1)$ Stronger

FDR (false discovery rate): $E\left(\frac{V}{R}\right)$

Controlling error rate in multiple tests

- **Bonferroni correction:** controls FWER; too strong if tests are dependent.

⇒ use a level α/N for each of the N tests

- **Westfall-Young:** step-down permutation procedure for FWER (Dudoit et al. 2000)
- **FDR?** First attempt in Tusher et al. 2001.

Non-parametric methods—Summary

- **No dangerous model assumptions.**
- **No great use of information on experiments and slides (which pair of slide should be similar than others).**
- **Serious problem of multiple comparisons.**

Some slogans

- Design the experiments to avoid confounding
- Pool information across genes whenever possible
- Need replicates
- Be aware of problem of multiple comparisons
- Watch the literature.

Statistical topics

Design of Experiments → Freedman et al.

ANOVA → Rice

T-statistics → Rice

Bayesian hierarchical models → Gelman et al.

Test of Hypothesis → Rice

Multiple testing → Rice

Statistical Reference

Rice “Mathematical Statistics and data analysis”, 2nd ed.,
Duxbury

Freedman, Pisani, Purves “Statistics”, 3rd ed., Norton

Gelman, Carlin, Stern and Rubin “Bayesian Data Analysis”,
Chapmann Hall.

References

<http://www.stat.ucla.edu/~sabatti/statarray/index.html>