The Use of Statistics in Microarray Studies

Prof. Ernst Wit
University of Groningen
e.c.wit@rug.nl
http://www.math.rug.nl/~ernst

Outline

- Pooling experimental material
- Dual-channel microarray designs
  - Loop designs vs. reference designs?
  - Dye-swap designs?
- How to detect differentially expressed genes?
  - How to deal with nuisance factors in the exp'1?
  - How to deal with thousands of p-values?

Why do a microarray experiment

- To discover
  - "genes that are differentially expressed between tissues"
  - "a typical genomic profile of a particular disease"
  - "differences between phenotypic similar pathologies"
- *i.e.*, to find some differences
The use of statistics in microarray studies
Prof. Ernst Wit

Why design a microarray experiment?

- We want to be sure that the differences we find are
  - due to biological differences (e.g. different tissue, different diseases)
  - not due to other experimental conditions (e.g. microarray batch)
  - not due to chance variation
- Our exp’t should maximize the difference between biological conditions

4 known design principles

1. Replication: Vary the conditions of interest, with replicates that are
   - as many as possible
   - biological rather than technical replicates
   - as evenly distributed across the conditions of interest
2. Restriction: keep conditions not of interest constant
3. Blocking: keep track of uninteresting conditions that do vary
4. Matching: channels are similar units for effective comparisons

A statistical toy model for gene expression

\[ \log_{10} \text{expression} = \delta_{g} + \beta_{b} + \alpha_{g} + \beta_{a}(x) + \delta_{b}(\theta_{b}) + \beta_{e} + \gamma_{l} + \varepsilon_{r} + \varepsilon_{d} \]

- \( \delta_{g} \): true expression for gene \( g \) under condition \( c \)
- \( \beta_{b} \): RNA Batch effect (experimenter, time of day, temperature)
- \( \alpha_{g} \): array effect (scanning level, pre/post-washing)
- \( \beta_{a}(x) \): location effect (chip, coverslip, washing)
- \( \delta_{b}(\theta_{b}) \): dye effect (dye, unequal mixing of mixtures, labelling, intensity)
- \( \gamma_{l} \): print pin effect
- \( \varepsilon_{r} \): spot effect (amount of DNA in the spot printed on slide)
- \( \varepsilon_{d} \): biological variation for individual \( k \)
- \( \varepsilon_{d} \): within-replicate, variation of technical replicate \( r \)
Normalization

The proposed model:

\[
\log_{\text{base}} \text{signal} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5 + \epsilon_6 + \epsilon_7 + \epsilon_8 + \epsilon_9 + \epsilon_{\text{other}}
\]

It is hoped that the structural artifacts,

\[
\beta_0, \beta_1, \beta_2, \beta_3, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6, \epsilon_7, \epsilon_8, \epsilon_9, \epsilon_{\text{other}}
\]

can be eliminated by means of “normalization”

This leaves us a model for gene \( g \) at condition \( c \) on array \( a \) for the \( j \)th technical replicate of individual \( k \):

\[
\log_{\text{base}} \text{signal} = \delta_{jc} + \epsilon_{jc} + \epsilon_{\text{other}},
\]

Where \( \epsilon_{\text{other}} \) includes non-structural artifacts from nuisance effects.

Rephrasing our three questions

From the model:

1. How to deal with biological variation \( \epsilon_{\text{bio}} \) ?
2. How to estimate differential expression \( \delta_{\text{diff}} \) ?
3. How to deal with bias \( \delta_{\text{bias}} \) ?

- Simple answer: “taking log-ratios”
- Slightly less simple: “insert a random effect” (see analysis)

Pooling reduces biological variation

Minimize error \( \min_{\theta} (\theta)^2 \) subject to budget \( B = n_k G_0 + n_k G_0 \)

- \( G_0 \) = costs individual sample extraction
- \( G_0 \) = cost of a measurement (microarray, dyes, etc.)

The screen versions of these slides have full details of copyright and acknowledgements
The use of statistics in microarray studies
Prof. Ernst Wit

Optimal design in 2-channel microarrays

Question: "which design is more efficient in estimating the contrasts?"

Reference Design

Loop Design

Note: each arrow stands for 1 microarray (green dye red dye)

Use od: optimal (interwoven loop) designs in R

The function

\[ \text{Od}(\text{nt}, \text{ns}, \text{optimality} = \ldots, \text{method} = \ldots) \]

- uses simulated annealing to find
  - L-optimal/D-optimal
  - interwoven loop/all designs
  - for any number of conditions
  - and any number of slides
- Can be obtained from smida library at:
  

Some conclusions...

Loop designs are often optimal...

- L-optimal designs for 7/8 conditions with 7/8 slides, respectively
The use of statistics in microarray studies
Prof. Ernst Wit

...but not always

- L-optimal design with 9 conditions and 9 microarrays

Close-to-optimal alternative: interwoven loop-designs

- An interwoven loop design is defined by:
  - number of conditions (HERE: 15)
  - number of loops (HERE: 3)
  - jump size for each loop (HERE: 1, 4, 6)

Advantages of interwoven loop designs

- Easy to implement – Compare left with right
- Highly efficient – Left is only 0.6% more efficient than right
- Automatic "dye balance" – Conditions measured equally often with Cy3/Cy5
The use of statistics in microarray studies
Prof. Ernst Wit

Dye swap or dye balance?

➢ The simple model needs to be expanded for a possible dye effect:

\[
\log_{2}(Y_{i}) = \beta + \gamma_{i} + \epsilon_{i}
\]

➢ Dye swap designs have been proposed as a way to control the dye effect

➢ HOWEVER, interwoven loop designs are more efficient!

<table>
<thead>
<tr>
<th>D</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 cond. 10 arrays</td>
<td>64% 80%</td>
</tr>
<tr>
<td>6 cond. 12 arrays</td>
<td>50% 75%</td>
</tr>
<tr>
<td>7 cond. 14 arrays</td>
<td>37% 60%</td>
</tr>
<tr>
<td>8 cond. 16 arrays</td>
<td>25% 62%</td>
</tr>
</tbody>
</table>

Relative D- and L-efficiency of dye-swap versus interwoven loop

Analysis

➢ So, you have to let yourself be convinced by a statistician to do a more complicated microarray design

➢ It is “optimal”, alright, but how can you estimate the effects of interest?

➢ Aim: a step by step guide to analyze the data

• Create a variable with fixed effects of interests

• Create one or more variables with random nuisance effects

An example: microarray analysis

➢ A microarray exp’t, consisting of 5 dual-channel slides, uses a loop design without biological replication; The data are suitably normalized

Loop Design

We create:

• Conditions = design matrix

• Arrays = random spot effect matrix, belonging to 5

• Bio.reps = random biological effect matrix, belonging to 5
The use of statistics in microarray studies
Prof. Ernst Wit

...keeping track of the model indices!

- The experiment involves only technical replicates, i.e.:
  - \( s_i = s_{ij} + s_{ik} + s_{il} \)
  - \( b_k = b_{ik} + b_{il} + b_{ij} \)
  - \( m_j = m_{ij} + m_{ik} + m_{il} \)
  - \( r_{ij} = r_{iij} + r_{iik} + r_{iil} \)

- Where
  - \( s_i \) is the spot effect for array \( i \)
  - \( r_{ij} \) is the \( i \)th biological replicate in the exp't

- Therefore:
  - \( \text{conditions} = (1,2,3,4,5,5,1) \)
  - \( \text{arrays} = (1,1,2,2,3,3,4,4,5,5) \)
  - \( \text{bio.reps} = (1,2,3,4,5,5,1) \)

Analysis of the data in R

- For a particular gene, the ten channel expression values are given as:
  - \( y = (0.62, 1.16, 1.51, 3.05, 2.78, 3.61, 3.61, 4.93, 5.12, 0.62) \)

Loop Design

- In R we use the library nlme with the commands:
  - \( \text{ex1<-lme(y~conditions, data=dat, random = list(grp = pdBlocked(list(pdIdent(~ -1 + bioreps), pdIdent(~ -1 + arrays)))))} \)

Results

- Mixed-effects model fit by ML:
  - Residual standard error: 1.084 on 19 degrees of freedom
  - Number of obs: 20, groups: 5, number of groups: 5

  Fixed effects:
  - \( y \sim \text{conditions} \)
    - Value Std.error t-value p-value
      - condition1 1.14905 0.68861 1.683 0.115
      - condition2 0.29273 0.68861 0.426 0.668
      - condition3 0.29273 0.68861 0.426 0.668
      - condition4 0.29273 0.68861 0.426 0.668
      - condition5 0.29273 0.68861 0.426 0.668

The screen versions of these slides have full details of copyright and acknowledgements
The use of statistics in microarray studies
Prof. Ernst Wit

And what does that mean?

- For this one gene, it means that at a 5% significance cut-off:
  - Condition 2 is not obviously different from condition 1
  - Conditions 3-5 are all significantly different from condition 1

And if you look at many genes simultaneously...?

- P-values from 10,000 genes for comparison between condition 1 and 2:

By looking at the right-hand figure, you can deduce:
- Total number of (non-)differentially expressed genes
- False Discovery Rate at each arbitrary cut-off

Conclusions

- We have met several microarray design issues and concluded:
  1. There are several design principles to obey when considering sources of variation of a microarray experiment
  2. Carefully assigning samples to conditions can improve estimation:
     - optimal designs are important in order to reduce cost/time
     - interwoven loop designs are a good compromise
     - pooling can be practically helpful
  3. Mixed effect models are excellent tools for analyzing microarray data
The use of statistics in microarray studies
Prof. Ernst Wit

More of this in: Statistics for Microarrays
- Ernst Wit & John McClure
- John Wiley & Sons
- All statistical aspects of
  - microarray design
  - microarray analysis
  - microarray inference
  - "microarray myths"

Many thanks for your attention!