Designing Microarray Experiments

Naomi Altman
Penn State University
Dept. of Statistics
Bioinformatics Consulting Center
naomi@stat.psu.edu

For the Arabidopsis 2005 Workshop
June 13, 2005
Outline

• Replicability
• 2 objectives of Microarray Experiments
• From organism to data
• Steps in Designing a Microarray Study
  – 1. Determine the objective of your study.
  – 2. Determine the experimental conditions (treatments).
  – 3. Determine the measurement platform
  – 4. Determine the biological sampling design.
  – 5. Determine the biological and technical replicates.
  – 6. Determine the RNA pool.
  – 7. Determine the hybridization design.
Understanding Variation

The basic tenet of scientific inference:

Phenomena worthy of scientific investigation are replicable.

(But just because something is replicable does not mean it is worth investigating.)
Detecting Replicable Expression

- **False Discovery Rate (FDR)**
  The percentage of genes declared significant which are actually just noise

- **False Nondiscovery Rate (FNR)**
  The percentage of genes which are significant but are missed in the study

- **Power**
  The probability that a significant gene is declared significant in the study
Study Objectives

• Inference about the technology

• Biological Inference
Study Objectives

- Inference about the technology
  - Technical Replication
- Biological Inference
  - Biological Replication
From Organism to Data

1. Grow Organism
2. Sample Tissue
3. Extract RNA
4. Spike and Label Sample
5. Hybridize
6. Scan
From Organism to Data

1. Grow Organism
2. Sample Tissue
3. Extract RNA
4. Spike and Label Sample
5. Hybridize
6. Scan

Biological Sample

Biological Subsample

Technical Replication
Steps in Designing a Microarray Study

1. Determine the objective of your study.

2. Determine the experimental conditions (treatments).

3. Determine the measurement platform

4. Determine the biological sampling design.

5. Determine the biological and technical replicates.

6. Determine the RNA pool.

7. Determine the hybridization design.

8. And for custom arrays ...
1. Determine the objectives of your study.

The level of replication should match the objectives:

For differential expression, we compare the mean difference between treatments to the variability of replicates at the correct level.

For gene expression clustering, we average among replicates at the correct level.
Where do we replicate?

1. Grow Organism
2. Sample Tissue
3. Extract RNA
4. Spike and Label Sample
5. Hybridize
6. Scan

}{ Biological Inference

}{ Individual

}{ Platform
Doing it right:

The cost of doing a properly replicated experiment is high, but not as high as the cost of doing a bad experiment.
2. Determine the experimental conditions (treatments) under study.

Genotypes – if comparing mutants, useful to have the background wild types

Tissues - if possible, take all from the same individuals

“stimuli” such as hormone treatments or exposure
What is the appropriate “control” condition?

Time course – multiple observations on the same individuals?
  - synchronization
  - when does the “action” occur?
2. Determine the experimental conditions (treatments) under study.

1 factor (e.g. genotype or tissue or stimuli or time)
Typically the experiment is a “one-way layout” and the treatments or conditions are just the unique levels of the factor.

2 or more factors
Typically the experiment is arranged as a factorial design using all combinations of factor levels.

The design is said to be balanced if there are equal numbers of replicates at all levels for all conditions.
2. Determine the experimental conditions (treatments) under study.

e.g. 2 genotypes Gg, 2 tissues Tt, 3 times, abc

The treatments are

<table>
<thead>
<tr>
<th>GTa</th>
<th>GTb</th>
<th>GTc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gta</td>
<td>Gtb</td>
<td>Gtc</td>
</tr>
<tr>
<td>gTa</td>
<td>gTb</td>
<td>gTc</td>
</tr>
<tr>
<td>gta</td>
<td>gtb</td>
<td>gtc</td>
</tr>
</tbody>
</table>

If we have 3 replicates, we will need 36 arrays.
2. Determine the experimental conditions (treatments) under study.

It is not necessary to have balanced experiments or to use factorial designs, but it simplifies the analysis.

Doing a pilot experiment is extremely useful for picking the appropriate experimental conditions.
Doing it right:

The cost of doing a pilot experiment is (almost) always repaid in a better experiment design. And pilot experiments are great for training.
3. Determine the measurement platform

-Affymetrix, radionucleotide, cDNA, oligonucleotide
-custom or off-the-shelf

The cost of arrays is high, but not as high as the cost of:
Repeating the entire experiment.
Validating all findings on another platform.
Losing a publication, grant opportunity or precedence.

The costs include:
Personnel time
Reagents
Analysis time
3. Determine the measurement platform

-Affymetrix, radionucleotide, cDNA, oligonucleotide
-custom or off-the-shelf

The cost of arrays is high, but not as high as the cost of:
Repeating the entire experiment.
Validating all findings on another platform.
Losing a publication, grant opportunity or precedence.

The costs include:
Personnel time
Reagents
Analysis time

“But I can get arrays free from my buddy …” It’s not free.
Doing it right:

The cost of microarrays is high, but not as high as the cost of redoing the experiment.
3. Determine the measurement platform

My current advice:

Use off-the-shelf whole genome arrays when available.
- less variability
- better informatics
- easier to compare across studies

Seek the advice of your local microarray center when available.

Be consistent with your collaborators (and competitors) when possible.

But, be open to newer and better technologies as they arise.
4. Determine the biological sampling design.

A block is a set of samples that have a natural clustering that reduces variability.

E.g. tissues from the same plant, plants grown together in May (versus Dec), plants grown in Madison versus Gainesville.

The most power comes from:

A strong protocol for sample collection.
As complete set of conditions in every block.
Replicating blocks.
4. Determine the biological sampling design.

  e.g. A time-course study of leaf development in 2 genotypes.
  
  Protocol: All leaves collected at 8 a.m. at specified stages.

  Blocking: 1 tray of each genotype grown together starting every 6 weeks. (3 replicates)

  One plant collected from each genotype at each stage and leaves harvested.
5. Determine the number of biological and technical replicates.

The correct analyses of biological experiments differentiate between technical and biological replicates.

Technical replication reduces measurement error.

Biological replication is used for statistical inference. The more biological replicates you have, the higher your power.

You need at least 3 biological replicates for each experimental condition for biological inference.
From Organism to Data

1. Grow Organism
2. Sample Tissue
3. Extract RNA
4. Spike and Label Sample
5. Hybridize
6. Scan

\{ Biological Sample \}
\{ Biological Subsample \}
\{ Technical Replication \}
5. Determine the number of biological and technical replicates.

Technical replication is not necessary, but it can be cost effective if it is much cheaper than biological replication. (But you still need 3 biological replicates.)

(Dye-swaps can be done using biological replicates.)

Spot replication is very cheap and reduces missing data.
5. Determine the number of biological and technical replicates.

How many replicates?

Sample sizes can be determined from power analysis. This requires a measure of biological variation and measurement error.

These can be based on previous experiments or pilot experiments, and the correct ratio of technical replicates to biological replicates can be determined by ANOVA.

The minimum acceptable sample size:
3 biological replicates

preferably: 2 spot replicates (custom 2-channel arrays)
Doing it right:

Design defensively:

Spots fail. (Duplicate spots if possible.)

Hybridizations fail. (Store extra samples.)
5. Determine the number of biological and technical replicates.

With good statistical assistance, you can develop good designs that require fewer replicates but ...

If you go this route, it would be wise to make your statistician a collaborator – not “just” a consultant!
6. Determine the RNA pool.

a. Standardization
b. Pooling
c. Amplification
d. Spiking
6. Determine the RNA pool.

a. Standardization – Since we want to compare samples it would be nice if they were quantitatively similar #cells or total RNA or total mRNA

b. Pooling

c. Amplification

d. Spiking
6. Determine the RNA pool.

a. Standardization

b. Pooling – Pooling biological samples is like averaging. This decreases biological variation, without increasing the number of arrays, and thus increases statistical power.

But we still need 3 independent pools.

c. Amplification

d. Spiking
Pooling

e.g. We have 3 biological samples and 3 arrays:

1 sample per array
• reduces both technical and biological variance by 1/3
• provides measure of biological variability for biological inference.

Pool all samples, hybridize to 3 arrays
• reduces both technical and biological variance by 1/3
• provides a measure of technical variability NOT suitable for biological inference
Pooling

e.g. We have 6 biological samples and 3 arrays:

Using 3 pools of 2 samples each
• we reduce the biological variance by 1/6 and the technical variance by 1/3
• we have a measure of biological variation suitable for biological inference
Pooling

But pooling is not always wonderful:

if one of the samples entering the pool is “bad” the entire pool is contaminated.
6. Determine the RNA pool.

a. Standardization

b. Pooling

c. Amplification – Amplification can be used if the quantity of RNA is too small for hybridization. Amplification is supposed to preserve expression ratios for individual genes.

Rule of thumb: If you need to amplify one sample, amplify all.

d. Spiking
6. Determine the RNA pool.

a. Standardization

b. Pooling

c. Amplification

d. Spiking – Foreign RNA may be spiked into the sample for quality control or normalization. If you expect a high degree of differential expression, it is a good idea to have spiking controls to check your normalization. I recommend spiking several controls, each at a different concentration.
7. Determine the hybridization design.

1-channel arrays

2-channel arrays
7. Determine the hybridization design.

1-channel arrays

Usually simple balanced ANOVA layout. (We learn this material in Stat 100.)

Analysis of differential expression proceeds by t-tests or ANOVA or variations such as SAM, LIMMA, Wilcoxon ...
Hybridization design for 1-channel arrays

Typical experiments will be balanced one-way layouts or factorial designs.

Analysis is somewhat more complicated if there are both technical and biological replicates.
Hybridization Design for 2-channel arrays

Many choices for hybridization schemes.

Control design: 1 treatment, 1 control

Reference design: many treatments, 1 control

Loop design: many treatments

Replicated Loops: many treatments

Others: yet another statistical collaborator needed
Control Design

• Purpose: Compare 2 conditions A and C

Both conditions are on every array. Equal numbers have A red and A green (which can be biological replicates).

Analysis is usually done on A-C using parametric or nonparametric “t-test”.

If A and C can be measured on the same biological sample, each biological sample is one array.
Reference Design

• Purpose: Compare many conditions A, B, C, ....

The reference sample is on every array.

Comparisons are among conditions, not reference.

Dye-swap is not needed.

Analysis uses X-R, and proceeds like 1-channel analysis.
Loop Design

• Purpose: Compare many conditions (or combinations, including time course)
  A, B, C, ....
  The dye-swap is between arrays.
  2 samples per condition per loop.
  Analysis is done for each channel.
Replicated Loop Design

- Purpose: Compare many conditions (or combinations, including time course)

A, B, C, ....

\[ \begin{align*}
A & \overset{\rightarrow}{\longrightarrow} B \\
A & \overset{\uparrow}{\longrightarrow} C \\
D & \underset{\leftarrow}{\longleftarrow} C \\
D & \underset{\downarrow}{\longleftarrow} B
\end{align*} \]

Each pair of replicates is in a different loop.

These can be designed to minimize variance of certain comparisons.
The Best Design (2-channel)

Control designs are always fine for 2 condition studies.

Reference designs always have less power than loop designs for the same number of arrays (and platform) but:

are simpler to run and to analyze
E.g. 8 tissues on 16 arrays.

Reference design:
2 arrays per tissue – 2 biological replicates. Reference sample is either genomic DNA or tissue pool.

Loop Design:
2 loops of 8 – 4 biological replicates.
Other Designs

- E.g. Time course with 4 conditions.

```
A → B  A → B  A → B
↑  ↑  ↑
D ← C  D ← C  D ← C
```

Time 1       Time 2       Time 3

The loops do not need to be connected for analysis. This is much simpler than the loop of 12 combinations of condition and time.

Principles of optimal design can be used to pick the loops.
Doing it right.

Plan the statistical analysis at the same time you plan the hybridization design.

The correlation structure induced by the design is critical to the analysis.

Correlation is induced by:
• sample pairing on 2-channel arrays
• repeated measurements on the same organism
• blocking
Handling the Correlation Structure

• We have the most power for detecting differential expression between samples with high correlation – e.g. samples on the same array, from the same individual or the same block

• We build the correlation into our analysis by including the blocking structure and hybridization design in the analysis of the experiment
8. And for custom arrays

Platform

cDNAs or oligos

Choice of genetic material (gene/oligo selection)

Spatial layout on array

Control spots

Number of spot replicates
8. And for custom arrays

Platform – The cost is coming down!

cDNAs or oligos

Choice of genetic material (gene/ oligo selection)

Spatial layout on array

Control spots

Number of spot replicates
8. And for custom arrays

Platform – The cost is coming down!

cDNAs or oligos - ??

Choice of genetic material (gene/ oligo selection)

Spatial layout on array

Control spots

Number of spot replicates
8. And for custom arrays

Platform – The cost is coming down!

cDNAs or oligos - ??

Choice of genetic material (gene/ oligo selection)  
You can do some very fancy experiments

Spatial layout on array

Control spots

Number of spot replicates
8. And for custom arrays

Platform – The cost is coming down!

cDNAs or oligos - ??

Choice of genetic material (gene/ oligo selection)
   You can do some very fancy experiments

Spatial layout on array – at random, separating spot, oligo replicates. DO NOT arrange by EST set.

Control spots

Number of spot replicates
8. And for custom arrays

Platform – The cost is coming down!

cDNAs or oligos - ??

Choice of genetic material (gene/ oligo selection)
You can do some very fancy experiments

Spatial layout on array – at random

Control spots – if used for normalization, should be placed at random

Number of spot replicates
8. And for custom arrays

Platform – The cost is coming down!

cDNAs or oligos - ??

Choice of genetic material (gene/ oligo selection)  
   You can do some very fancy experiments

Spatial layout on array – at random

Control spots – if used for normalization, should be placed at random

Number of spot replicates – 2 to 4 if possible, widely spaced
If you have done all this ...
If you have done all this ...

You are ready to do your microarray experiment.
References

Some examples of nicely done microarray studies
General Principles of Experimental Design
Comparisons of Measurement Platforms
Sample size
Sample pooling
Sample Amplification
Use of spiking controls
Experiment Design for 2-channel Arrays
Questions??
References

Some examples of nicely done microarray studies
• Caldo, Nettleton and Wise (2004) Plant Cell

General Principles of Experimental Design (Textbooks)
• Montgomery (2005) *Design and Analysis of Experiments*
• Kuehl

Comparisons of Measurement Platforms
• Cao et al (2003) Alliance for Cellular Signaling

Sample size
• Black and Doerge (2002) Bioinformatics
• Cui and Churchill (2003) *Methods of microarray data analysis*
References

Sample pooling
• Kendziorski, Zhang and Attie (2003) Biostatistics

Sample Amplification

Use of spiking controls
• Altman (2005) Applied Bioinformatics
• Yang et al (2001) Microarrays: Optical Technologies and Informatics,

Experiment Design for 2-channel Arrays
• Kerr (2003) Biometrics
• Simon, Radmacher and Dobbin (2002) Genetic Epidemiology
• Vinciotti et al (2004) Bioinformatics