

Designing Microarray Experiments

Naomi Altman
Dept. of Statistics
Bioinformatics Consulting Center

naomi@stat.psu.edu

Sept. 9/04

Computational Genomics Journal Club
Penn State

What is a Microarray?

A microarray is a substrate upon which probes, representing one strand of part of a chromosome is printed.

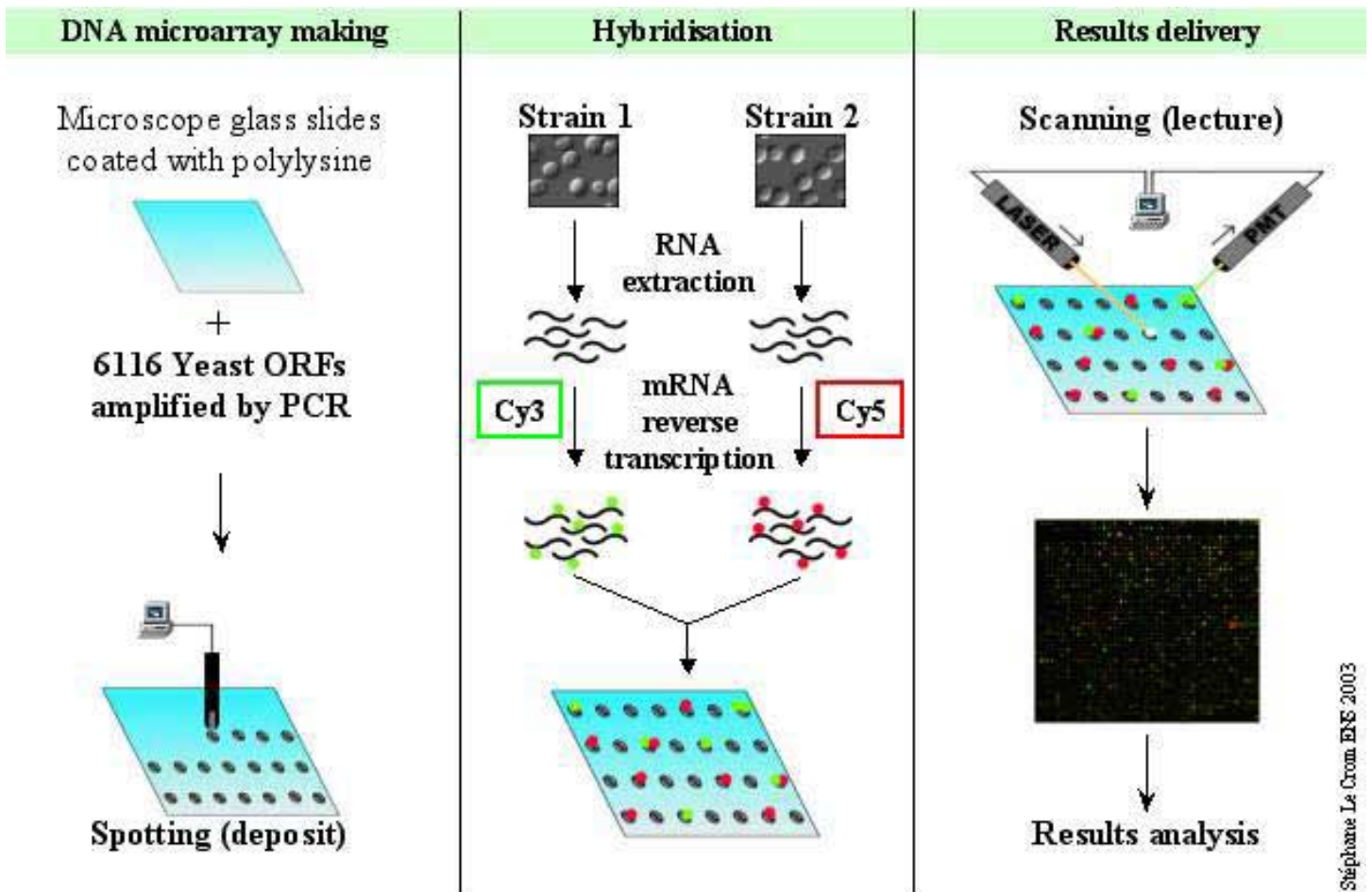
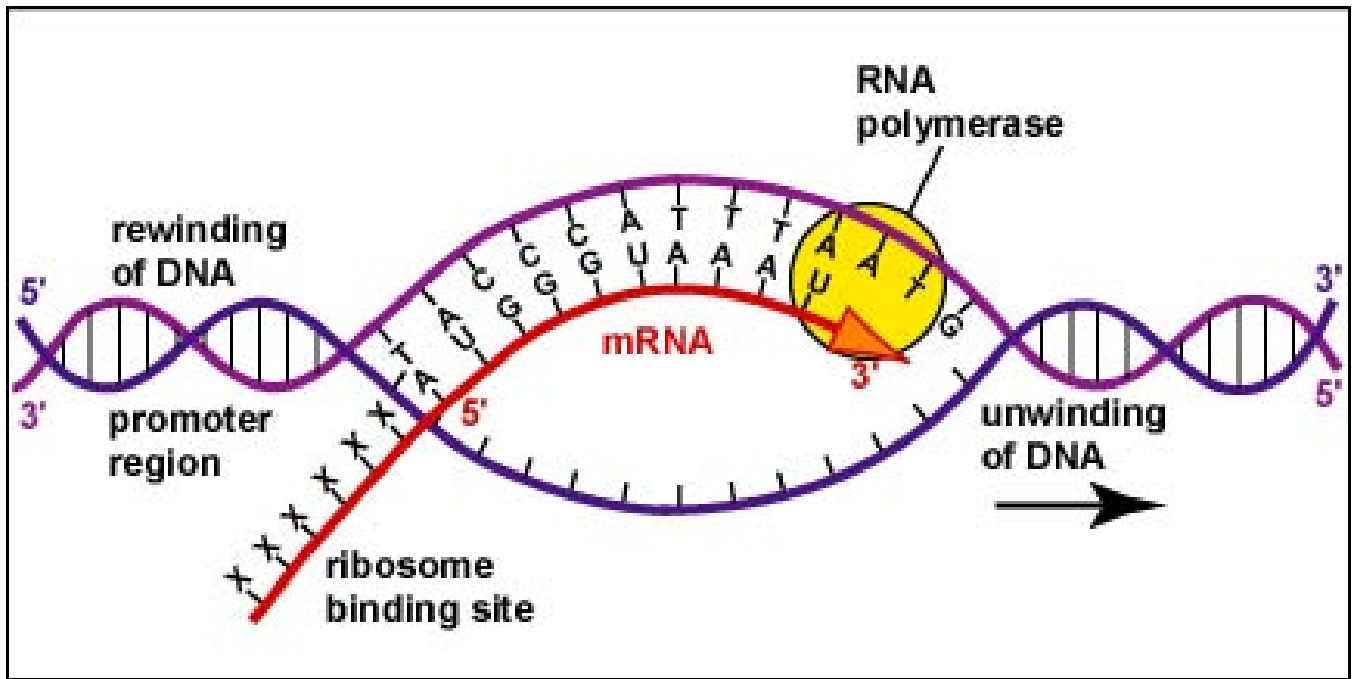
mRNA is removed from a group of cells (usually a tissue or organ or cell colony) converted to a single strand of DNA and labeled.

The labeled material is allowed to hybridize to the probes.

The labeled material will hybridize only to a matching probe.

The amount of label is quantified by using a laser scanner (on fluorescently labeled arrays) or from its radioactivity (on radionucleotide labeled arrays).

Usually, the array surface is divided into pixels. The probe locations are recognized by computer software. The pixels are converted into a probe summary that is called the intensity or expression level.



What are Microarrays Used For?

Current technology allows up to 40,000 or more probes on an array.

This allows us to determine which of the probes express in the samples (down to some detection limit) and to roughly quantify the relative abundance of the probes.

It allows us to determine which probes express differently in different samples.

It is used when we are interested in the expression of large numbers of genes.

Other technologies are available for handling smaller numbers of genes (up to 100 or so) with:

More sensitivity

More accuracy

More detail about location in the organism

Some Problems in Using Microarrays

High measurement error

Cross-hybridization

Differences among RNAs in labeling and hybridization to probes

-
-
-

Microarrays cannot currently be used to compare the actual expression level of different genes in the same sample.

Steps in Designing a Microarray Study

(in approximate order)

1. Determine the objective of your study.

A study to describe the genes expressing during tumor development differs from a study to determine which genes can be used to classify tissues into cancerous versus non-cancerous, which differs from using the microarray to classify a particular tissue sample.

2. Determine the experimental conditions (treatments) under study. (e.g. genotypes, “stimuli” such as hormone treatments or exposure to chemicals, times of measurement)

3. Determine the measurement platform

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

4. Determine the sampling design.

e.g. - multiple tissues on the same individual?

- apply the treatments to different tissues on the same individual or to different individuals?

- hybridize the same RNA sample to more than one microarray?

- pool samples?

- amplify the RNA?

5. Determine the hybridization design (for multi-channel arrays)

Each channel (currently 2) gets one sample.

How will the samples be matched on the arrays.

6. For custom arrays:

Where will you get the probe material?

How should the material be laid out spatially on the array?

And for oligo arrays:

- oligo selection

- spotted or printed on array?

2 Important Principles of Microarray Design

1. Since microarrays are expensive, and there are lots of choices to be made, plan carefully.
2. Since microarray technology is developing incredibly fast, do not order your arrays and hybridization kits until you absolutely must.

And the Most Important Principle
of Microarray Experiment Design

BIOLOGICAL REPLICATION

Selecting Experimental Conditions

Step 1:

Determine the objectives of the experiment.

Step 2:

Suppose there are “no surprises”. What do you expect to happen?

Step 3:

Determine minimum number of factors and “levels” of interest

Step 4:

Usually, if there are 2 or more factors (e.g. genotype and time) a complete factorial arrangement is used.

Additional Step for Time Course Studies

When does the “interesting action” occur?

Are you interested mainly in what happens when the stimulus is applied, or when equilibrium is reached?

Are you primarily interested in determining at what time an event occurs, or how gene expression changes over time?

Determine How the Sampling Will be Done

Replicate, Replicate, Replicate

We are almost always interested in the properties of a biological system, not an individual sample. So, we are looking for features on the arrays that persist across many individuals.

1. You need at least 3 and preferably 5 or more **independent biological samples**.
2. You need to understand the sources of variability in your samples (e.g. – ear drums, mice, litters, labs ...) and how this impacts the inferences from the study.

3. If you are spotting your own arrays, it is useful to have each probe on 2 spots. This reduces loss of biological information if a spot “fails”.
4. If you are relating your microarrays to other studies, you should use the same individuals for both. If you are doing other genomic analyses, plan to reserve some of the RNA.
5. Splitting RNA samples to make 2 or more arrays (or taking multiple samples from 1 individual) is technical replication. It reduces measurement error, greatly increases the complexity of the statistical analysis, and does not change your need for at least 3 biological replicates.
6. Pooling RNA samples is like averaging. It reduces biological variance, but reduces your ability to spot aberrant samples. You still need at least 3 independent pools.

It is better to hybridize 3 samples to 3 arrays then to pool the samples, and then split the same pool onto 3 arrays.

7. RNA samples can be amplified. This may introduce length-specific biases in detection, but may be OK for differential expression (which is based on ratios) as long as the RNAs are copied sufficiently so that they still hybridize to the array. (Kits that amplify 20 samples are in the \$700 range.)

Choice of Platform

The choice of platform is basically one of cost: in \$\$, labor and time.

Often investigators think of cost being the price of the array + reagents.

The actual cost is the price of

- Materials
- Lab time (creating cDNA library, collecting samples, extracting RNA, flagging spots)
- Investigator time (data analysis, redoing studies to include more genes or other methods, writing grants ...)

If you are looking at less than 100 genes, you should not be using microarrays.

If you are using microarrays, are you planning to use other methods as well (e.g. PCR, protein arrays, QTL studies)? If so, plan your experiments to integrate this information.

Choosing A Platform

Affymetrix Arrays

Arrays and reagents are expensive (\$700/array).

Gene density is high (22000 genes or more).

Informatics are available from Affy and other sources and appear to be pretty good.

Failure rate is low.

Replicability (technical precision) is high.

Statistical analysis is fairly simple.

Available for 19 “model species”.

Custom arrays can be made (at very high price, but this is coming down.)

cDNA 2-color Arrays

These are spotted from a cDNA library.

The cost of creating the library is high.

The cost of printing is low. (Labs are often willing to “share” at \$200/slide or less – but that is not the “true cost” even if they are free.)

The per slide or per gene cost can be higher than expected because failure of array or individual spots is frequent. (You should plan for failure by having extra RNA.)

Typical arrays have 3000 – 25000 spots.

Within array replicability is usually high (i.e. duplicates of the probes) but the technical variability between arrays may be high.

There are 2 samples hybridized on each array. This makes for some interesting statistical design issues – you can use simple but inefficient designs (meaning you need more arrays) or more efficient designs (like “loops”) which are harder to analyze but use 50% less arrays.

Comparison of different genes on the same array is problematic, because the cDNA probes have different hybridization efficiencies.

Dye-swap is essential because some cDNAs have more affinity for Cy-5 and some for Cy-3. Dye-swapping can (and should) be done using different biological samples (so you need an even number of replicates).

Oligo 2-color Arrays

- Oligos (usually 50 or 70 “mers”) are fabricated from the known gene sequence.
- These may be fabricated directly on the array (Agilent) or used like a cDNA library for spotting.
- It may be cheaper to fabricate oligos from known sequence than to create or purchase a cDNA library.
- Oligos can be selected for uniform hybridization characteristics, minimal cross-hybridization, selection of gene families ... The companies that fabricate oligos usually offer the service to do this.
- Commercial oligo arrays are supposed to be very replicable.

- Locally produced oligo arrays are supposed to be more replicable than locally produced cDNA arrays.
- All other caveats about spotted arrays apply.
- Commercial custom arrays are comparable in cost to spotting locally for up to about 40 arrays.
- We have found some commercial suppliers are willing to “bundle” to form a bulk order – e.g. arrays from 2 projects, even different species, count as 1 bulk order.
- As for other platforms, prices keep changing.

Radionucleotide Arrays

These are usually either spotted cDNA arrays, or on some substrates you can spot the e. coli containing the library.

They are cheaper and noisier than 2-color spotted arrays, but you can put only one sample on an array.

In some systems, the samples can be “stripped” and the array can be reused with another sample.

Analysis (following spot summary) is similar to 2-color arrays with reference sample.

There are still issues regarding spot summaries. High intensity spots may overlap other spots.

Hybridization Design for 2-color systems

Since you can put 2 samples on an array, you need to worry about how to best pair the samples.

The most common design is the reference design, which puts the same control sample on every array. Dye swap is not necessary and statistical analysis is simple, but ... you need twice as many arrays as in designs that use both channels.

There are lots of possibilities for designing more efficient hybridization pairs.

Get some help before you start – it is easy to make a mistake that will be much more expensive than a couple of hours of BCC time!

Custom Spotted Arrays from Start to Finish

Steps to making a custom array:

1. Assemble the genes:

This requires either a cDNA library or sequence information to create oligos. (We have lots of experience with this on campus.)

2. Assemble the probe material (either cDNAs or oligos) for spotting.

3. Determine cDNAs or oligos to be used as foreign controls. (Genes from another species or a kit.)

4. Determine the spot design including layout and replicate spotting (2 or 4 is usual).

Spot Design and Layout for Statistical Analysis

Normalization for 2-color arrays is based on 2 assumptions:

1. most genes do not differentially express
2. genes are randomly distributed on the array surface

In unsequenced organisms, information for the probes is likely to come from expression studies, which may be specific to tissues or conditions. So many of the genes may differentially express. It is important to keep the coverage of the array broad in the sense that genes come from many different expression studies. It is important to include “house-keeping” genes on the array.

Genes which are known to express some condition should not be clustered on the array.

Resources on Campus

DNA Microarray Facility (Craig Praul)

<http://hils.psu.edu/stf/dnama/home.html>

design of microarrays and microarray studies.

Spotting arrays

Labeling and Hybridization

Scanning

Center for Computational Genomics (Izabela Makalowska)

<http://www.cbio.psu.edu>

bioinformatics searches

development of genomics databases

oligo design

Bioinformatics Consulting Center (Naomi Altman, Gary Chase, Wenlei Liu)

<http://bcc.cbio.psu.edu>

Design of microarray experiments
Hybridization design
Normalization of microarray data
Statistical analysis of microarray data

LionDB Project (Istvan Albert)

<http://www.liondb.org/index.html>

Storage of genomics data
Web-based analysis of microarray data

Remember: The cost of one “bad” array pays for about 5 hours of facility time. So consultation is cost-effective.