Microarrays
The technology
Goal:

Goal: To measure the amount of a specific (known) DNA molecule in parallel.

“In parallel”: do this for thousands or millions of molecules simultaneously.

Main components of microarrays:

1. Hybridization
2. The ability to attach a known DNA molecule to a solid surface.

Expression: To measure RNA

1. Convert RNA to DNA
2. Measure DNA

This is a common trick: to measure X in a cell, convert X to DNA and measure DNA.
We try to measure a specific (short) DNA sequence called “the target”, which we measure using “a feature”.

1. Take a known DNA sequence and attach many single stranded copies in a specific place on a solid surface. This is a “feature”.

2. In a sample we have an known number of copies of this DNA sequence (“the target”). Label of all these copied (attach a fluorescent molecule).

3. Put the labelled sample onto the feature. Each labelled copy of the target will hybridize to the feature.

4. Measure how many copies have hybridized by measuring the amount of light from the feature.
Two main types of microarrays

- One color
- Two color
Before Labeling

Sample 1

Sample 2

Array 1

Array 2

Each array has 4 features
Before Hybridization: One Channel

Sample 1

Sample 2

Array 1

Array 2
After Hybridization

Array 1
Array 2
Quantification

Array 1

Array 2
Microarray Image
Nomenclature

Target: the molecule we want to measure.

Probes: the molecules we use to measure the target.

Feature: contiguous region on an array with the same probes.

In common usage we ignore that each feature has multiple copies of the probe, and just use “the probe of the feature”.

We distinguish probe and feature because one array can have multiple features with the same probe.
Feature level data

After scanning, image analysis software produces one (two) measure(s) for each feature.

These measures are usually stored as one (or two) files per array. Common formats are CEL (Affymetrix), IDAT (Illumina), XYS (Nimblegen), GPR (Genepix). Sometimes binary, sometimes text.

This is usually where we start. It is uncommon to do the image analysis, although there is historical work on this.
Image Analysis

Addressing

Segmentation

Data Extraction

http://www.epibiostat.ucsf.edu/biostat/cbmb/bmi209.fall05/lectures/bmi209fall05-lecture2.pdf
A complex measurement process lies between mRNA concentrations and intensities

- RNA degradation
- amplification efficiency
- reverse transcription efficiency
- hybridization efficiency and specificity
- labeling efficiency
- quality of actual probe sequences (vs intended)
- scratches and spatial gradients on the array
- cross-talk across features
- cross-hybridisation
- optical noise
- image segmentation
- signal quantification
- signal "preprocessing"
A complex measurement process lies between mRNA concentrations and intensities

- RNA degradation
- quality of actual probe sequences
- image segmentation
- amplification efficiency
- scratches and spatial gradients on the array
- signal quantification
- reverse transcription efficiency
- cross-talk across features
- signal "preprocessing"
- hybridization efficiency and specificity
- cross-hybridisation
- labeling efficiency
- optical noise

The problem is less that these steps are ‘not perfect’; it is that they vary from array to array, experiment to experiment.
Types of microarrays

cDNA spotted microarrays: “homemade”. Pretty much extinct. Poor manufacturing quality; source of much interesting very problematic data.

Affymetrix: Uses 25mers, but use multiple different probes (a “probeset” to measure a single target. Great manufacturing quality using semi-conductor technology. Very expensive to make a new array.

Agilent: Uses 50-60mer. Uses print-jet printing technology. Cheap to manufacture a new array design. Typically one probe per target.

Illumina bead arrays: around 50-60mers. Uses interesting chemistry to improve the hybridization signal and contains a (more) random number of probes per feature.
Oligonucleotide microarrays

1.28 cm

Actual size of GeneChip® array

500,000 locations on each GeneChip® array

Millions of DNA strands built up in each location

Actual strand = 25 base pairs
spotting pin quality decline
after delivery of $5 \times 10^5$ spots

after delivery of $3 \times 10^5$ spots

H. Sueltmann DKFZ/MGA
Probes and Probesets

Typically 11 probe(pairs) in a probeset
Latest GeneChips have as many as:
  54,000 probesets
  1.3 Million probes
Two Probe Types

Reference Sequence

TAGGTCTGTATGACAGACACAAAGAAGATG

CAGACATAGTGT^CTGTGTTTCTTCT

CAGACATAGTGT^GTGTGTTTCTTCT

PM: the Perfect Match

MM: the Mismatch

No-one uses these anymore

Idea: Use PM-MM

Issue: $E(X-Y) = E(X) - E(Y)$

but $V(X-Y) = V(X) + V(Y)$