

# Microarrays

Miler T. Lee

21 January 2005  
revised 19 February 2005

## 1 Biology

1. Microarrays are like large-scale reverse Northern blots
  - Northern: RNA is run on a gel, labeled DNA is hybridized onto gel
  - Microarray: DNA is spotted onto the array, labeled RNA is hybridized onto array
2. Nomenclature
  - **probe** = DNA that is spotted into each well in the array
  - **target** = labeled RNA (or DNA) that is being analyzed by hybridizing to the array
3. Spotted array
  - PCR of cDNA clones to get oligos, which are robotically spotted onto the array – can't control amount that is being spotted in each well
  - Denature to make single stranded DNA
  - Label 2 mRNA samples - experimental and control (reference) or 2 experimental conditions - with different dyes and cohybridize
  - Labeling: **direct** (one dye molecule per nucleotide; fluorescence increases proportional to length) or **indirect** (one fluoro per oligomer)
  - typically, Cy3 = experimental sample (green), Cy5 = reference sample (red)
  - Expression levels measured as ratio between the two dye intensities (figure); if it looks green, upregulation with respect to baseline reference; if it looks red, downregulation with respect to baseline; yellow, no change
4. Affymetrix array (photolithographic)
  - Probes are 25 mers, usually 8-20 sets per gene; typically extracted from 3' UTR
  - Probes are sequenced directly onto the chip:
  - for a given probe, the 3' OH is protected with a large functional group, preventing addition of a nucleotide
  - at each cycle, the nascent probe is either masked or exposed to light
    - if masked, no addition of a nt
    - if exposed to light, 3' OH becomes deprotected, so sequence can be extended by exactly one nucleotide by adding protected nucleotides to the mix
  - repeat cycle 4x25 times
  - Presence/absence calls per gene depend on comparing intensity of match probes with intensities of mismatch probes - differ by one nucleotide in the middle of the probe; presumably strong binding to the mismatch probe indicates non-specific binding

	<b>Spotted</b>	<b>Affy</b>
probe size	40-60 nt	25 nt
cost	less	expensive
co-hybridization	yes b/c of the variable amt of DNA on each spot	no need b/c DNA is synthesized directly on chip in a controlled fashion
reagents	need cDNA clones to PCR, or can work off of sequence	sequence is required
sensitivity	slightly less	slightly more

Table 1: Summary of Spotted vs Affy arrays

#### 5. Contrast with **SAGE** - serial analysis of gene expression

- from cDNAs from cell sample, extract sequence tags (10-14 bp), which are presumably sufficient to identify the source gene
- form concatemers by linking many tags together
- sequence the concatemers and count up the number of times each tag appears - directly proportional to the amount of RNA for the corresponding gene present in the original sample
- unlike microarrays, does not require explicit knowledge of the sequences expected (i.e., what to spot on the array)

## 2 Applications

### 1. Gene expression profiling

- Look for different upregulated / downregulated genes in different conditions, e.g., tumor vs non-tumor
- Cluster genes with similar expression profiles and look for statistically significant overrepresentation of particular classes of genes (using, e.g., Gene Ontology terms) within the clusters as compared to random

### 2. Genome-wide location analysis

- **ChIP on chip** = chromatin immunoprecipitation on chip, assays which DNA sequences are bound by proteins, e.g. transcription factors

### 3. Array-based comparative genomic hybridization (array-CGH)

- Detect changes in copy number of chromosomal regions characteristic e.g. of tumor cells (does not detect transposition)
- Genomic DNA clones rather than cDNA clones are spotted onto the array

### 4. Tissue microarray - tissues spotted on the array

### 5. SNP genotyping

## 3 Analysis

### 1. Normalization

- linear scaling (normalize expression levels across conditions based on median or median and variance)
- with respect to housekeeping genes, which should have the same expression levels across conditions

- LOESS

2. Testing significance of overrepresentation of a gene class in a cluster - Fisher Exact Test

3. t-test, FDR, FWER