Introduction to Microarrays

Lecture outline

- What are microarrays?
- Affymetrix oligonucleotide microarray
  - Technology
  - Analysis
- What else can be done with micorarrays
  - Alternative splicing
  - ChIP-chip
  - Highthroughput genotyping & copy number analysis

What are microarrays?

- Miniaturized high-throughput experiments on glass slides
- Based on principles of hybridization
- Many types
    - DNA microarrays
    - Protein microarrays
    - Chemical compound microarrays
    - Tissue microarrays

This lecture will focus on DNA microarrays only

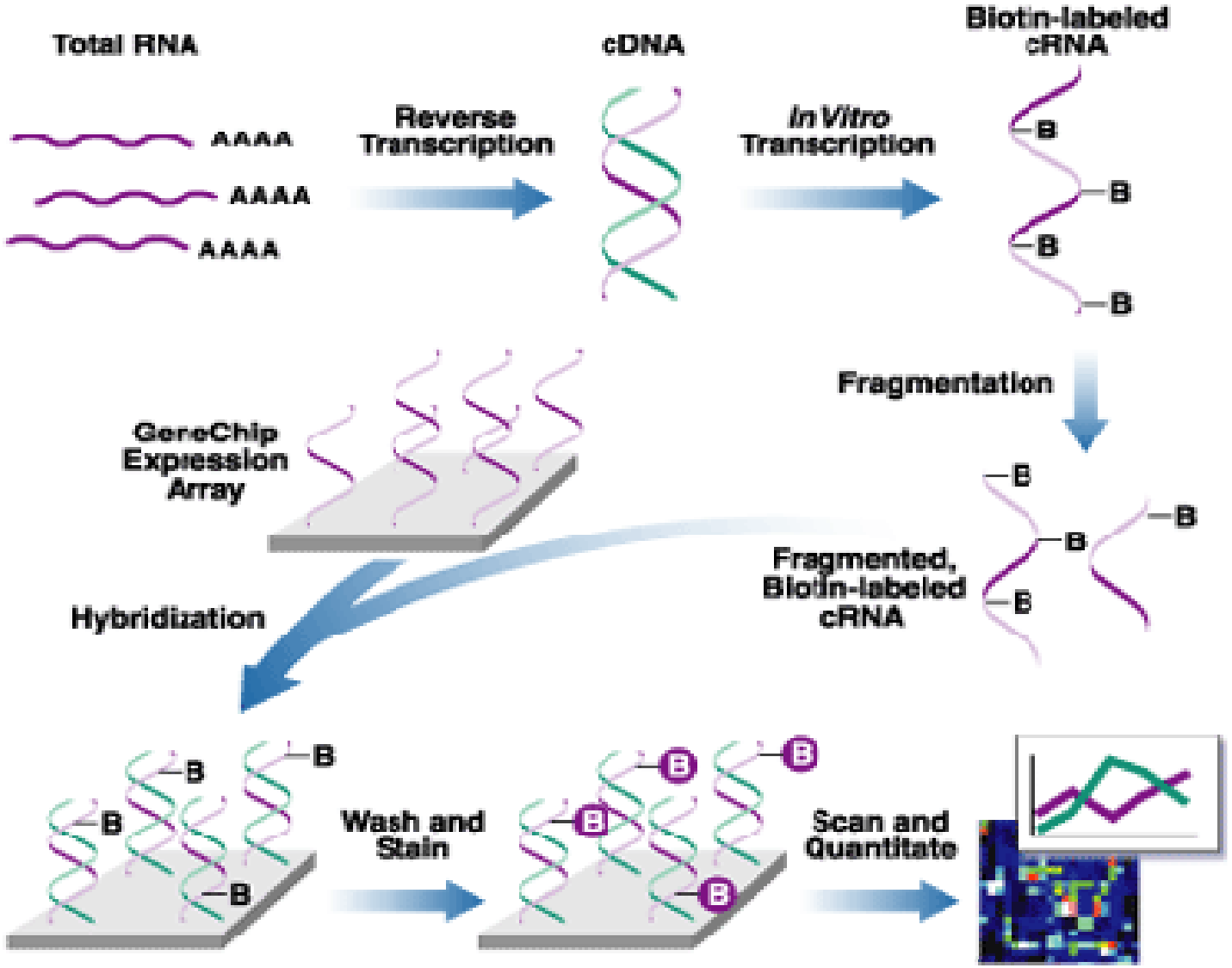
What are DNA microarrays?

- Originated in Stanford in early 1990's (Schena *et al.*)
- Analagous to Southern Blotting
- Usually glass slides on which there are many thousands of 'spots' consisting of DNA 'probes'
- A dye-labelled cDNA/cRNA sample is washed over the slide
- Labelled cDNA/cRNA hybridizes to the probes
- Hybridization is detected using lasers
    - Laser causes die to fluoresce
    - Fluorescence signal measured
- Depending on the type of array may require one or two dyes

Affymetrix oligonucleotide microarray as an example

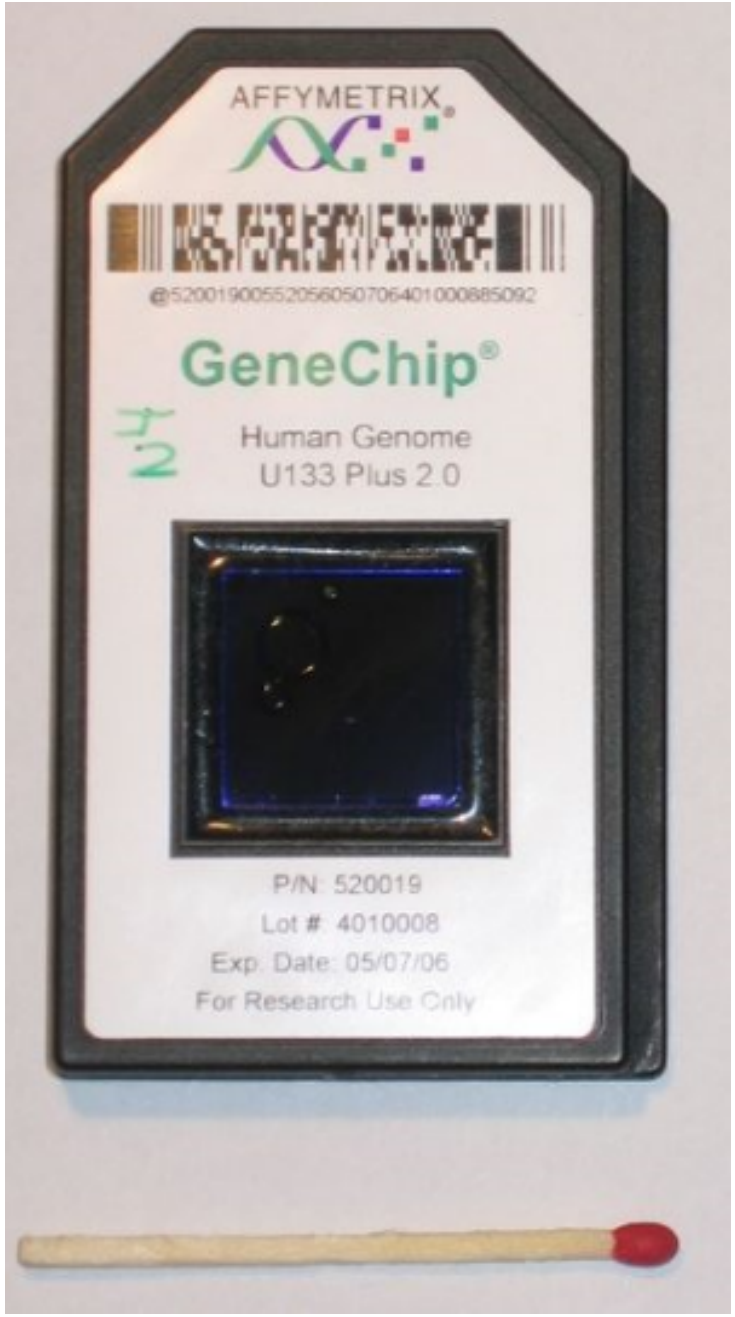
- DNA oligonucleotides synthesized on the array
- Probes usually around 25bp
- Single colour (i.e. RNA labelled with one dye)
- Originally targeted against 3' gene ends
- Latest versions target along the length of the transcript (e.g. Exon array and Whole transcript array)
- Collections of probes (called probesets) target features (e.g. gene or exon)

Affymetrix oligonucleotide microarray technology



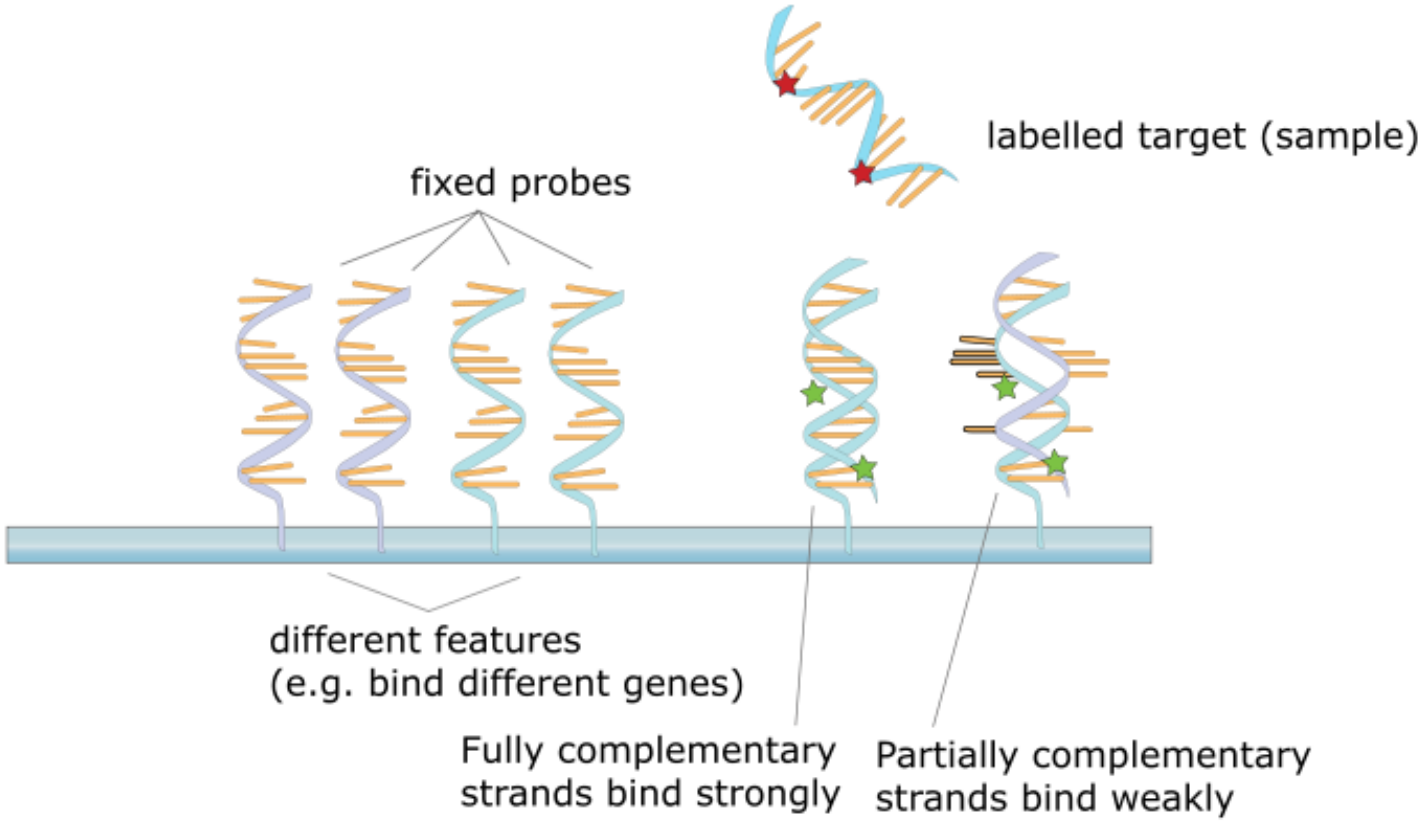
From a presentation by Gareth Palidwor

Example of a Gene Chip from Affymetrix



The basis of microarrays: hybridization

- RNA hybridizes to complementary (or partially complementary DNA)
- Microarrays consist of an array of ‘spots’ containing DNA
- Each spot has DNA from a different target (e.g. gene)
- RNA which has been labelled with a dye is washed over the array
- RNA from a given gene hybridizes to the DNA at the corresponding spot on the array
- The amount of RNA hybridized to the spot expected to be proportional to the amount of RNA in the sample
- This can be detected from fluorescence of the dye when stimulated with a laser


labelled target (sample)
fixed probes
different features
(e.g. bind different genes)
Fully complementary
strands bind strongly
Partially complementary
strands bind weakly

Fundamentals of microarray analysis: Scanning

- Lasers cause dye bound to the RNA in the sample to fluoresce
- Fluorescence captured
- A fluorescence value can now be associated with each spot on the microarray

Fundamentals of microarray analysis: Quality control

- Microarrays sensitive to several artefacts
- Within array artefacts can be visualized by plotting an image of the array
- Descriptive statistics can be used to look for evidence of systematic differences between arrays (e.g. boxplots can reveal differences in overall brightness between arrays)
- The goal of microarray preprocessing is to correct these artefacts (to the extent possible)

Fundamentals of microarray analysis: Preprocessing: Background subtraction

- There may be some fluorescence in between the spots on the array (this needs to be subtracted from the signal coming from the spots).
- There are various approaches to background subtraction (see later)

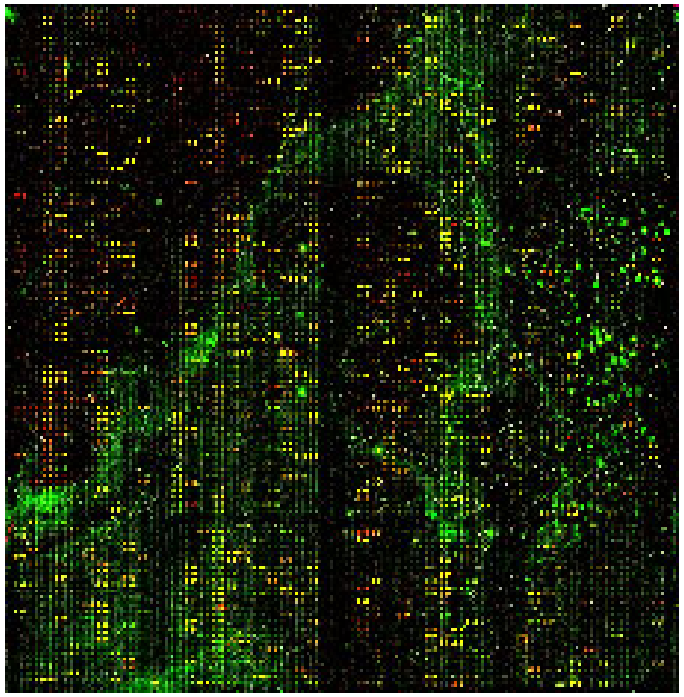


Image source: http://www.systemsbiology.nl/datgen/transcriptomics/preprocessing/background.html

Fundamentals of microarray analysis: Preprocessing: Normalization

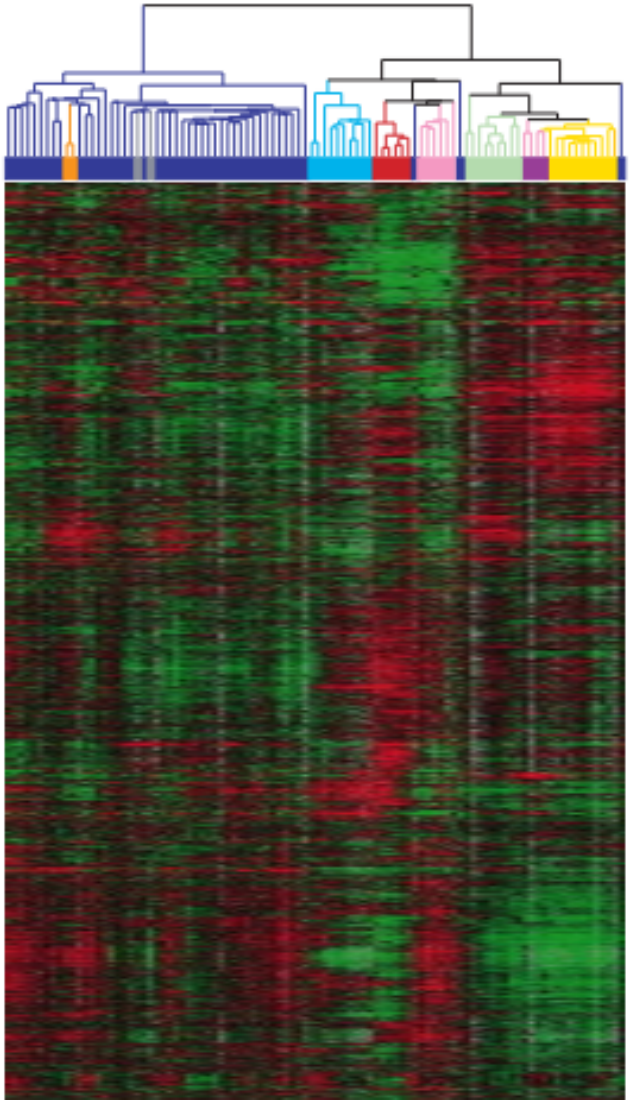
- There will be systematic differences between different slides – e.g. different amounts of sample, different quality of RNA, different temperature at which the array was run
- These need to be corrected for, in a process called normalization
- Methods:
    - Housekeeping genes – whose expression is relatively constant
    - Spiked in mRNAs at a known concentration
    - Quantile normalization: Manipulate the intensity distributions so that they are matched across experiments (see later)

Fundamentals of microarray analysis: Unsupervised clustering

- A first step in microarray analysis is often to use unsupervised clustering techniques to identify groups/clusters of samples
- This can reveal problematic/outlier samples that have not been corrected in preprocessing steps
- Unsupervised clustering can illustrate how the variability induced by the experimental groups compares to other sources of sample variability
- Examples of unsupervised clustering techniques: Principal Component Analysis (a standard dimensionality reduction technique); Hierarchical clustering.

Fundamentals of microarray analysis: Finding patterns in the data

- Unsupervised clustering (are there genes with similar patterns of expression across samples, or samples with similar genes expressed)

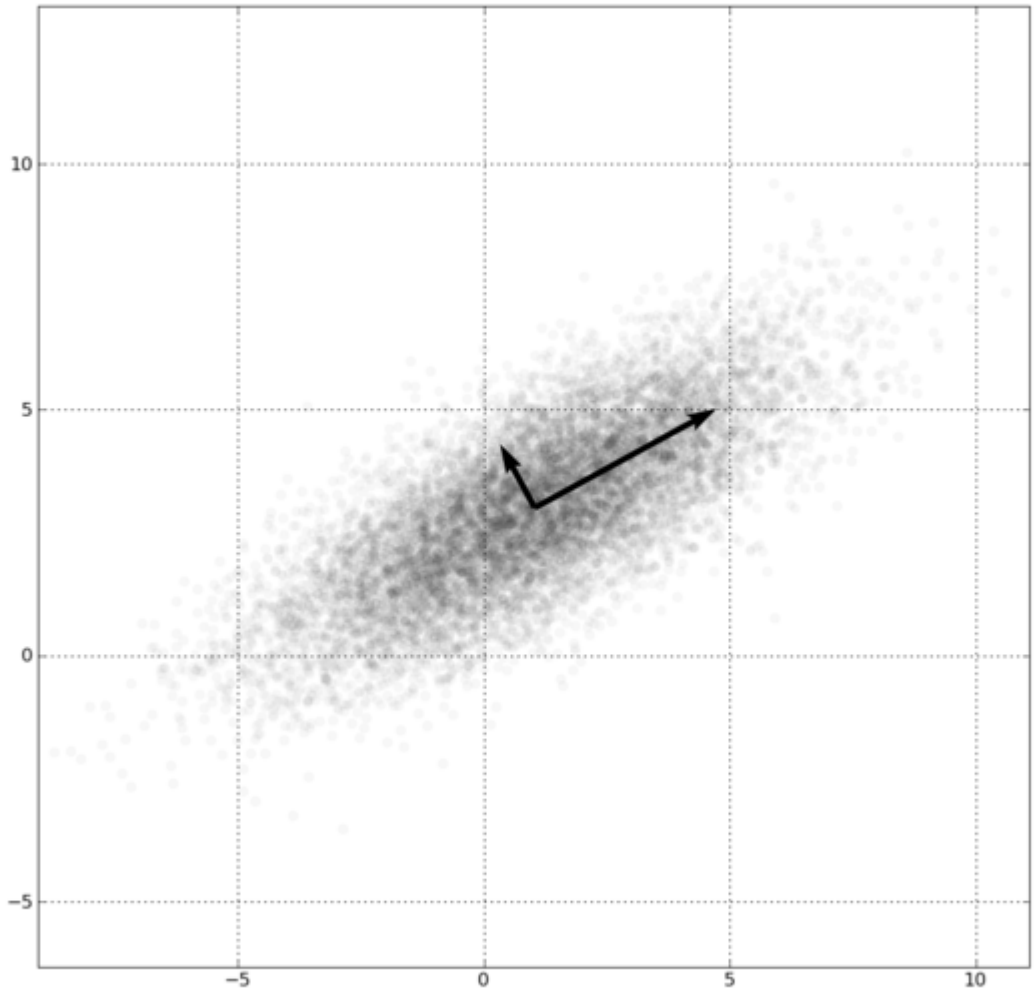


Alizadeh *et al.* Nature 2000

Principal components analysis

- ‘Dimensionality reduction technique’
- Used to visualize high-dimensional data
- Allows high dimensional points to be plotted in a lower dimensional space (e.g. 2D) in a way that separates the points well (by focusing on the directions of greatest variability)
- Technically: the principal components are the directions of greatest variability in the n-dimensional cluster of points

Principal components



Source: Ben FrantzDale

Fundamentals of microarray analysis: Gene expression analysis

- Differential expression analysis (find all the genes that are differentially expressed, e.g. between drug-treated and untreated samples)
- Finding patterns – supervised/unsupervised learning (e.g. find the patterns that discriminate between different types of tumours with different prognoses)

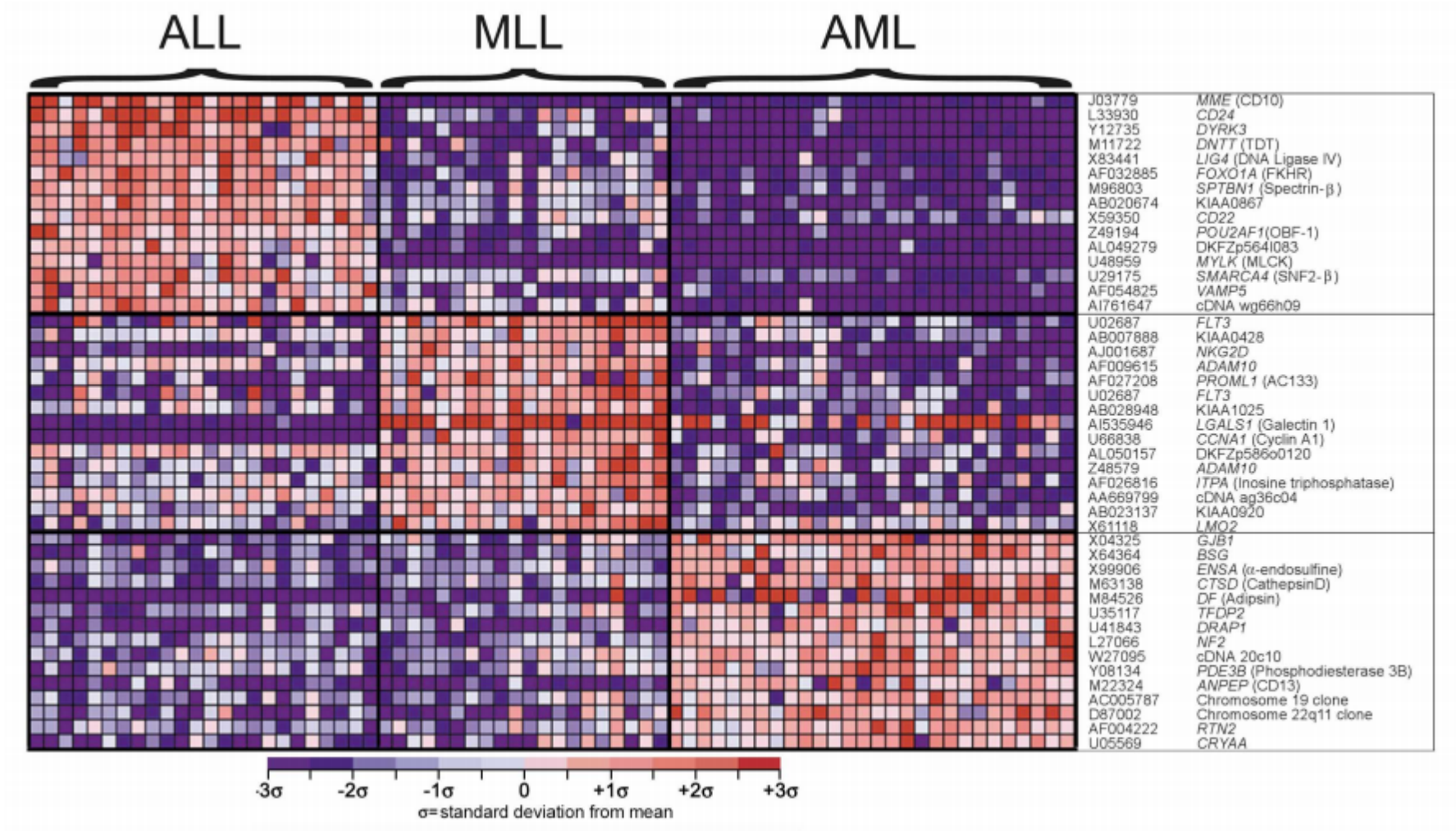
Fundamentals of microarray analysis: Differential expression analysis

- Can use standard statistical tests (e.g. t-test)
- But necessary to correct for multiple hypothesis testing

Fundamentals of microarray analysis: Classification

- Supervised learning: e.g. Armstrong *et al.* 2001
  - Acute lymphoblastic leukemias (ALL) involving translocation in mixed-lineage leukemia (MLL) gene associated with poor prognosis
  - Considered three types of leukemia: MLL (as above); other ALLs and Acute Myelogenous Leukemia (AML)
  - Were able to classify patients accurately on the basis of gene expression (a viable alternative to existing tests)
  - Provides clues about the gene expression 'programs' associated with each leukemia type and why prognosis differs.

Expression patterns of the subset of genes that were used to classify leukemia samples



Armstrong et al. *Nature Genetics, 2001*

Applications of microarrays

- Measure gene expression level

(recall DNA -> mRNA -> Protein)

- Measure alternative splice isoforms
- Molecular diagnostics
- Measure binding of proteins to DNA (ChIP-Chip)
- High-throughput genotyping

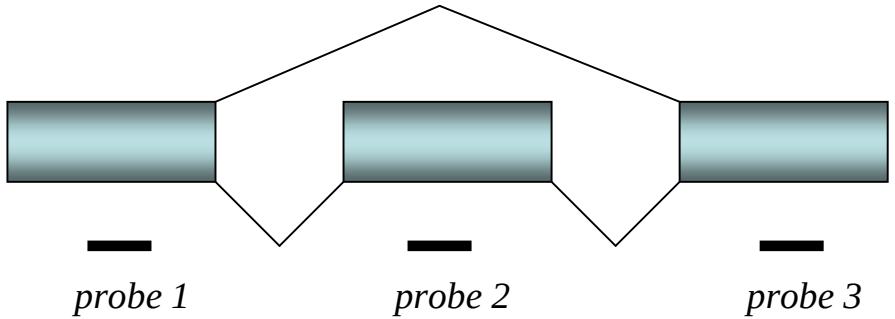
Applications of microarrays: measuring gene expression

- Regulation of gene expression is critical for gene function
- Microarrays provide a means of measuring the expression level of all genes simultaneously
- Comparing gene expression in different samples (e.g. drug-treated and untreated cells) provides information about the biological differences between the samples (e.g. the effects of the drug on the cell)
- Because they are high-throughput microarrays provide information not just on individual genes but on whole sets of genes (e.g. in a biological pathway) that distinguish between samples
- Differences in gene expression between individuals, is implicated in many human diseases and complex phenotypes

Applications of microarrays: Alternative splicing

• The majority of multi-exon human genes produce more than one transcript through alternative splicing

• Expression levels of different transcripts can be measured by comparing the ‘signal’ from different parts of the gene



• Relative expression of different transcript isoforms can be important for biological processes and involved in human diseases and phenotypes

Applications of microarrays: Molecular diagnostics

- E.g. different cancers associated with different ‘molecular signatures’ (gene expression profiles)
- Can be used to type a cancer and evaluate prognosis and guide treatment
- Microarrays have been proposed for use in diagnosis of infectious diseases (microarrays with probes for different pathogen genes)

Applications of microarrays: ChIP-Chip

- Used to find out where proteins (e.g. transcription factors) are bound to DNA
- Helps to identify functional transcription factor binding sites
- DNA fractionated; Antibodies against the protein of interest used to ‘pull down’ DNA fragments to which proteins bound; microarrays used to measure the amount of each DNA fragment that has been pulled down (i.e. the extent to which this DNA fragment is bound by the protein in the sample)

Applications of microarrays: High-throughput genotyping

- Microarray contains probes corresponding to both alleles of each human SNP (e.g. Affymetrix million SNP chip)
- The relative intensity of the signals allows SNPs to be genotyped
- Has application in current wave of huge genome-wide association studies
- Is the basis of personalized medicine