# Lecture #1

Introduction to microarray technology

# Outline

- General purpose
- Microarray assay concept
- Basic microarray experimental process
- cDNA/two channel arrays
- Oligonucleotide arrays
- Exon arrays
- Comparing different array technologies
- Microarray community standards
- Interpreting the biology correctly

# Purpose of microarrays

- Understand the processes and associations both within and between genes (functional genomics)
  - Genetic diseases (many disorders are multifactorial)
    - Regulation patterns
    - Splice variants (exon splicing events)
    - Single nucleotide polymorphisms (SNPs)
  - Pathogens
  - Drug discovery
  - Personalized medicine (pharmacogenomics)

- Gene interactions are complex in nature, such that it is necessary to assay many simultaneously to understand dependencies in expression patterns

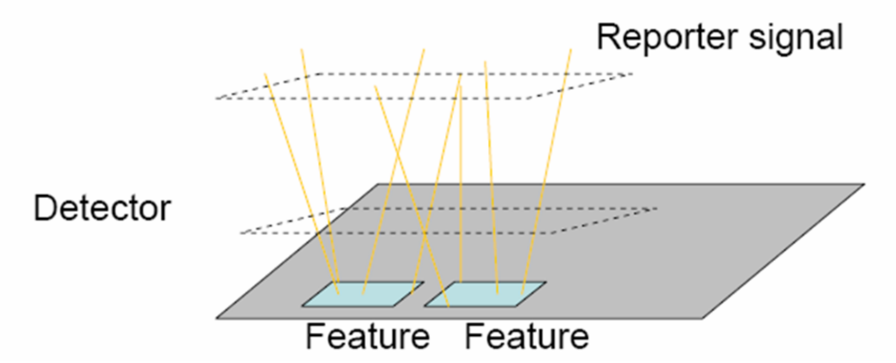
- Requires high-throughput technology

# Some definitions

- Microarray: chip, genechip, array
- Transcript: message, mRNA
- Gene expression: number of mRNA molecules
- Signal intensity: measure of the number of mRNAs present in a sample
- Probe: DNA sequence (usually) that is fixed to the microarray and used to measure the target; also known as ‘reporter’
- Target: sequence from sample that is labeled with some dye method
- Sample: patient/subject, animal, cell, array
- Feature: the (x,y) location of probes on an array

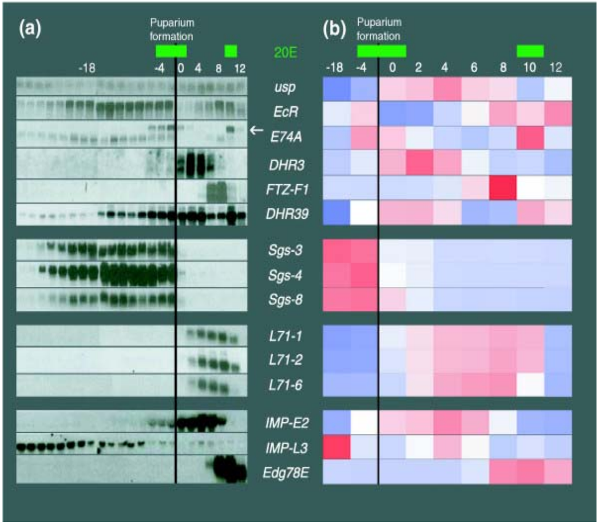
# What is a microarray?

- At the most general level, a microarray is a flat surface on which one molecule interacts with another
  - Location and signal produced provide characteristics about the interacting partners
  - Probe: relatively short sequence of DNA fixed to the surface
  - Target: the sequence entity that we are measuring (typically labeled)



# Microarray concept

- Quantitative measure of mRNA
  - Since most changes in cell states are associated with mRNA
- Similar to a Blot
  - Intensity-based measurement generated from some label is used to represent amount of material present within a sample



*http://genomebiology.com/content/figures*

# Basic experimental process

http://www1.istockphoto.com/

1) Select or design an appropriate microarray

2) Collect sample, extract and purify the target (DNA or RNA)

3) Label the target with some dye and fragment

4) Hybridize the target to the array; target sequences specific to the probes hybridize

5) Remove non-hybridized material using some wash condition

6) Laser is applied to labeled target (in a duplex with the probe) which fluoresces at a specific wavelength (dependent upon the dye used), proving a quantitative intensity value

7) Collect the signal, process, normalize, and proceed to analysis

***The signal intensity is assumed to correlate with the amount of mRNA in the sample***

http://www.bio.davidson.edu/courses/genomics/chip/chip.html

# Where are microarrays used?

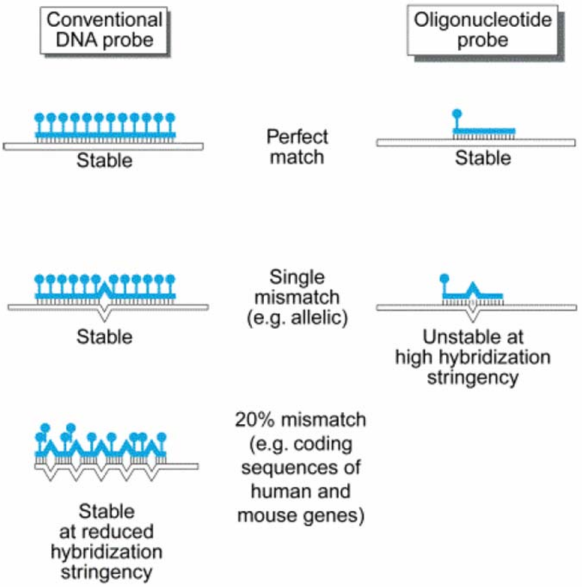
- Anywhere to better understand the associations both within and between genes (functional genomics)
  - Biomarker discovery: identify genes that are differentially regulated in a disease state
  - Pathogen detection: one of multiple methods currently used by DHS (and other agencies) to detect the presence of certain pathogens (‘Biosensing’)
  - Drug discovery: identify genes that are 1) regulated in response to drug treatment (pharmacodyamic markers), 2) predictive of a patient’s prognosis, 3) diagnostic for a patient’s condition
  - Genetics: identify large chromosomal or loci deletions or amplifications (i.e. copy number variations)
  - Many more purposes…

# Probe selection for array design

- I won’t go into too many details here, but probes are selected based on numerous characteristics
  - Specificity to target sequence
  - Melting temperature ($T_m$)
  - Propensity to not form secondary structure (free energy of folding)
  - 3’ bias for complement to target sequence (tends to be more unique)
  - Length and uniqueness
  - Minimal homology to other targets in a gene family

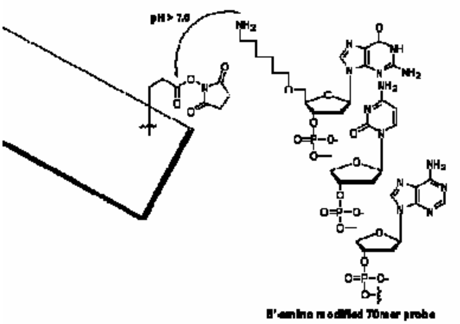
# Two primary microarray designs

- cDNA/two channel arrays
  - 50-70 bp probes
  - Two-color hybridization used for each probe
  - Poor specificity due to mismatch tolerance
- Oligonucleotide arrays
  - 8-60 bp probes
  - Single color hybridization used for each probe
  - Good specificity but poor sensitivity


Conventional DNA probe
Oligonucleotide probe
Stable
Perfect match
Stable
Stable
Single mismatch (e.g. allelic)
Unstable at high hybridization stringency
Stable at reduced hybridization stringency
20% mismatch (e.g. coding sequences of human and mouse genes)

# cDNA/Two channel Arrays

- cDNA/two channel arrays
  - Less expensive technology
  - Complete sequence is attached to chip (probe)
  - Two-color hybridization used for each probe
    - Internal control
    - cDNA is PCR’d with random 6mer primers and dCTP-dye conjugates
      - **Cy5** abs=650 nm; emm=667 nm
      - **Cy3** abs=552 nm; emm=568 nm
- cDNA/two channel array probe spotting
  - Utilizes spotting technology to attach probes to chip (printer robots)
  - A solution is picked up with a pin and this touches the array surface, depositing the probe solution



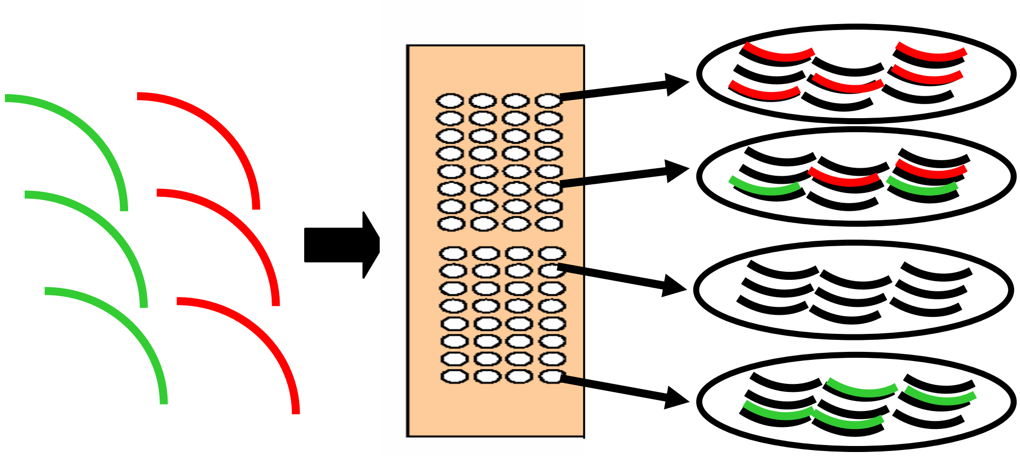
http://pga.mgh.harvard.edu/Parabiosys/images/figs/coupling.gif

# cDNA Technology

A glass slide, membrane, or polymer that has been spotted with non-labeled DNA probes designed to hybridize specific complementary DNA's of interest (cDNA).

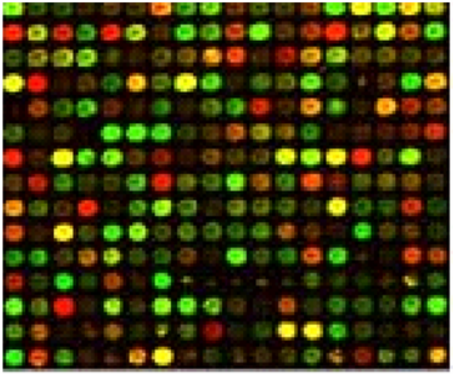
# cDNA Technology (cont.)

Samples are prepared from both an experimental sample, (e.g., malignant tumor) and a control sample, (e.g., normal tissue) and are then overlaid on the array and allowed to hybridize to each spotted probe:



# cDNA Technology (cont.)

Following hybridization, the array is scanned and the resulting gene expression information for each spotted probe on the array is reported



A green intensity = control Only expressed gene
A red intensity = experiment Only expressed gene
A mixed color = gene expressed in both control & experiment

# cDNA Technology (cont.)

The gene expression information that is actually reported are ratios of the amount of experiment sample to control sample that has hybridized to each spotted probe on the array

$$\text{Probe 1}\left[\frac{\text{Amount of experiment sample hybridized}}{\text{Amount of control sample hybridized}}\right] = \text{Probe 1 Expression Ratio}$$

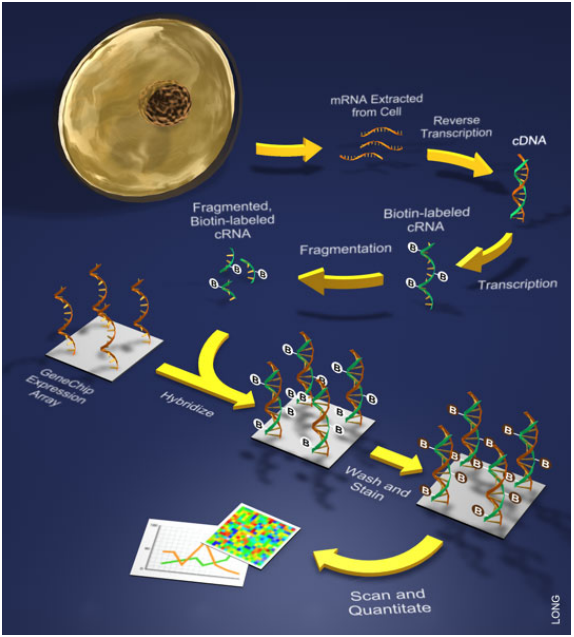
$$\text{Probe 2}\left[\frac{\text{Amount of experiment sample hybridized}}{\text{Amount of control sample hybridized}}\right] = \text{Probe 2 Expression Ratio}$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\text{Probe n}\left[\frac{\text{Amount of experiment sample hybridized}}{\text{Amount of control sample hybridized}}\right] = \text{Probe n Expression Ratio}$$

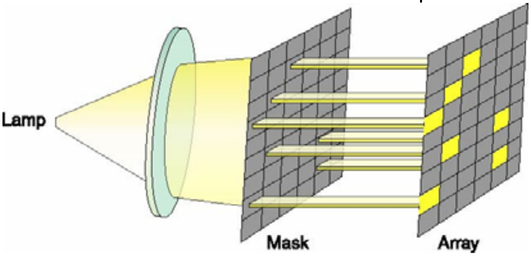
# Oligonucleotide Arrays

- Oligonucleotide array – e.g. Affymetrix
  - More expensive technology
  - Small (11-25 mers) or large (50-70 mers) sequence is attached to chip (probe)$^2$
  - Allows for non-repetitive or unique probe design for a particular gene
  - Multiple probes represent same gene/EST with overlap method
  - Each probe has a mismatch complement with single bp mutation
    - Cross-hybridization
    - Background correction
  - Utilizes photolithography technology to attach probes to chip

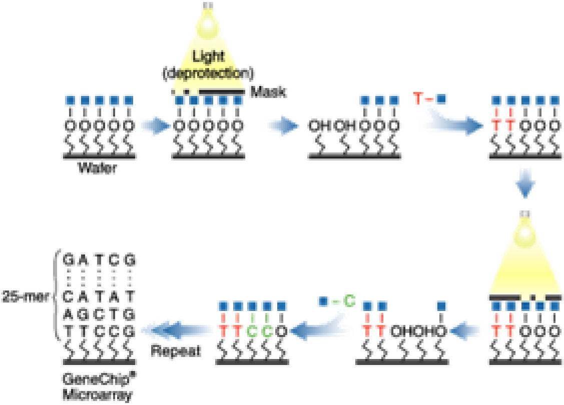


# Oligonucleotide array design

- In situ synthesis is used to build probes bp-by-bp using synthetic organic chemistry
  - Mask is designed for each array type (e.g. species) once
  - Affymetrix example shown on the right and described in next few slides
  - Photolithography technology (attachment using light)



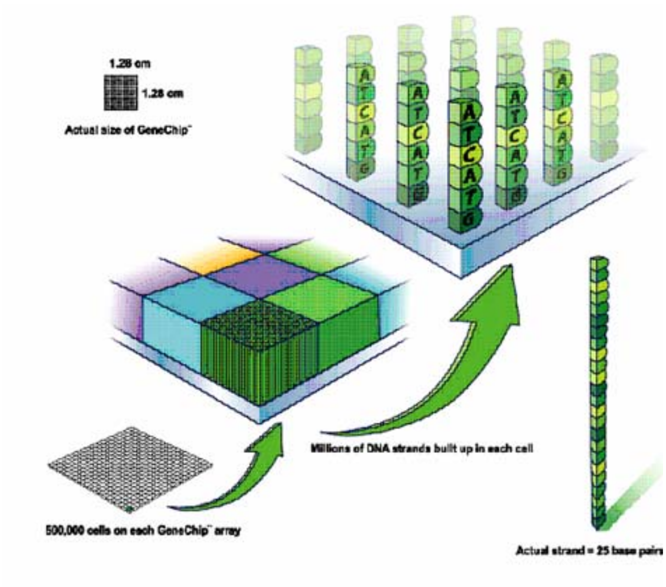
*http://www.affymetrix.com/corporate/media/image_library/low_res/photolitography.jpg*



*http://www.affymetrix.com/technology/manufacturing/index.affx*

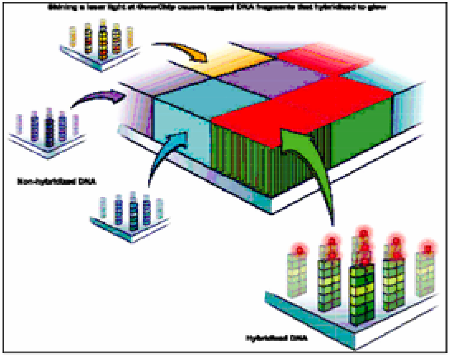
# Oligonucleotide technology

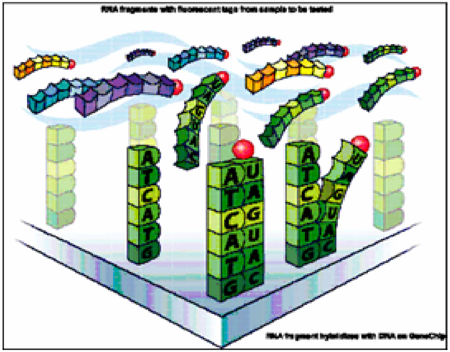
- Every microarray has up to 500,000 individual probe-cells, each 18µm across and containing millions of identical DNA molecules.[2]
- The human U133A array, for example contains over 260,000 different probes that together measure the expression of 22,283 different transcripts at once. [2]
- Chips exist for a variety of organisms including human, mouse, yeast, arabidopsis, and rat


1.28 cm
1.28 cm
Actual size of GeneChip™
A T C A T G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

# Oligonucleotide technology

- Fragmented RNA is labeled with a fluorescent tag and run over the chip
- Wherever there is a complementary probe sequence on the chip, the RNA can hybridize to it. [2]
- Since there are millions of oligos for each probe-sequence, the amount of labeled RNA that sticks corresponds to the amount in solution. [2]
- When the chip is scanned by a laser, the tagged fragments fluoresce, producing spots with a brightness proportional to the amount of RNA that has hybridized. [2]
- This is recorded by a camera and the array image processed by computer to produce expression levels for the different genes. [2]


Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA


RNA fragments with fluorescent tags from sample to be tested
A T C A T G
A T C A T G
U A G U A C
A T C A T G
U A G U A C
RNA fragment hybridizes with DNA on GeneChip

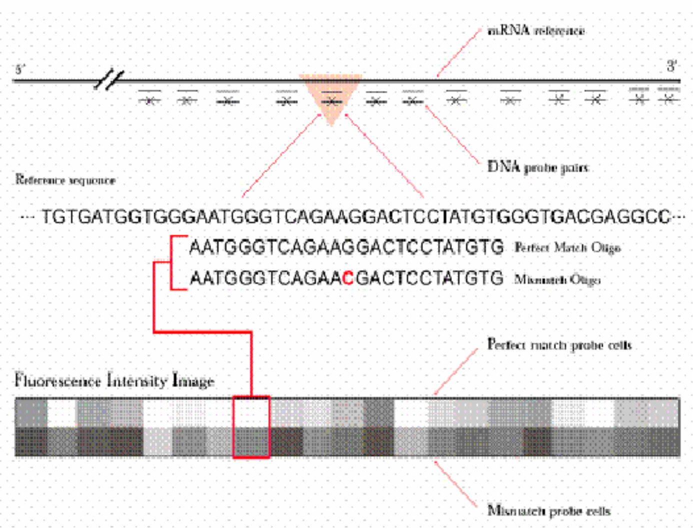
# Oligonucleotide technology

The chips are designed so that every transcript is represented by between 11 to 20 probes that match different parts of the 3' end of the mRNA sequence.[2]

Every chip probe consists of a pair 25 base oligos, one a perfect match (PM) to the transcript, the other a mismatch (MM) in which the middle residue has been changed.[2]

This probe-pairing strategy helps minimize the effects of non-specific hybridization and background signal.[2]

Figure 3: Oligonucleotide Probe Pair


mRNA reference
5'
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

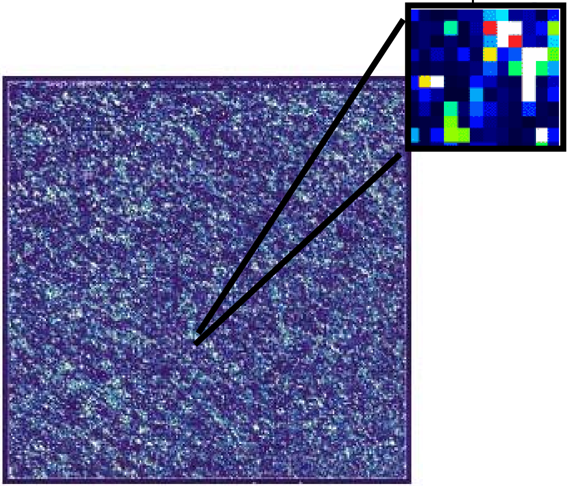
# Oligonucleotide array

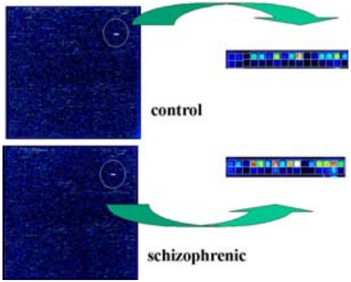
Once the probe has hybridized, chips are scanned to generate an image (dat file).[2]

Each spot, or feature, is ~ 20µm square and is scanned at a resolution of 3µm - giving an average of 49 pixels per spot.[2]

The array analysis software identifies individual features and overlays a grid separating each spot from its neighbors. [2]

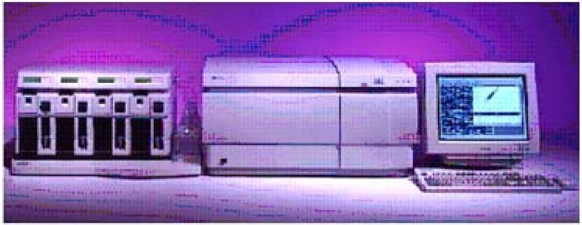
The expression level for a gene is calculated by subtracting the MM from the PM probes.[2]





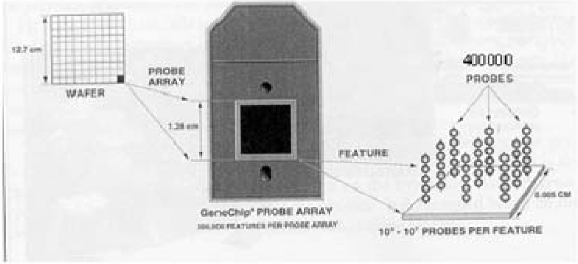
# Oligo Technology (cont.)

Fluidics machine, scanner, and software



An Affymetrix chip

Figure 2:Manufacturing GeneChip probe array

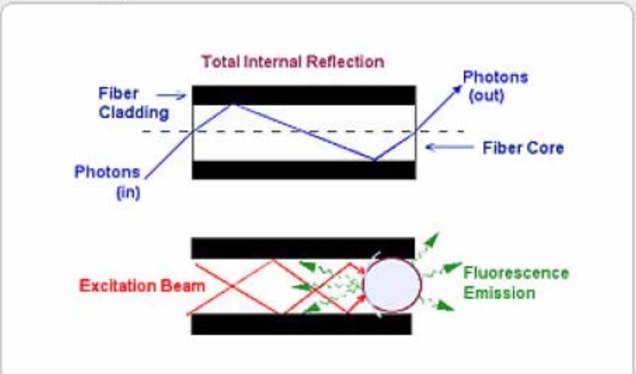


# Commercial array suppliers

- Affymetrix
- Nimblegen
- Agilent
- Nanogen
- Illumina
- CodeLink
- Many others…

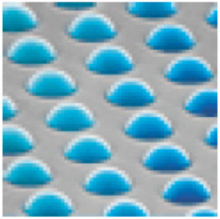
# Illumina arrays

- Oligonucleotides are linked to beads, each bead is color-coded. The beads are bound in the tips of optical fibers, which are then bundled together.

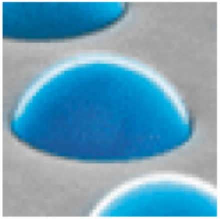


Individual fibers conduct light to enable data acquisition and quantitation of signal emitted by each bead.

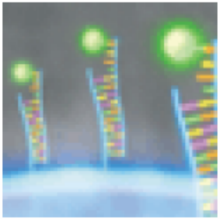
# Illumina bead process



Ten of thousands of wells with hundreds to thousands of bead types can be assembled into each array bundle

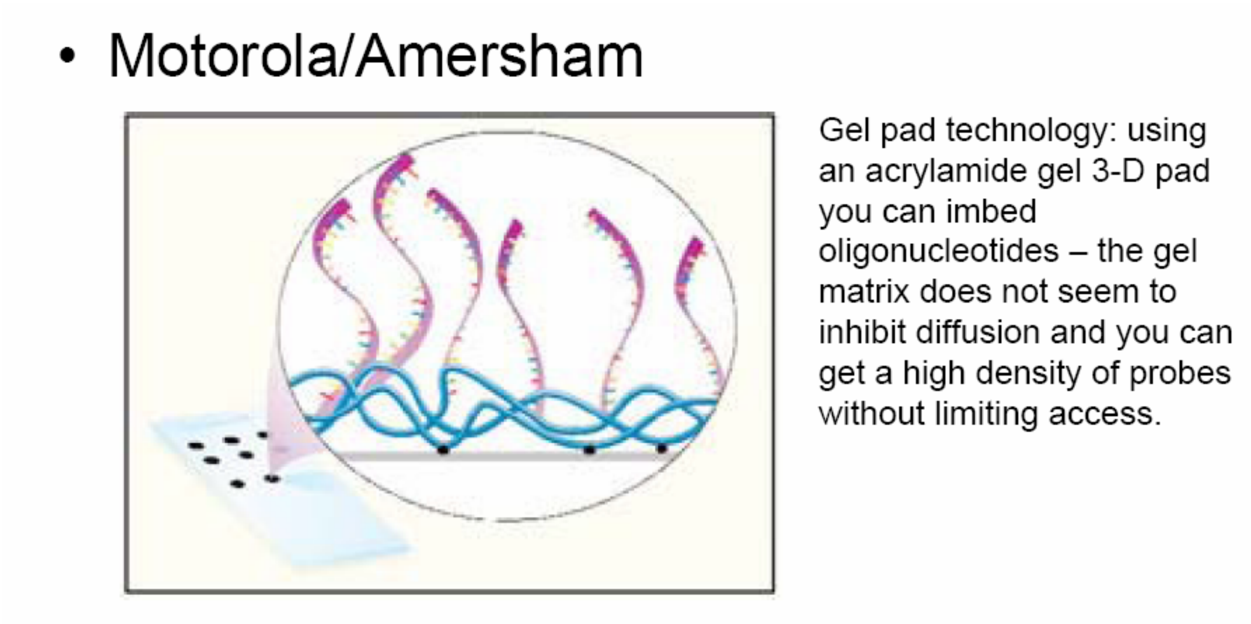


Determine which bead type occupies which well using a proprietary decoding process



The molecules in the sample bind to their matching molecules on the coated bead

# CodeLink arrays

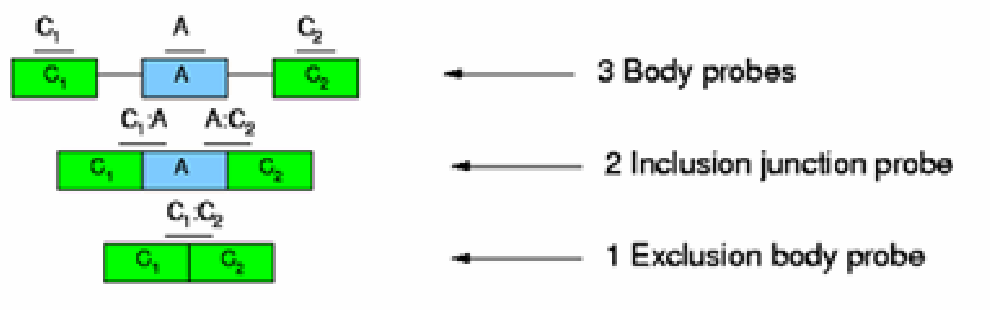


- Motorola/Amersham

Gel pad technology: using an acrylamide gel 3-D pad you can imbed oligonucleotides – the gel matrix does not seem to inhibit diffusion and you can get a high density of probes without limiting access.

# Exons and splice variants

- We know from molecular biology, that RNA undergoes processing prior to determining a final form
  - Introgenic regions can be spliced out, while exon regions are combined
  - These recombined transcripts are knwon as splice variants



# Exons and splice variants

- Why bother with looking for expression patterns in splice variants?
  - Many genes are products of splice variants
  - They can be strong indicators of certain diseases as well as specific drug response
  - Within a population, a specific gene may show high differential expression due to a disease condition
  - However, within a small subpopulation, those members that have an alternatively spliced version of the gene show not differential expression from the disease condition

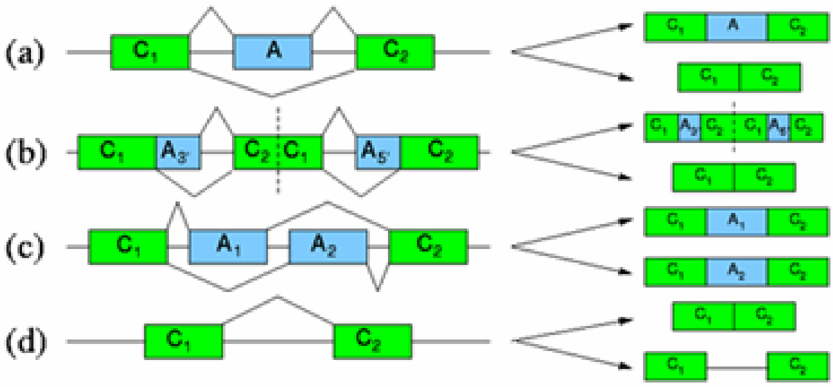
# Exon arrays

- Typical microarrays are designed with probes to target the 3’ end of the gene
- This type of design only measures the abundance of mRNA molecules that hybridize to the probe at the terminal region
- Such design does not take into account splice variants that can result from alternative splicing patterns
  - Requires probes that are complementary to multiple regions along the gene to measure possible splice variants
  - Requires information about known splice variant sequences to design probes for possible combinations of splice events

# Exon arrays

- The example below demonstrates 4 different exon (green blocks) patterns with introns (blue blocks) and without
  - Different splicing patterns can create multiple isoforms of the gene, depending on where the introns are spliced out and which exons are combined
  - Must measure such splice variants for each gene
  - Expression patterns can vary for multiple isoforms of the same gene

- Exon arrays are a relatively new technology so we will not spend a lot of time on this topic, however, it is important to be aware of the advantages that are provided with measuring alternatively spliced events
  - Primary companies: Affymetrix and Agilent (ExonHit sequences)



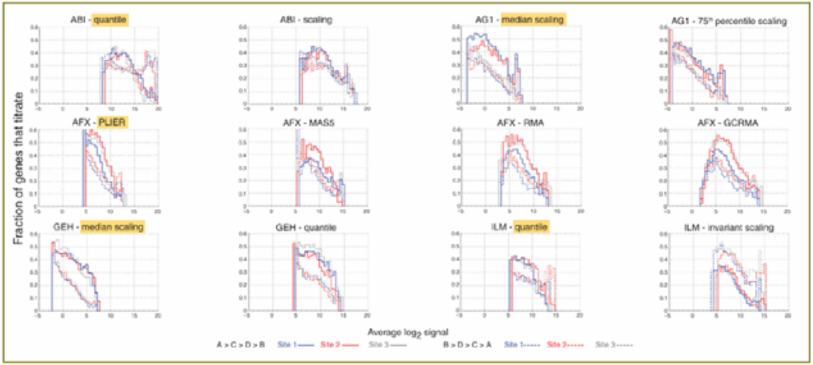
# Differences in array platforms

- Consistent comparisons across different array platforms have been difficult to make
  - Differences in vendors (e.g. Affymetrix and Agilent)
  - Differences between versions of an array from the same vendor (Affymetrix u95 vs u133 series arrays)

- Results have been shown to vary when comparing different array platforms
  - Different sensitivities in array technologies
  - Different probe selection designs
  - Different signal intensity normalization methods
  - etc.

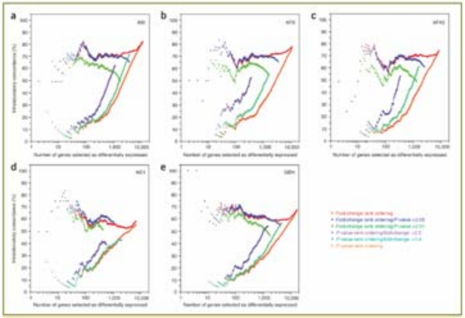
# Differences in array platforms

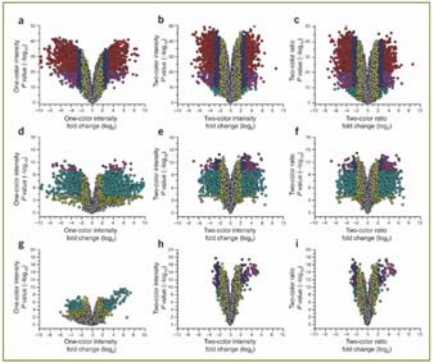
- The MicroArray Quality Control I (MAQC)[3]
  - A community led by the FDA to understand the use of microarray technology in clinical and regulatory settings
  - Published a series of studies to compare array platforms both to each other and quantitative results (RT-PCR)
  - Provide some guidance on methods that perform most consistently across platforms and identify those factors that do not tend to vary



# Differences in array platforms

- Studies published by MAQC sought to assess:
  - Comparison of microarray data to quantitative gene expression
  - Comparison of normalization methods
  - Use of external controls
  - One channel vs. two channel microarrays
    - One channel platforms: Affymetrix, Applied Biosystems, Eppendorf, GE Healthcare, Illumina
    - Two channel platforms: Agilent, TeleChem, CapitalBio
  - Intraplatform reproducibility
  - Analytical consistency across platforms





# Standards in the microarray community

- Minimal Information About a Microarray Experiment (MIAME)
  - Organization set up to provide standards in microarray experiments and analysis
  - Provide guidelines on the minimal necessary information for interpretable results
  - Encourage depositing data into public standard repositories
    - Journals and funding agencies
    - Most journals now require depositing data prior to publication

- Guideline examples
  - Experimental design
  - Array design
  - Samples
  - Hybridization parameters
  - Normalization methods

# Interpreting the biology

- Hybridization kinetics
  - Ideally, the probe that minimizes hybridization free energy is the optimal one to represent a gene
    - However, we cannot currently compute the free energy from the sequence alone
  - The hybridization free energy for a gene depends on the concentration of that gene
    - The less expressed gene with higher free energy can give a greater signal than the more expressed gene, if it is given in greater concentration
- mRNA expression vs. protein expression
  - Gene interactions sometimes do, but can also have minimal similarity to protein interactions
    - Kinases, Receptor-ligand binding, Protein docking, etc.
    - Half-lives of mRNA and its protein are not always proportional
- All gene expression events do not result in mRNA transcripts
  - tRNA, rRNAm snRNA
- Splice variants

# References

1) Li Fugen, Stormo D Gary,. (2001) Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics.* **17**,1067-1076.

2) The Paterson Institute: Onco-Informatics group

*http://bioinformatics.picr.man.ac.uk/mbcf/overview_ma.jsp*

3) Shi et al., (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotech*. **24**:1151-1161.

4) Yang Y, Dudoit S, Luu P, and Speed T. Normalization for cDNA Microarray Data. (2000*) UC Berkeley Tech Report.*

5) Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, and Speed T. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acid Research*. **31**.

6) Dudoit, S., Gentleman, R., Irizarry, R., and Yang, Y. (2002) Pre-processing in DNA microarray experiments. *Bioconductor short course.*

7) Bolstad BM, Irizarry RA, Astrand M, and Speed T. A comparison of normalization methods for high density oligonucleotide array based on variance bias. *Technical Report*.