# Moc/Bio and Nano/Micro

## Lee and Stowell

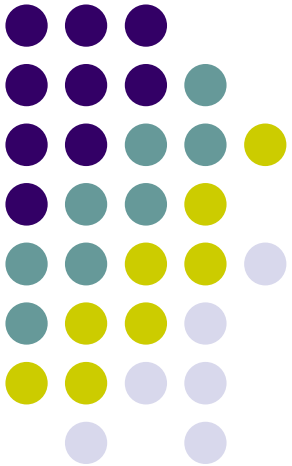
Moc/Bio-Lecture GeneChips

**Reading material**

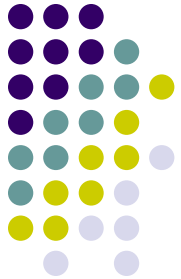
http://www.gene-chips.com/

http://trueforce.com/Lab_Automation/DNA_Microarrays_Industry.htm

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



# Gene Chips



- 1) Why do we want them
- 2) What are they and what can we do with them?
- 3) How are they made and implemented?
- 4) How is the data analyzed?

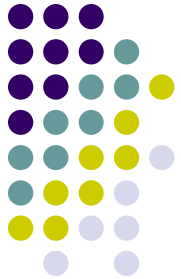
# 1) Why gene chips

- Although having the entire human genome is useful, it does not tell us much about functional interplay of genes
- We want to understand the complex interplay of various genes
- We need very high throughput technology to achieve this

# Some of These Databases Include

- The Human Genome is searchable at http://www.ncbi.nlm.nih.gov/genome/guide/human/
- A challenge facing researchers today is the ability to piece together and analyze the multitudes of data currently being generated through the Human Genome Project. NCBI's Web site serves an an integrated, one-stop, genomic information infrastructure for biomedical researchers from around the world so that they may use this data in their research efforts.
- MGC – http://mgc.nci.nih.gov/
- The goal of the Mammalian Gene Collection (MGC) is to provide a complete set of full-length (open reading frame) sequences and cDNA clones of expressed genes for human and mouse. The MGC is an NIH initiative that supports the production of cDNA libraries, clones and sequences. All the resources generated by the MGC are publicly accessible to the biomedical research community.

# What are they and what are they used for?

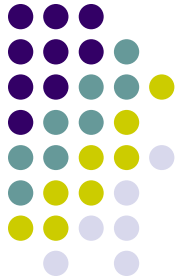


- A detector array for probing
  - Gene variation
    - Better RFLP (restriction fragment length polymorphism)
    - SNP (single nucleotide polymorphism) analysis
  - Gene expression levels
    - Disease cell versus normal
    - Environment stress
    - Host guest relationships
    - Pharmacogenetics

# They allow us to study

- Identification of complex genetic diseases
- Drug discovery and toxicology
- Polymorphism diseases (SNPs)
- Pathogen analysis (Host/Guest studies)
- Variable gene expression over time, disease states etc.

# And will have significant impact on

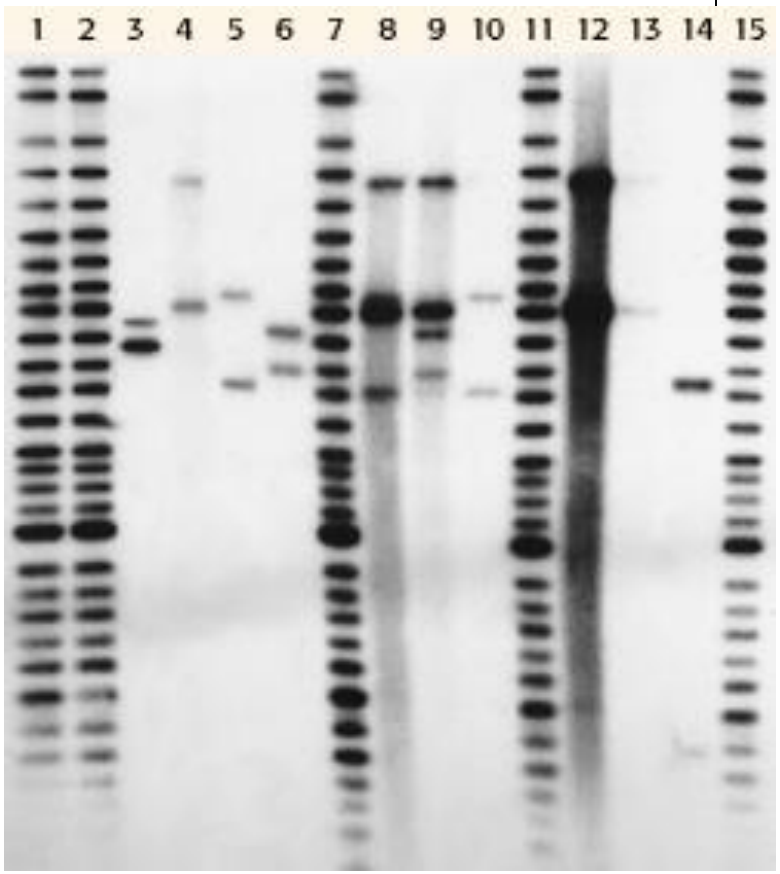


- Preventive medicine
- Disease sub typing
- Pharmacogenetics (optimized drugs based on genetic background) to maximize effectiveness and minimize side effects
- More effective anti-pathogen treatments
- And more

# 2) What can we study with these gene chips?

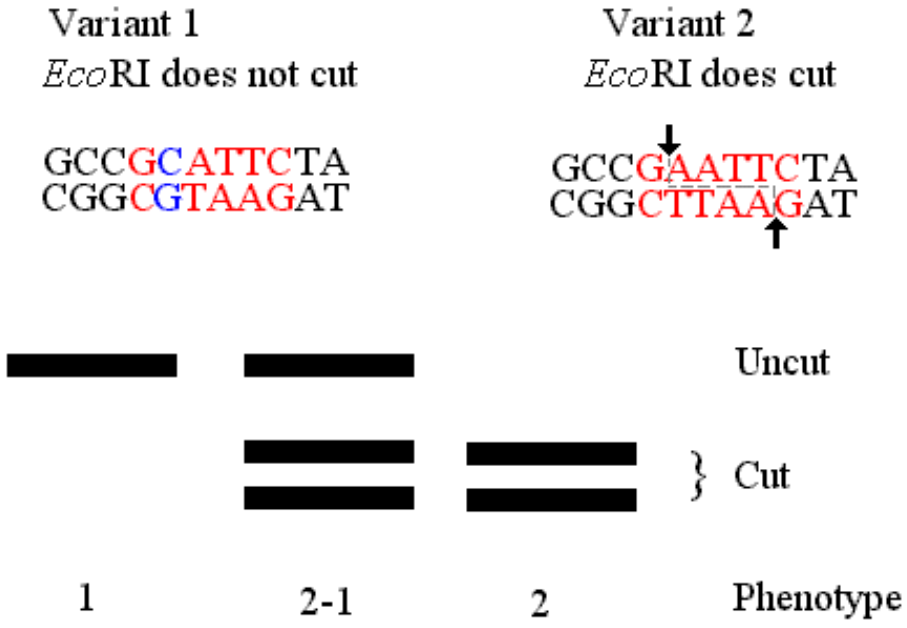
# RFLP-the first crude gene "chip"

- These are small and frequent differences in individuals' DNA that allow for specific patterns to appear when digested by enzymes or when sequenced
- "DNA fingerprinting"



# Restriction Fragment Length Polymorphisms

- Restriction enzymes are used to cut DNA at specific sites and create a distinct pattern for the DNA of each individual.
- These are used as markers on both physical and genetic linkage maps.



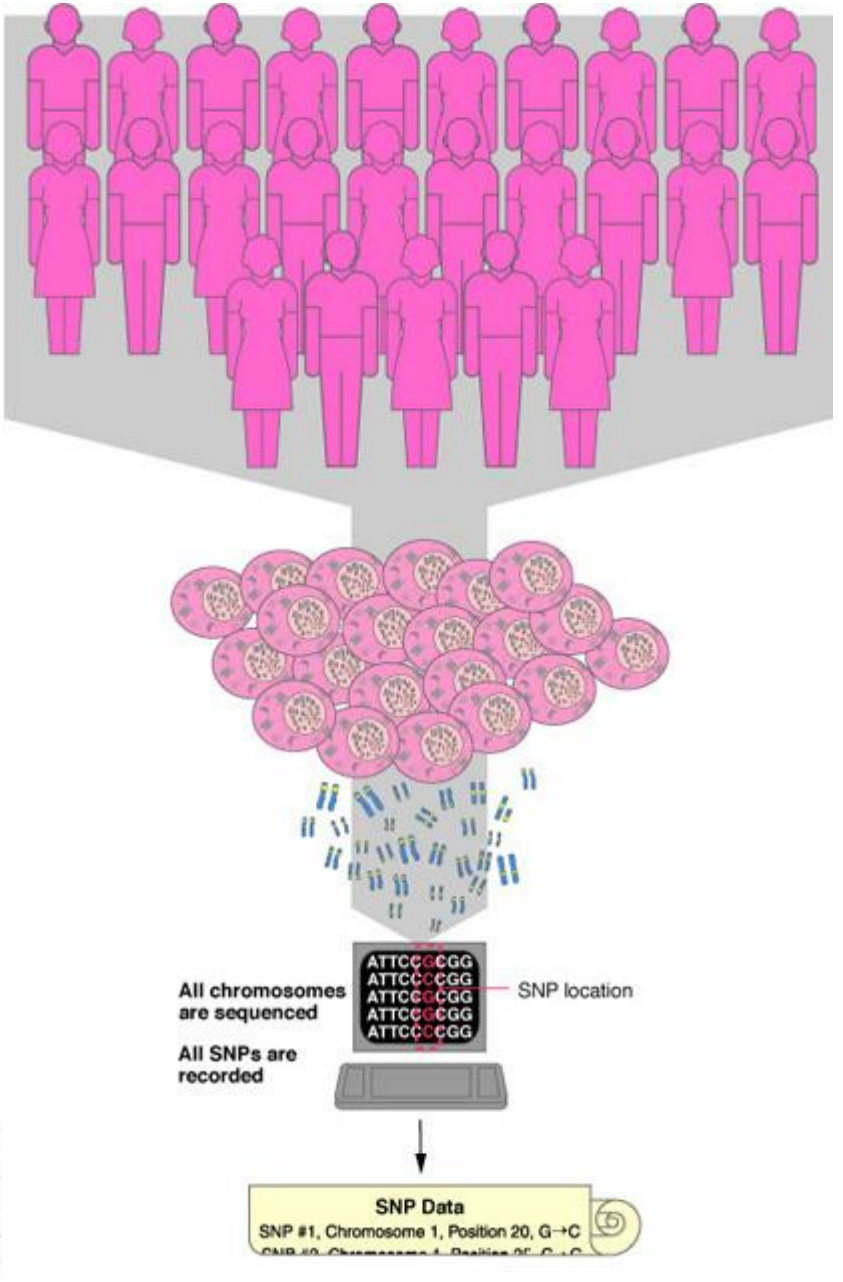
# Single Nucleotide Polymorphisms

- This is the most common genetic variation and occurs once every 100-300 bases, i.e. 1-3x$10^7$ SNP's in the human genome
- SNPs can be used to distinguish individuals or to track heredity.
- Researchers are looking for association between disease occurrence and specific changes in SNPs.
- Difficulty in getting enough data (patient samples)

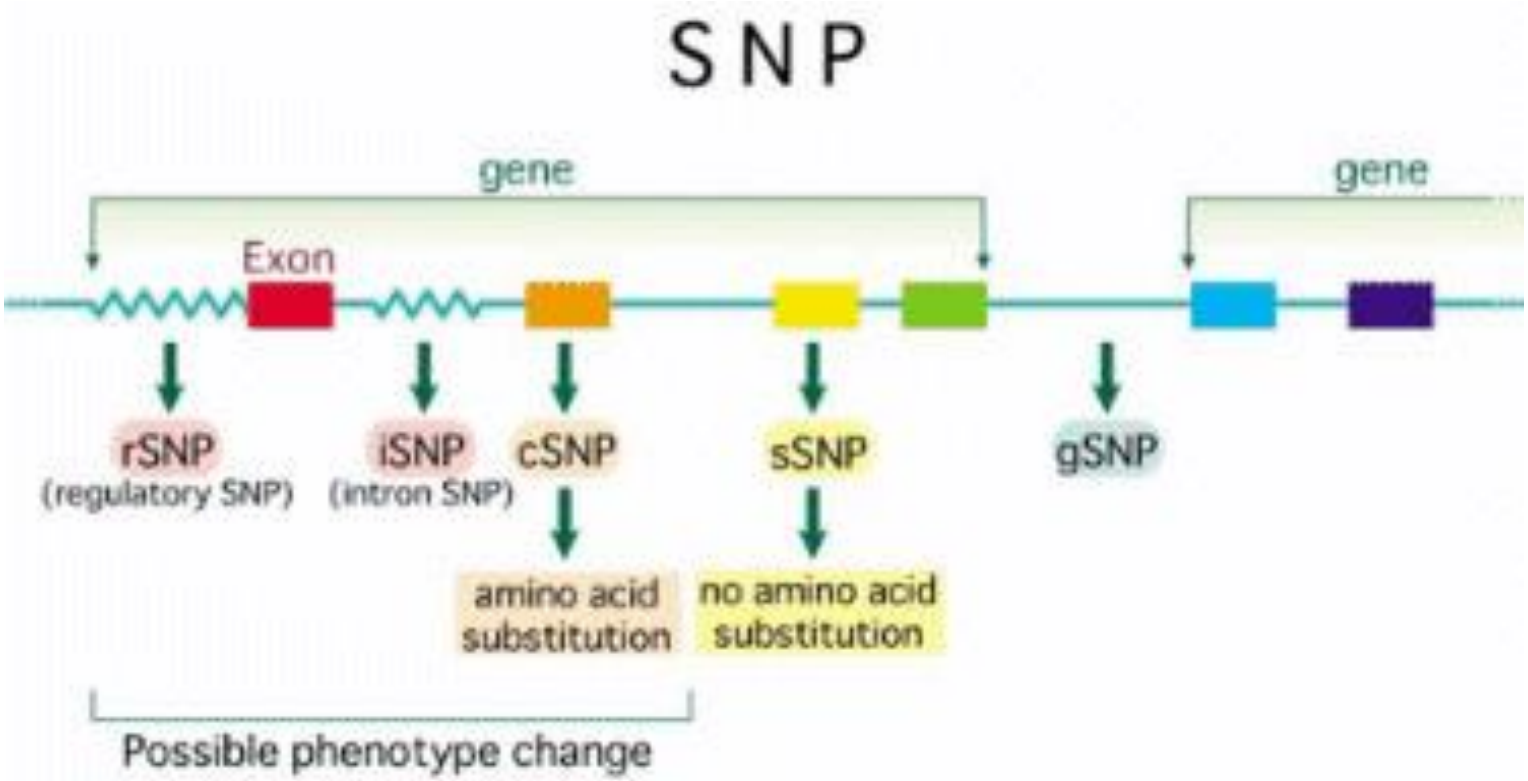
# Finding SNP's is laborious.

Currently about $2x10^6$ SNP's available

http://snp.cshl.org



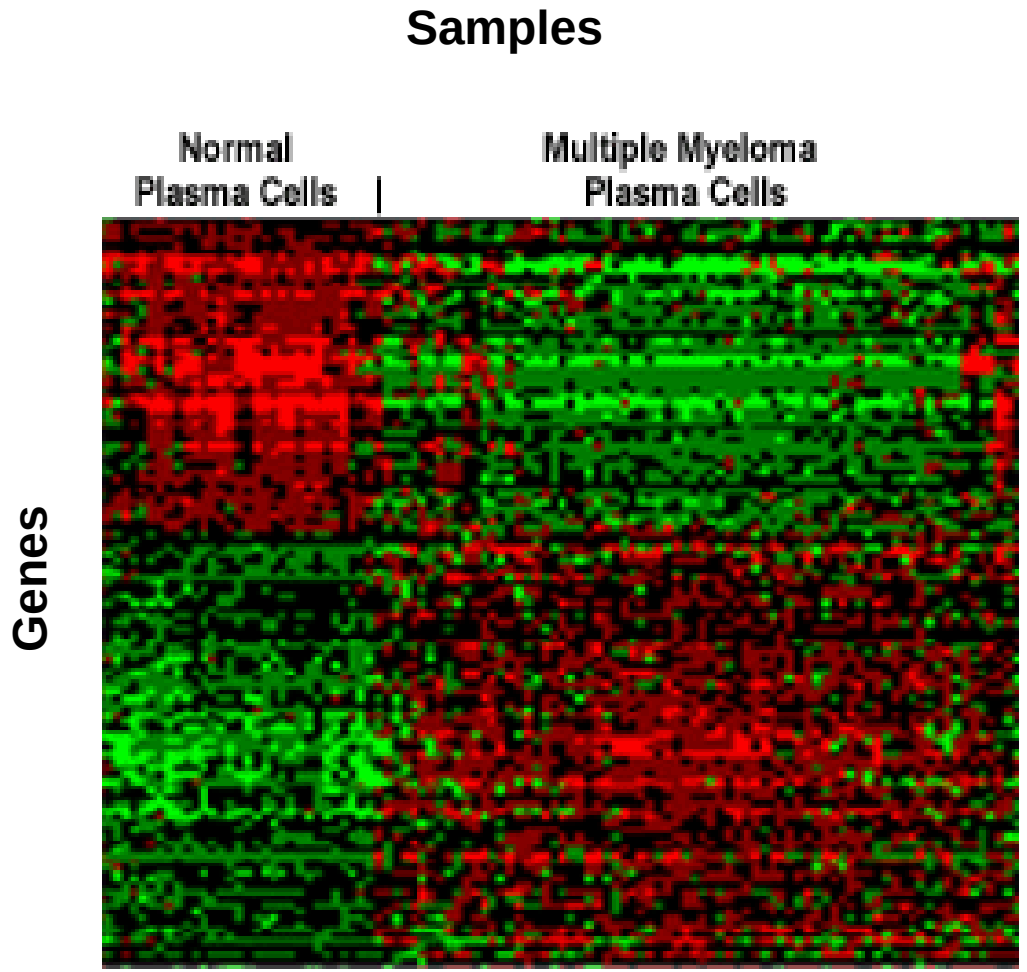
# Types of SNP's




cSNP = changeSNP
sSNP = silentSNP
gSNP = intergenicSNP

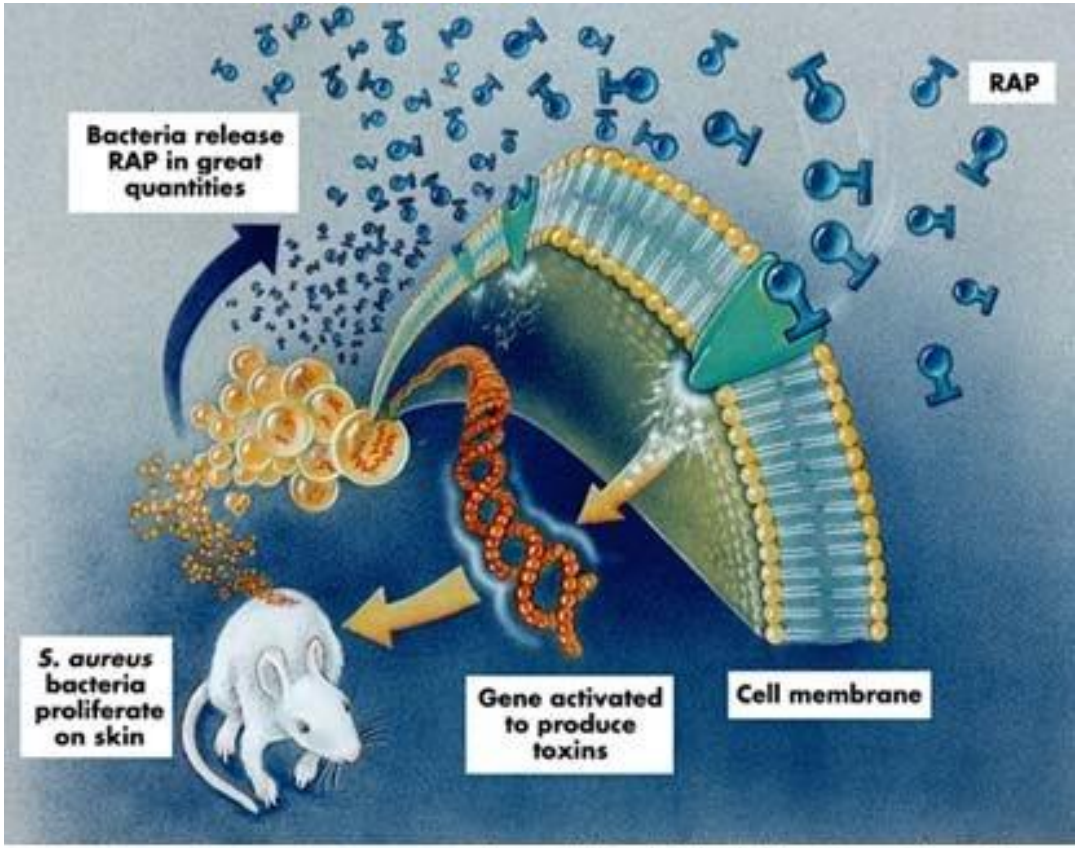
# Gene expression levels

- Compare normal and abnormal cells
- Compare stressed and unstressed cells
- Red = high
- Green =low



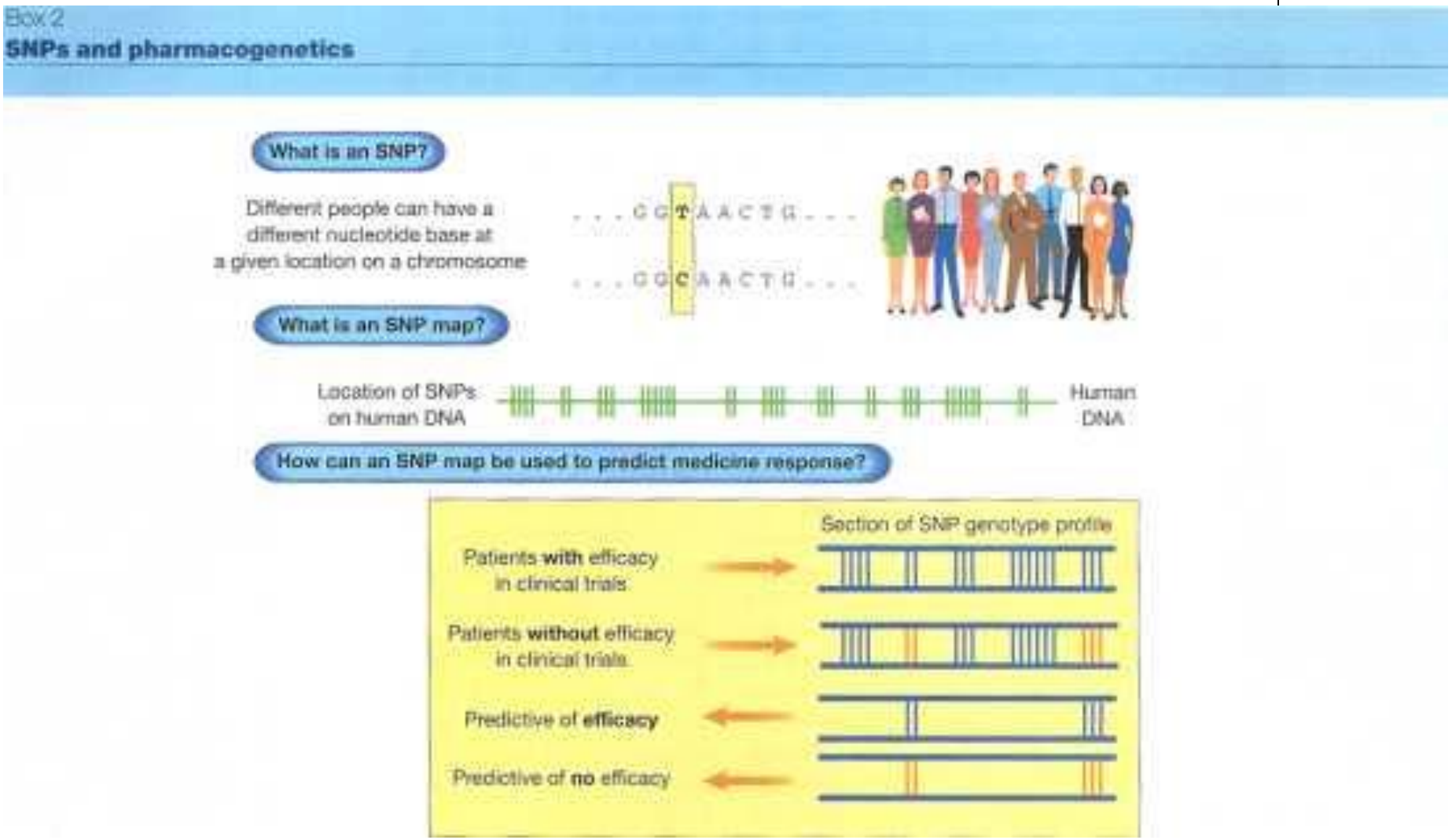
# Host guest relationships

- Study expression levels of the host to discover important response genes
- Study expression levels of the guest to discover important drug targets




Source: UC Davis School of Medicine and Medical Center
Artwork by Nelva B. Richardson

# Testing SNP's for relational properties (pharmacogenetics)



# Division of Information

- Genomics – DNA
  - SNP analysis
- Functional Genomics –
  - mRNA (DNA Arrays)
  - Proteomics - Protein

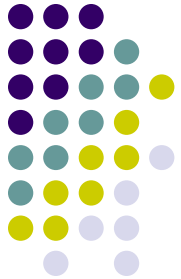
# Types of Gene chips

- DNA microarrays
  - Original for studying DNA sequence and
  - mRNA levels
- Aptamer microarrays
  - More recent and less developed
  - For studying protein levels
- Antibody microarrays
  - More recent and less developed
  - For studying protein levels

# What is a DNA Array?

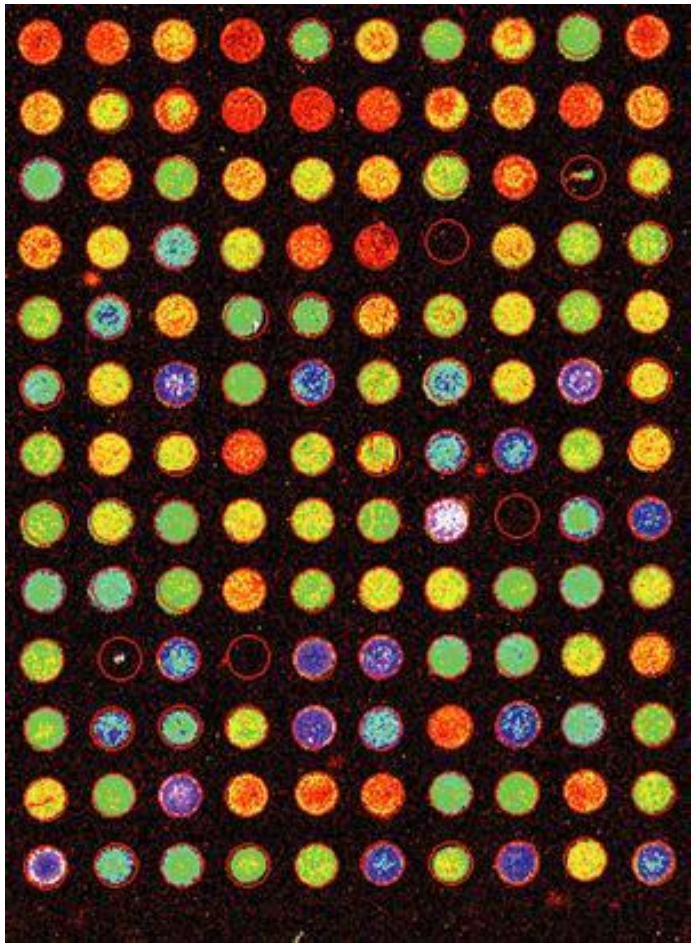
- Simply put, it is a device that allows for DNA to be bound to it for analysis with homologous cDNA or mRNA (usually via DNA)

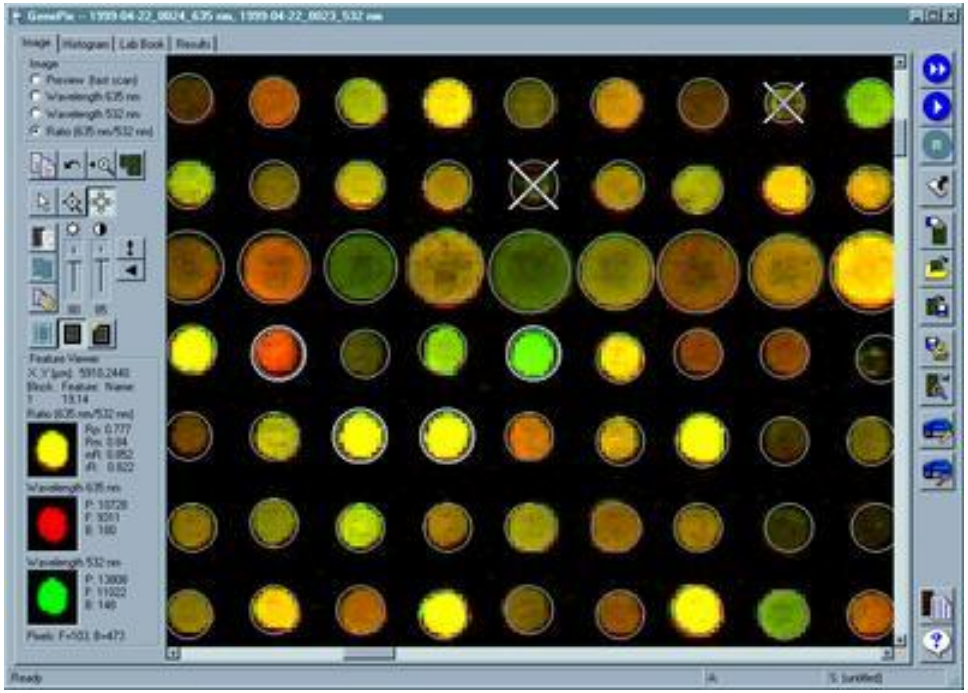
# Sizes of DNA Arrays



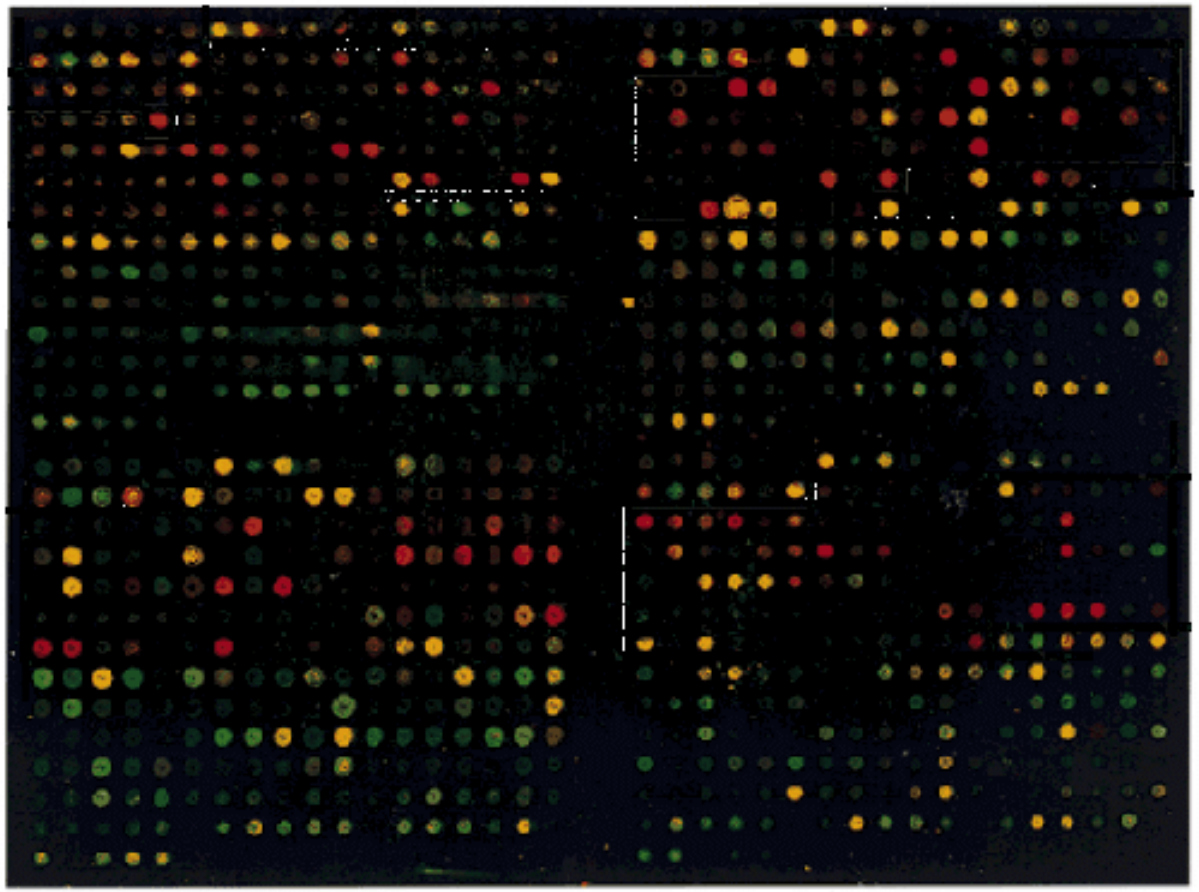
- Macroarrays – Membrane blots 1-10 K Genes
- Microarrays – Glass or polypropylene 10k+ genes
- High-density Oligonucleotide Arrays (Gene Chips) – Silicon 10-100k genes and potentially up to 4 million genes

# Macroarray

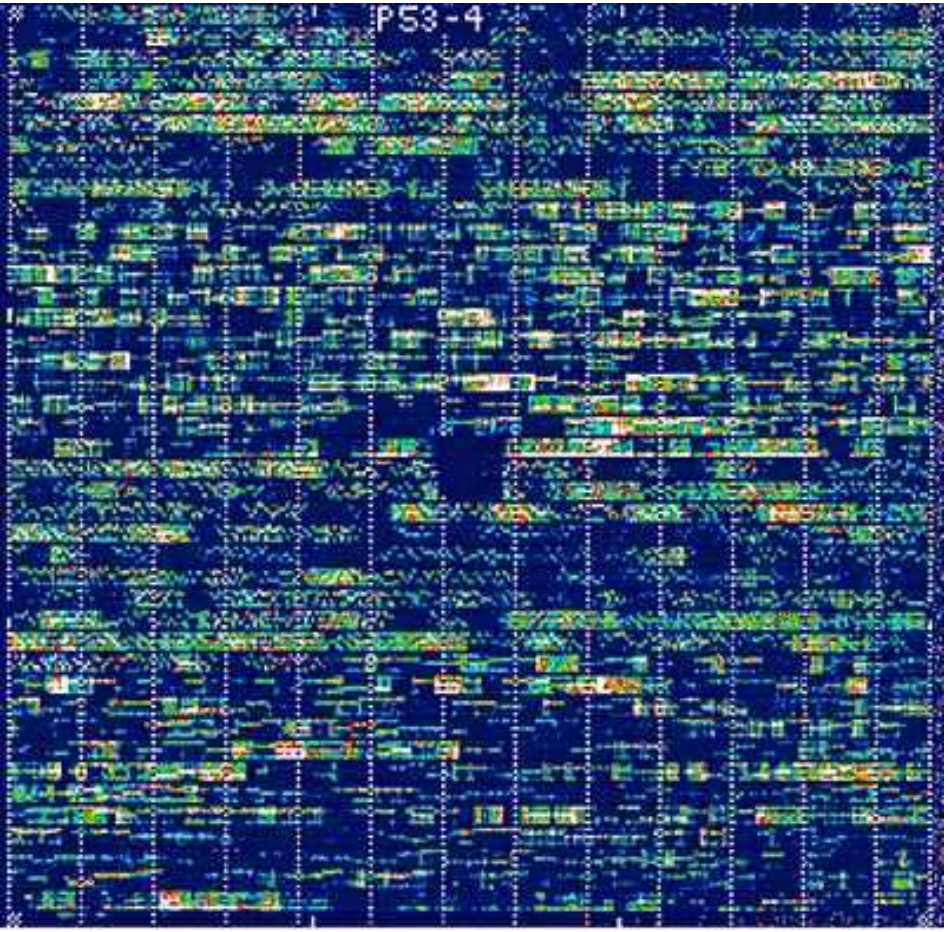


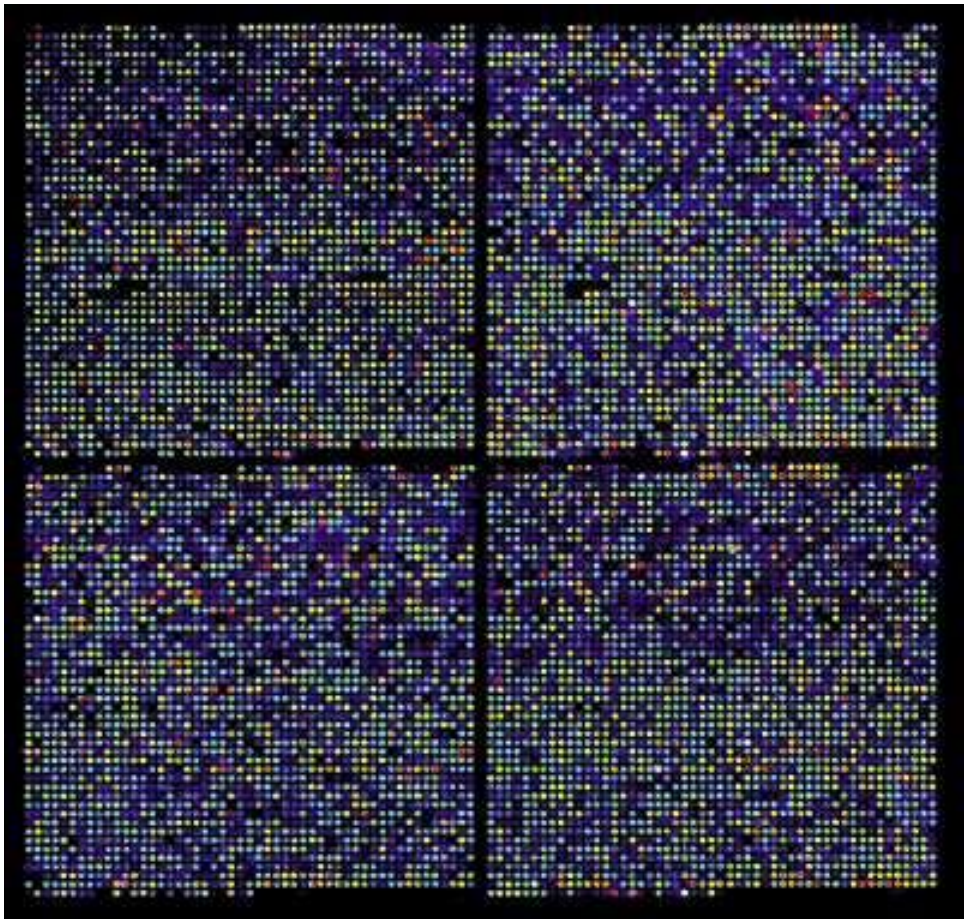


# Microarray

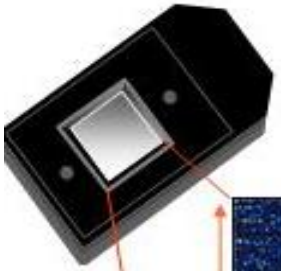


# Gene Chip

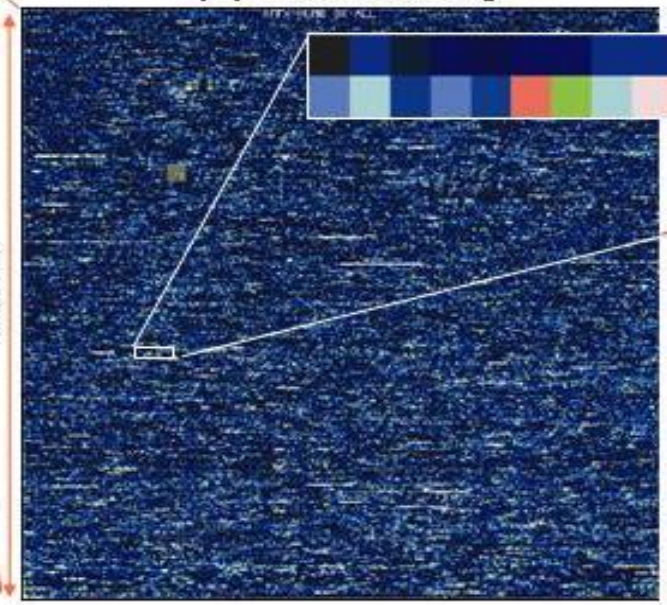


**Human Genome U133A GeneChip® Array**



**(1) Probe Array**




**(2) Probe Set**

Each Probe Set contains 11 Probe Pairs (PM:MM) of different probes

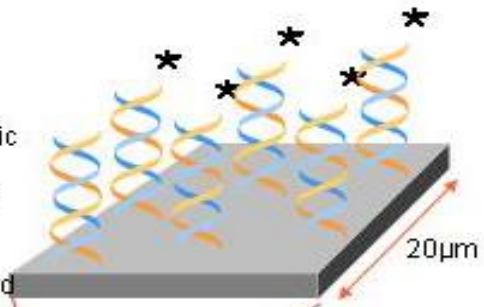
**(3) Probe Pair**

Each Perfect Match (PM) and MisMatch (MM) Probe Cells are associated by pairs

**(4) Probe Cell**

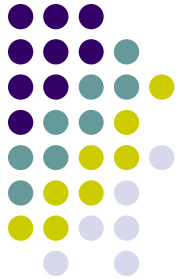
Each Probe Cell contains ~$40 \times 10^7$ copies of a specific probe complementary to genetic information of interest

probe : single stranded, sense, fluorescently labeled oligonucleotide (25 mers)




The Human Genome U133 A GeneChip® array represents more than 22,000 full-length genes and EST clusters.

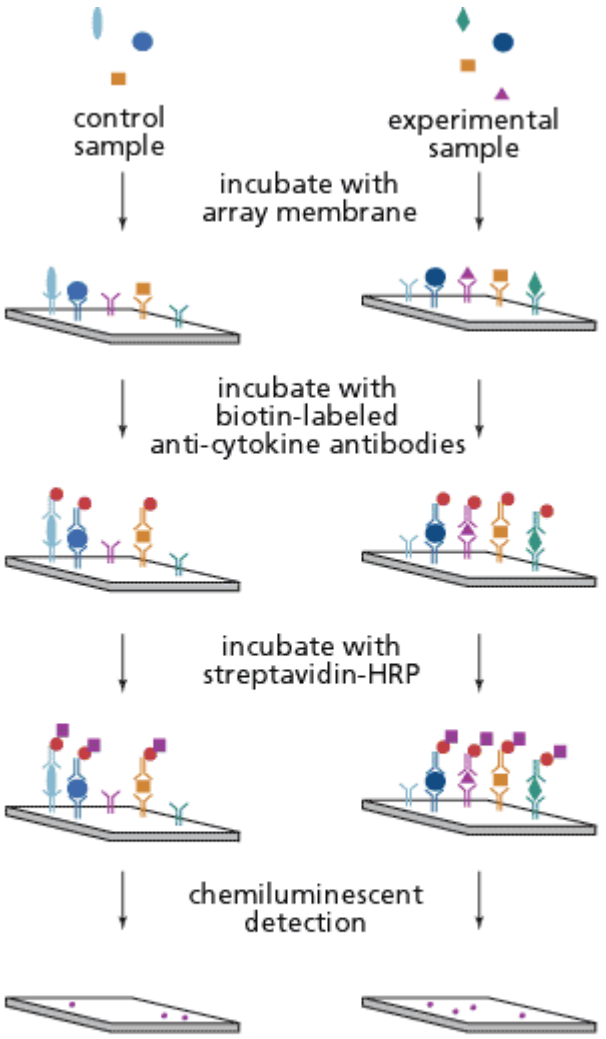
# What Can Fit on a Gene Chip



- The entire human genome can fit on a gene chip eventually
- For now all known human genes will fit. There are several companies that are selling these chips and they can be homemade.
  - Affymetrics (photolithography)
  - Agilent (printed)
  - Illumina (bead array)
  - Nanogen (electrostatics)

# Protein chips

- http://www.ciphergen.com/products/pc/
- http://www.bdbiosciences.com/index1.shtml



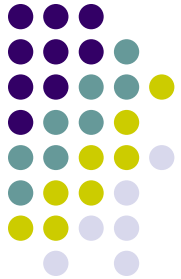
# Protein chips lag behind but

- Advantages are
  - they give direct readout of protein levels not mRNA. Remember gene regulation lecture
  - No intervening steps to produce samples
  - Can also look at splice variants and posttranslational modifications
- Disadvantages are
  - cross-reactivity  of antibodies or aptamers
  - Cost and difficulty to produce selective antibodies or aptamers for all proteins
  - Protein Extraction variations

# 3) How are they made and how do the work

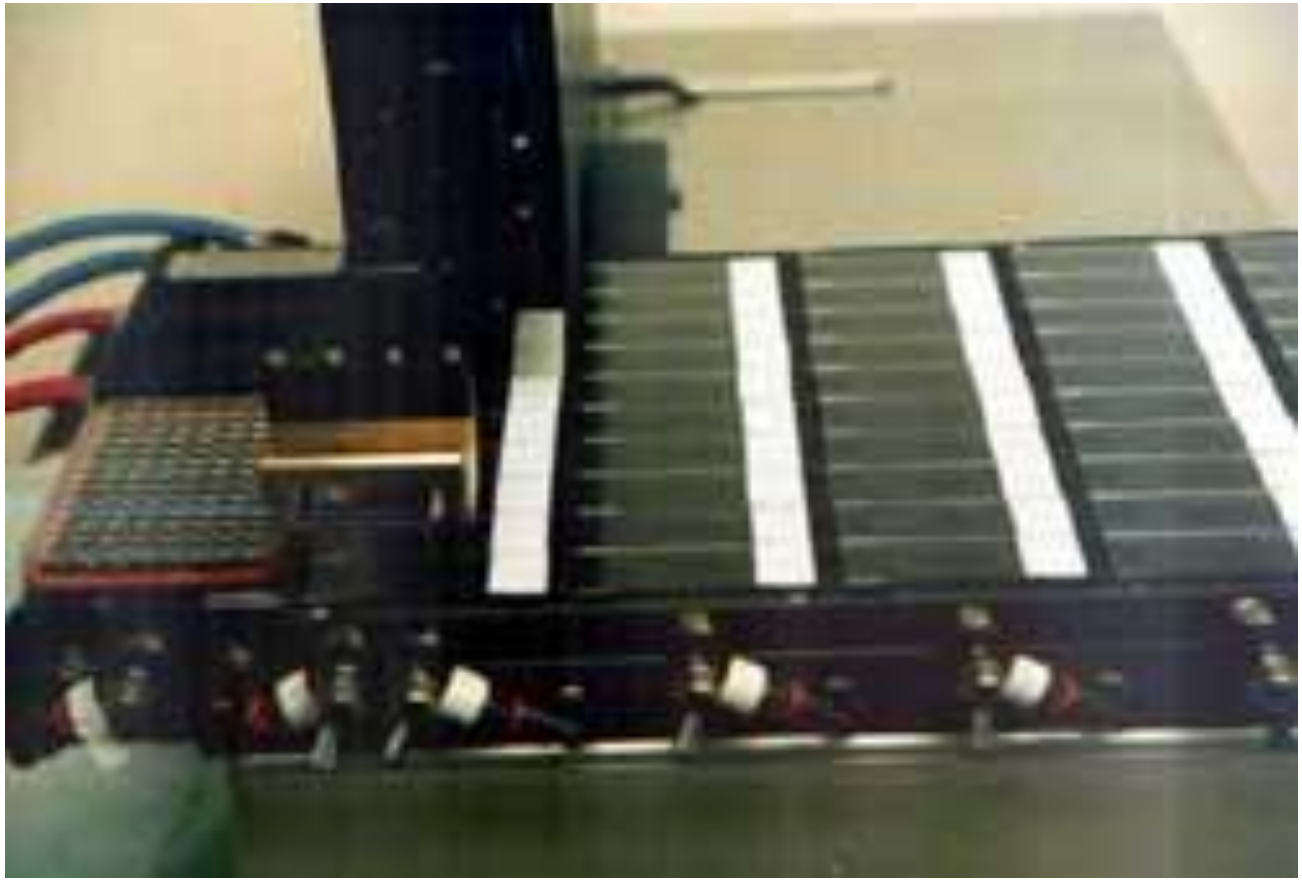
- Bonding DNA to a DNA chip
- Labeling the homologous DNA
  - Reverse transcription and/or PCR amplification
- Running the sample(s) over the chip
  - Simple hybridization
- Reading the chip
  - Fluorescence
  - Mass spec
- Analyzing the data

# Technology to Fabricate DNA Arrays



- Mechanical Micro-spotting – direct physical contact with a small pinhead (~0.1-0.5 micrometers)
- Ink-Jetting – electrically directs bases from jets
- Photolithography – uses semiconductor technology light directing of bases (Affymetrix Gene Chips)

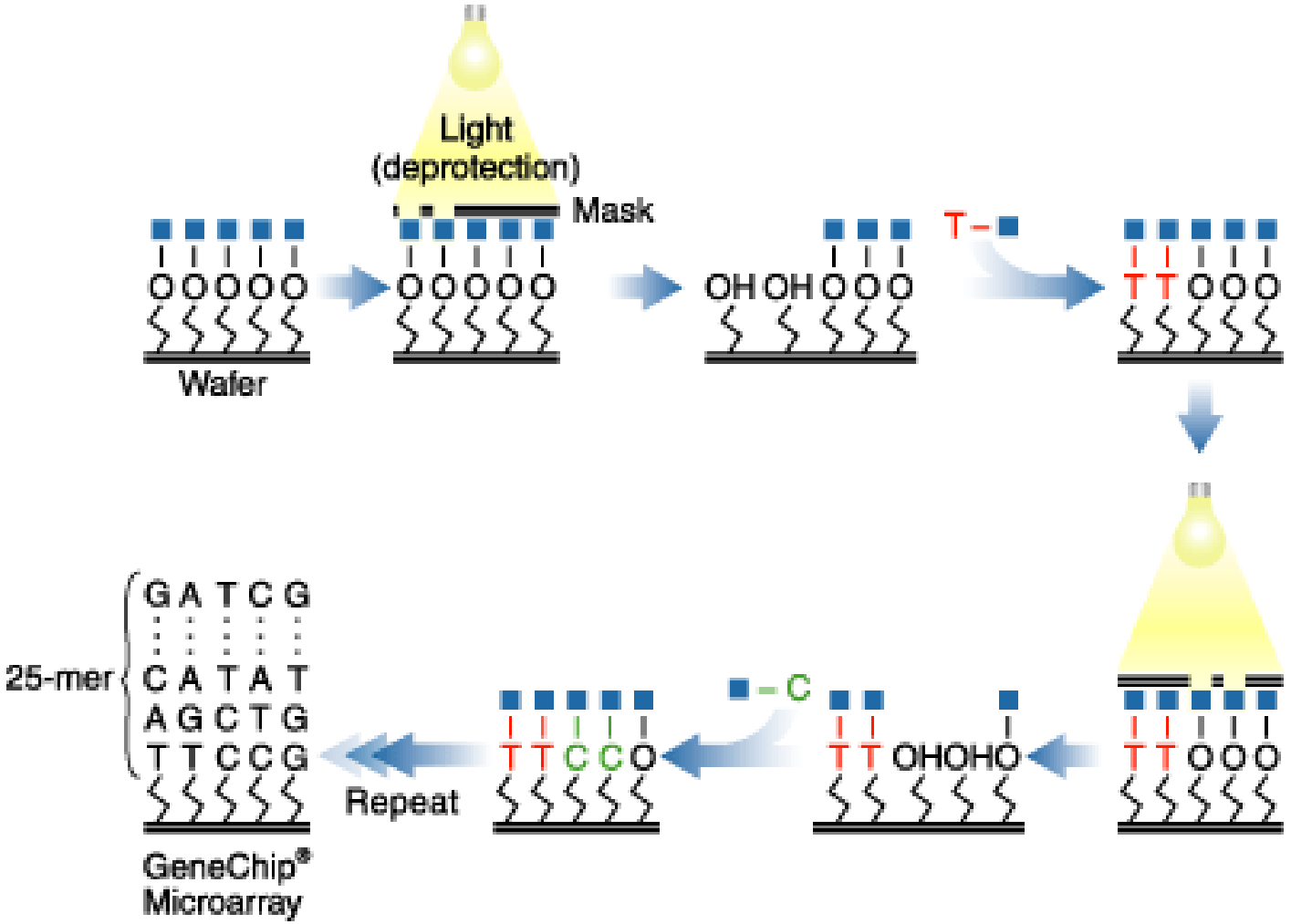
# Microarray spotting with a robot (100 micron) on glass



Synteni/Stanford

# Inkjet technology (micron)

# Photolithographic synthesis Micron or smaller

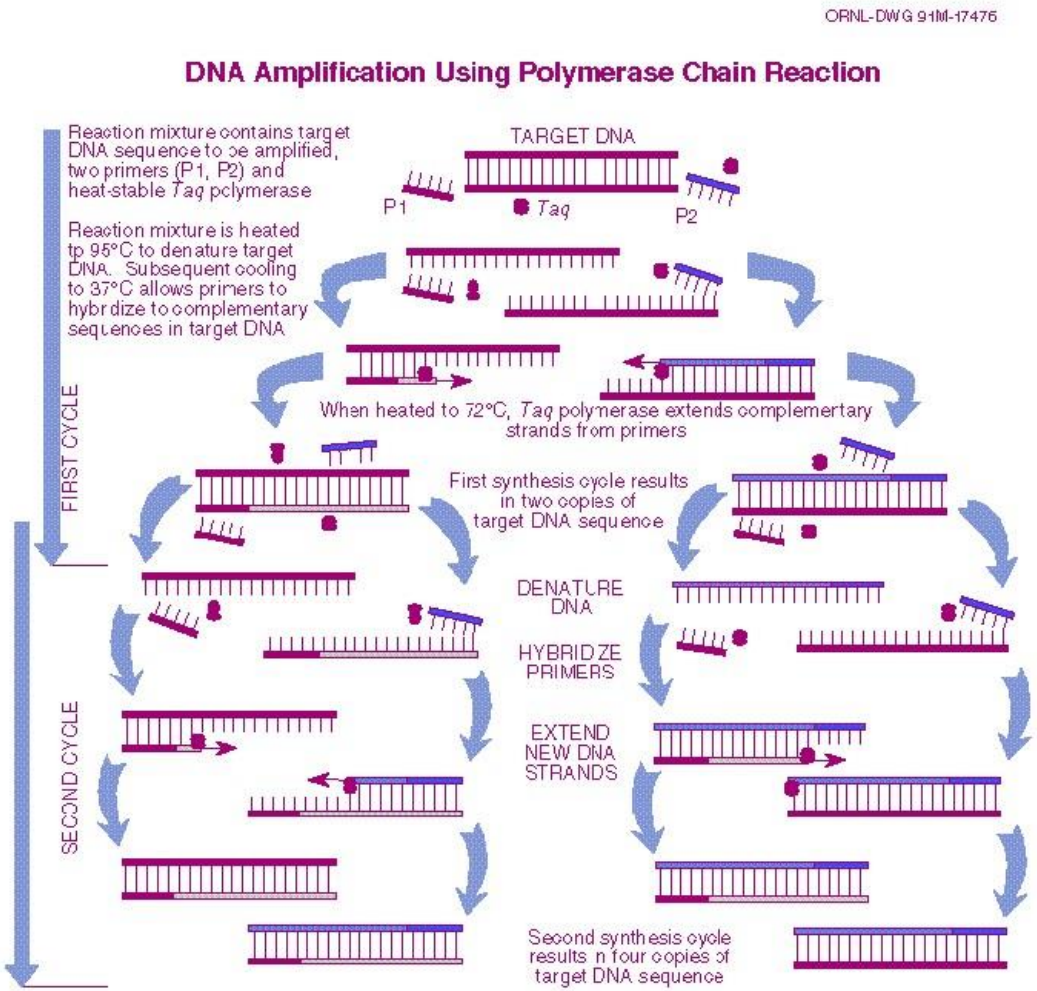



Affymetrix

# Protecting groups

- Nitrobenzoyl groups
  - Fast
  - Reactive side products
- Methoxybenzoins
  - Very fast
  - No reactive side products

# PCR Review

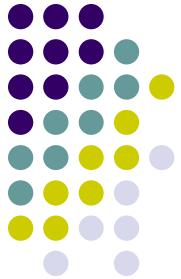



Source: *DNA Science*, see Fig. 13.

# PCR Steps

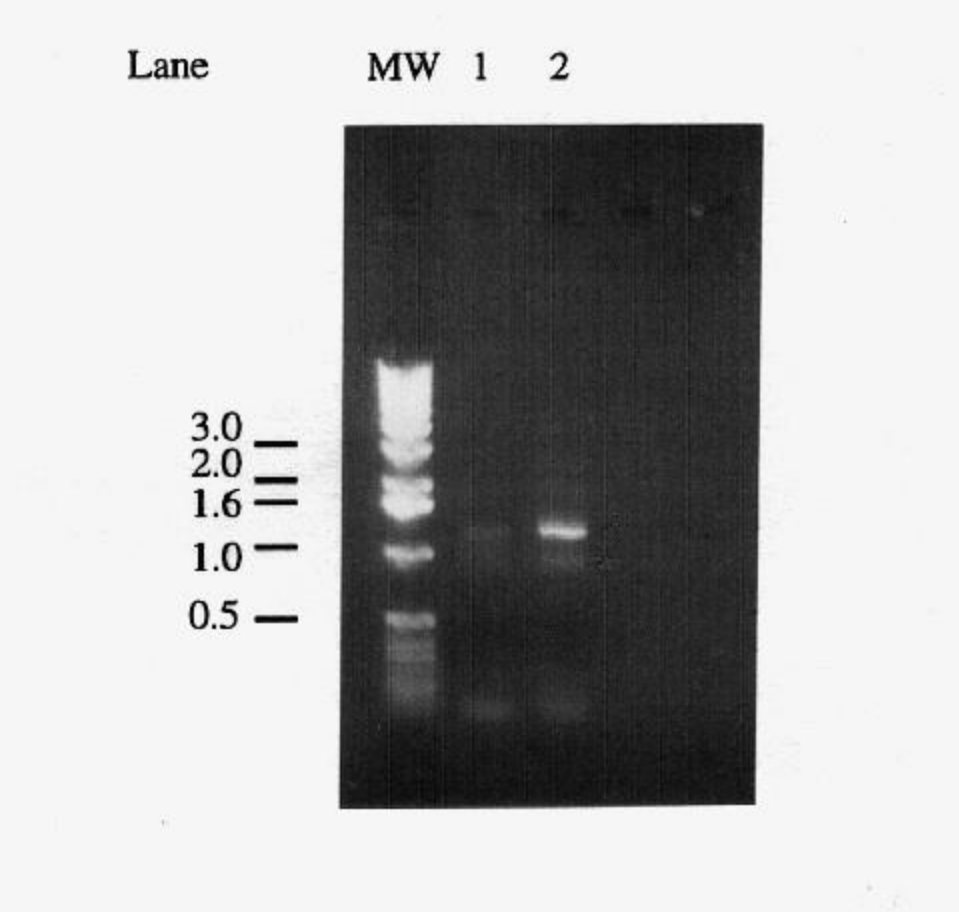
- Genetic material either homogenized or isolated is loaded into a micro-centrifuge tube.
- Oligonucleotides are added to be the “Primers” for the site-specific amplification.
- dNTP’s are added to provide the base pairs for the newly synthesized DNA/RNA.
- Heat stable polymerases are added to conduct the replication
- If the sense is (-) then a reverse transcriptase is added
- Heat is added in a cyclical manner to cause the splitting and annealing of the DNA for the replication assembly to bind and conduct the replication.

# Detecting Binding of Homologous Sequences

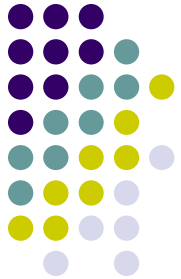


- The RFLPs or SNPs are labeled with fluorescent isotopes.
- Instead of standard nucleotides we introduce fluorescent mononucleotides
- Cy3 fluoresces green
- Cy5 fluoresces red
- These are used to look at genes under different conditions

# Polymerase Chain Reaction

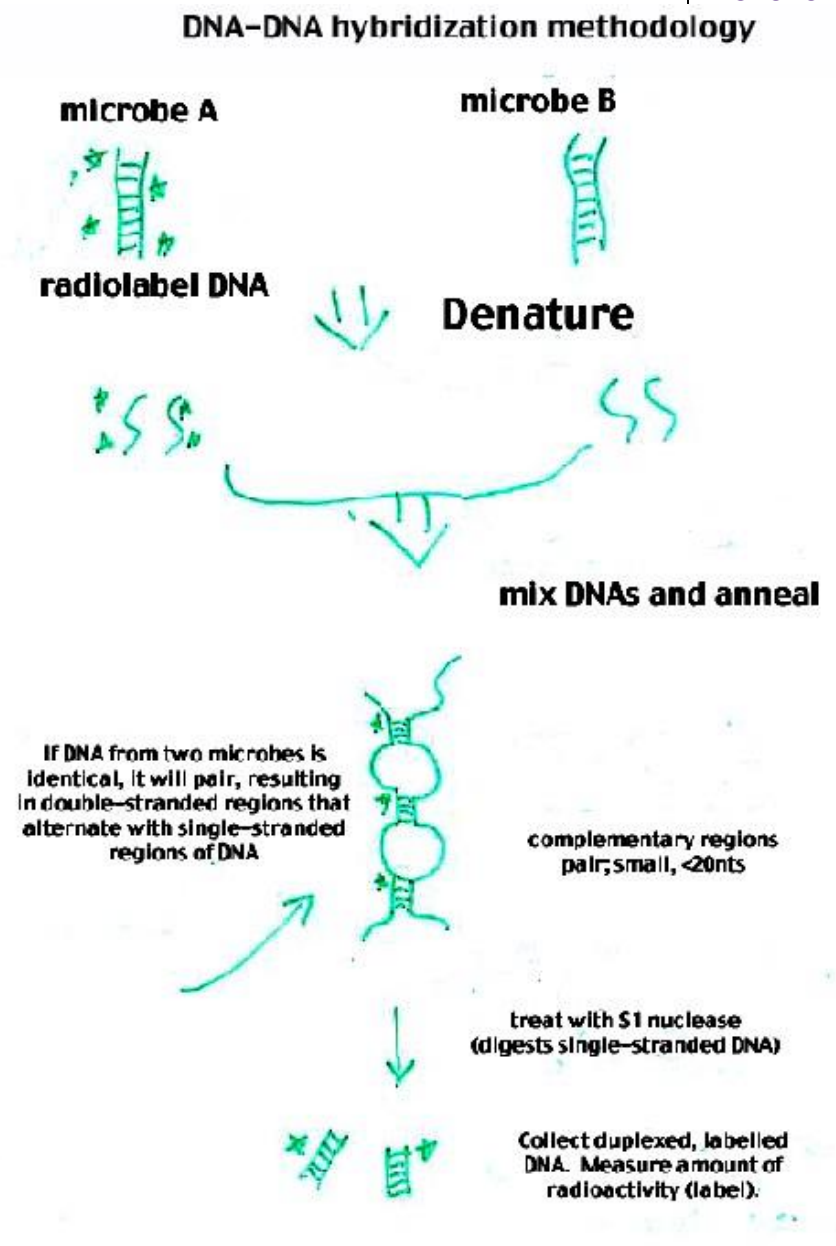


# Now we have a pool of different oligos



- Add them to the chip
- Perform hybridization reaction

# DNA Hybridization



# DNA Hybridization – Output from one pixel (spot)

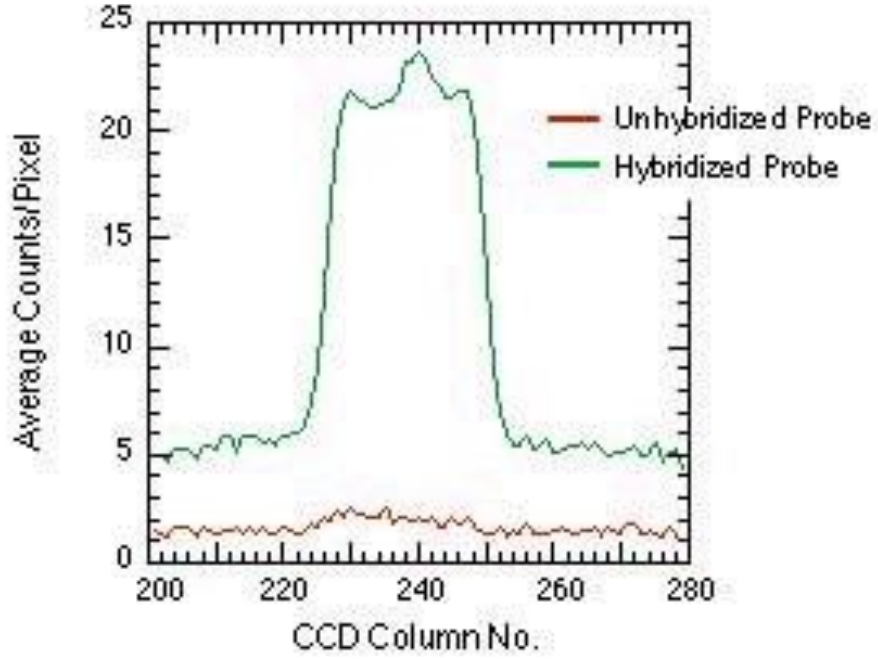
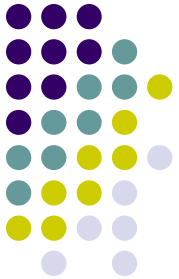



**Figure 5.** Profiles of laser-induced fluorescence in hybridization chamber containing immobilized 16-mer probe, before and after addition of complementary DNA. Both fluorescence images were taken in the presence of PicoGreen solution.

# 4) Recording and analyzing the Image



- Confocal Microarray Scanners are the most commonly used way to detect the fluorescence
- Pixel sizes are 5-20 microns, as probe cells have shrunk from 50 to 25 microns (5 micron probe cells appear attainable giving 4 million cells per 1 $cm^2$ chip)
- The image is normally stored as a *.tiff file and is analyzed for intensity of one or both colors of fluorescence

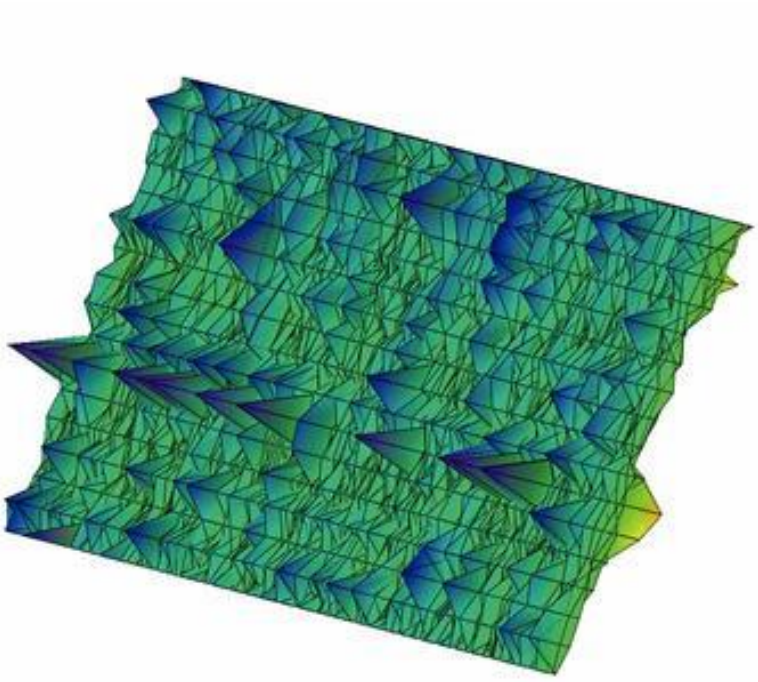
# Image Details

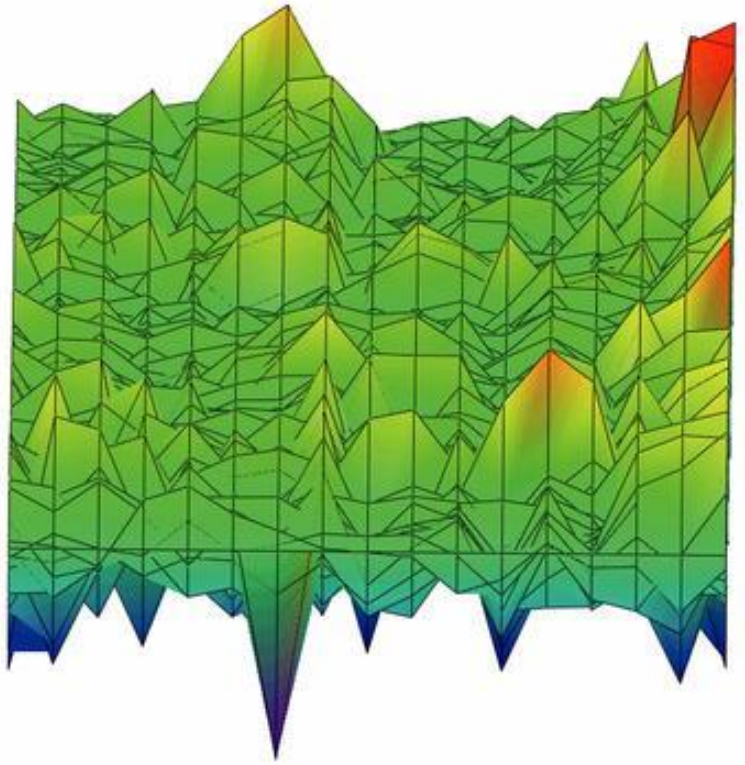
- Aligning a grid between the reader and the plate is not yet accurate so the total image is recorded, using markers in each corner as reference points
- A laser excites each cell and the fluorescence diffraction patterns are recorded
- When two samples (red and green) are used, the spot will appear as the resultant mixture of color intensities (if both are equal the spot will be yellow, while if no signal is noted the spot will be black)

# Image Analysis

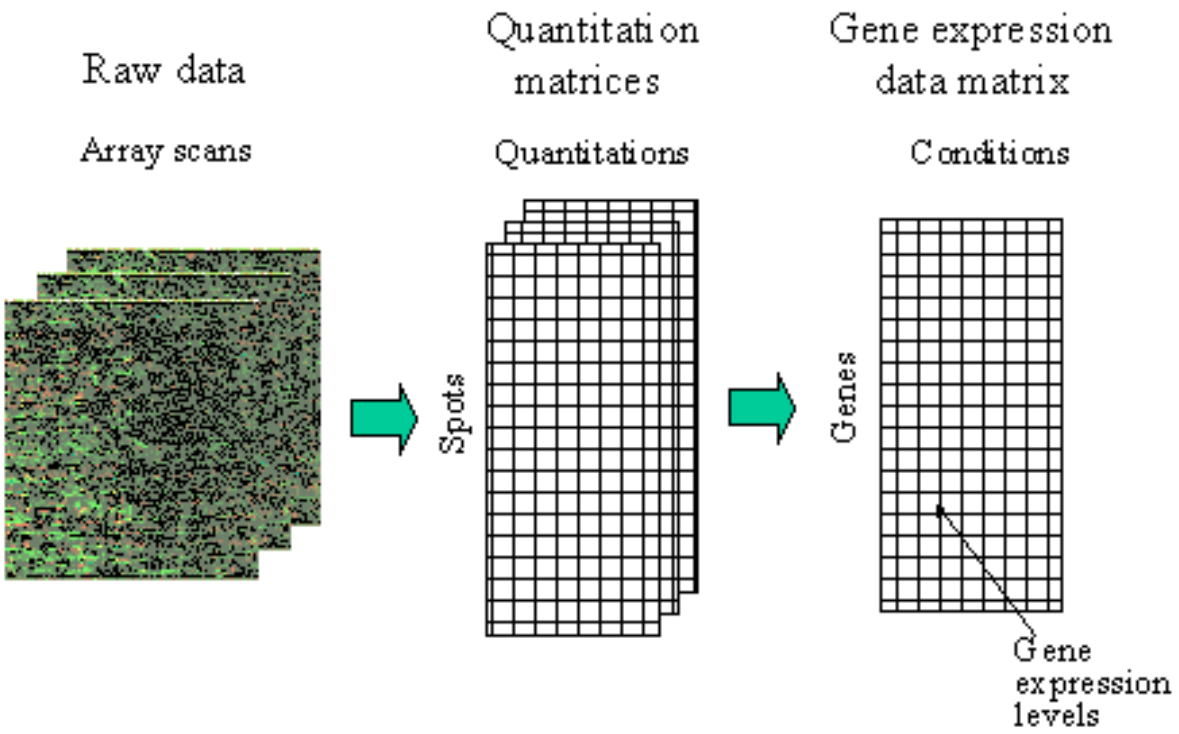
- The intensity of the image is analyzed using a relative measure (~ 300,000:1 S/N is a normal signal of 1 copy of a gene/cell.
- A background noise is recorded for a cell that has no fluorescence and that is subtracted from the signal (S/N ratio). Other normalization methods will be discussed later.
- This ratio can then be normalized
- This can be done in 2 or 3 dimensions depending on the number of samples used and the software program

# Now we have the data



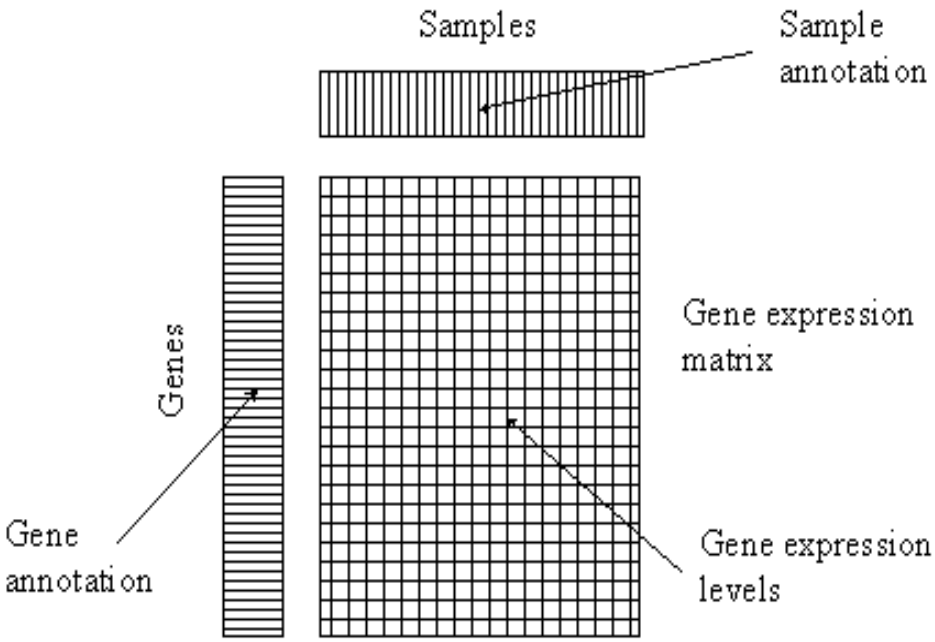


# Image Analysis



# Image Analysis

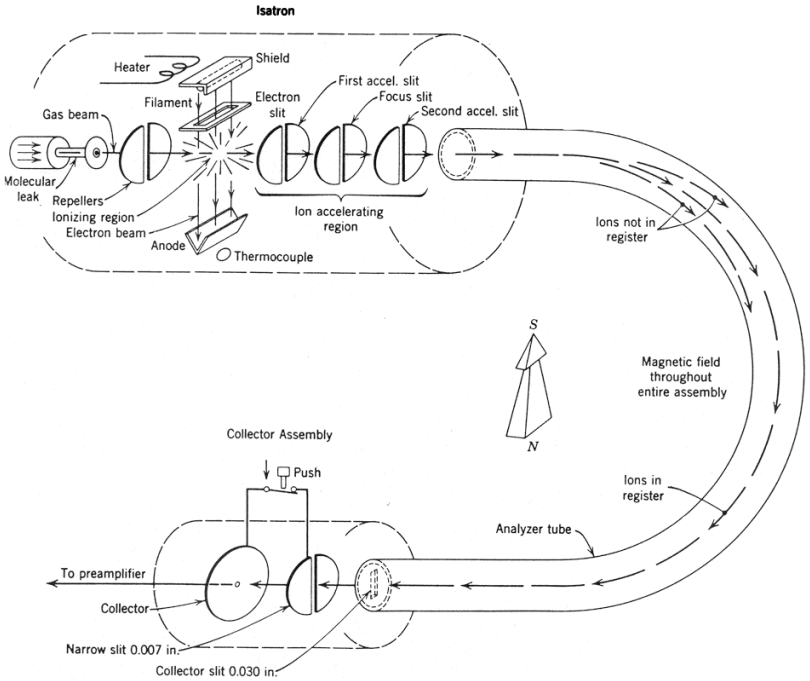
- Genes are measured in rows
- Samples are measured in columns

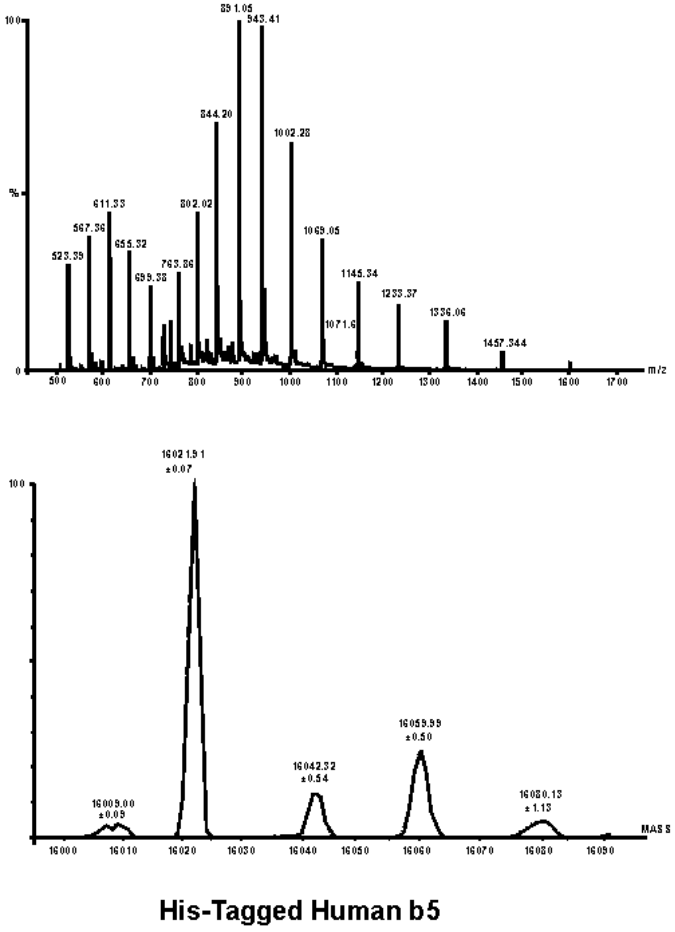


# Spot Analysis

- Spot intensity is measured
- Spot quality can be assessed
  - Absolute intensity in each channel
  - Uniformity of the individual pixel intensities
  - Shape of the spot
- Unfortunately there is currently no standard way of assessing the spot measurement reliability. If experiments have been done in replicates, they can be used to assess the standard errors in addition to the single measurement quality assessments

# Mass spec technology



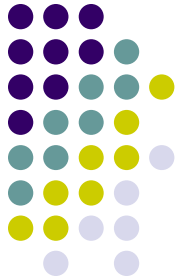

http://www.sequenom.com/

**His-Tagged Human b5**

# Workshop

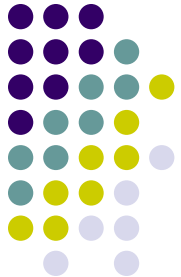
- Three methods for array fabrication were described (photolithography, inkjet synthesis, spotting). What are some advantages and disadvantages of each method?

# Analysis Techniques



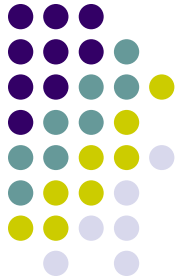
- Summarization and Characterization – overview of data for outliers or deviation detection
- Association – Linkage analysis techniques, association rules for later data mining
- Prediction or Classification Modeling
- Clustering
- Control Time Series and Expectation Ratio Likelihood (ERL)

# Gene Expression Matrix



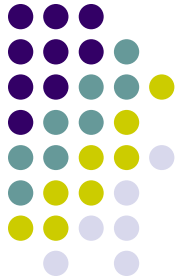
- By measuring transcription levels of genes in an organism under various conditions, at different developmental stages and in different tissues, we can build up `gene expression profiles' which characterize the dynamic functioning of each gene in the genome
- This enables us to understand gene regulation, metabolic and signaling pathways, the genetic mechanisms of disease, and the response to drug treatments
- For instance, if over expression of certain genes is correlated with a certain cancer, we can explore which other conditions affect the expression of these genes and which other genes have similar expression profiles. We can also investigate which compounds (potential drugs) lower the expression level of these genes.

# Gene Expression Matrix Problems



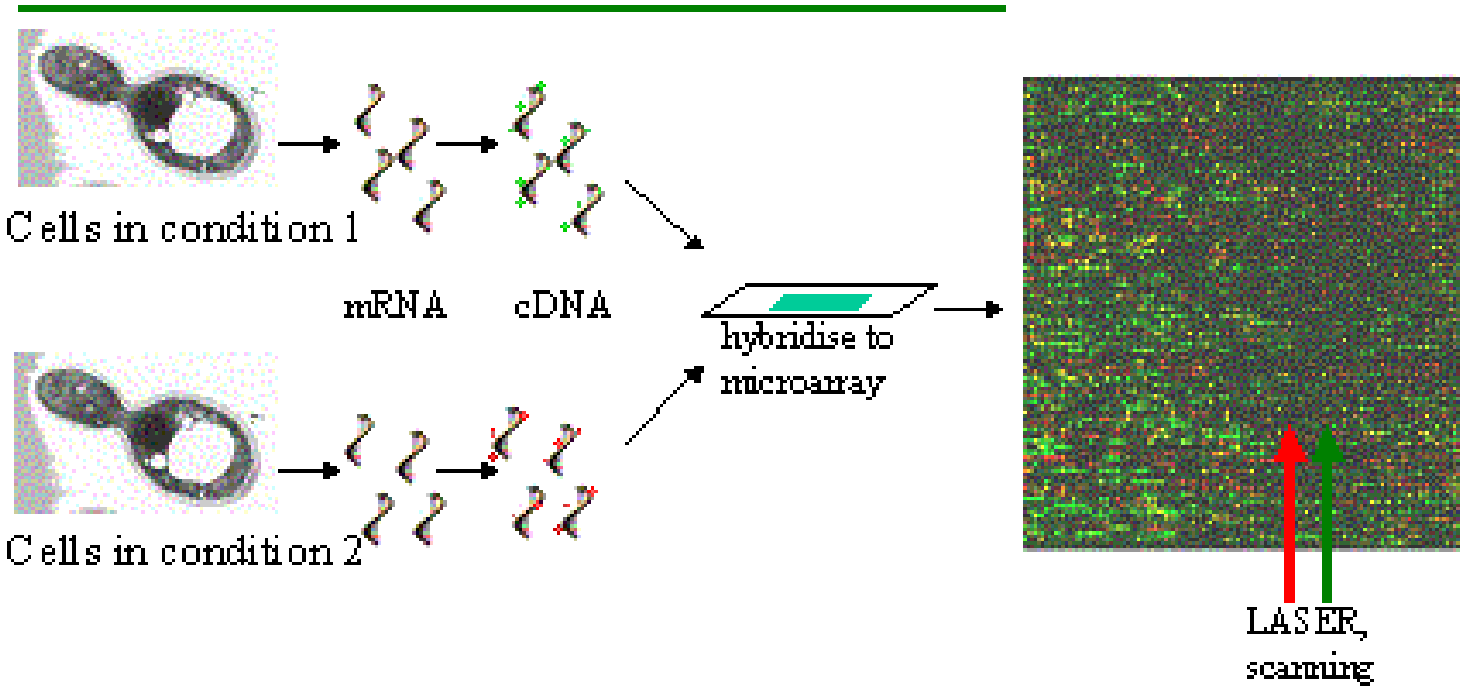
- The genes are located in different spots of the chip and are typically based on EST (expressed sequence tag) sequences and can be the same or slightly different sequence
- This makes linking the EST sequences more difficult

# How Genes are Grouped

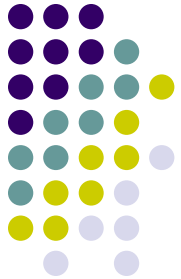


- Sequences can be grouped in many ways
  - Cellular function
  - Phylogeny
  - Regulation (Structural, Exon, Intron, Regulatory)

# Two different conditions



# Clustering Algorithm Application Order

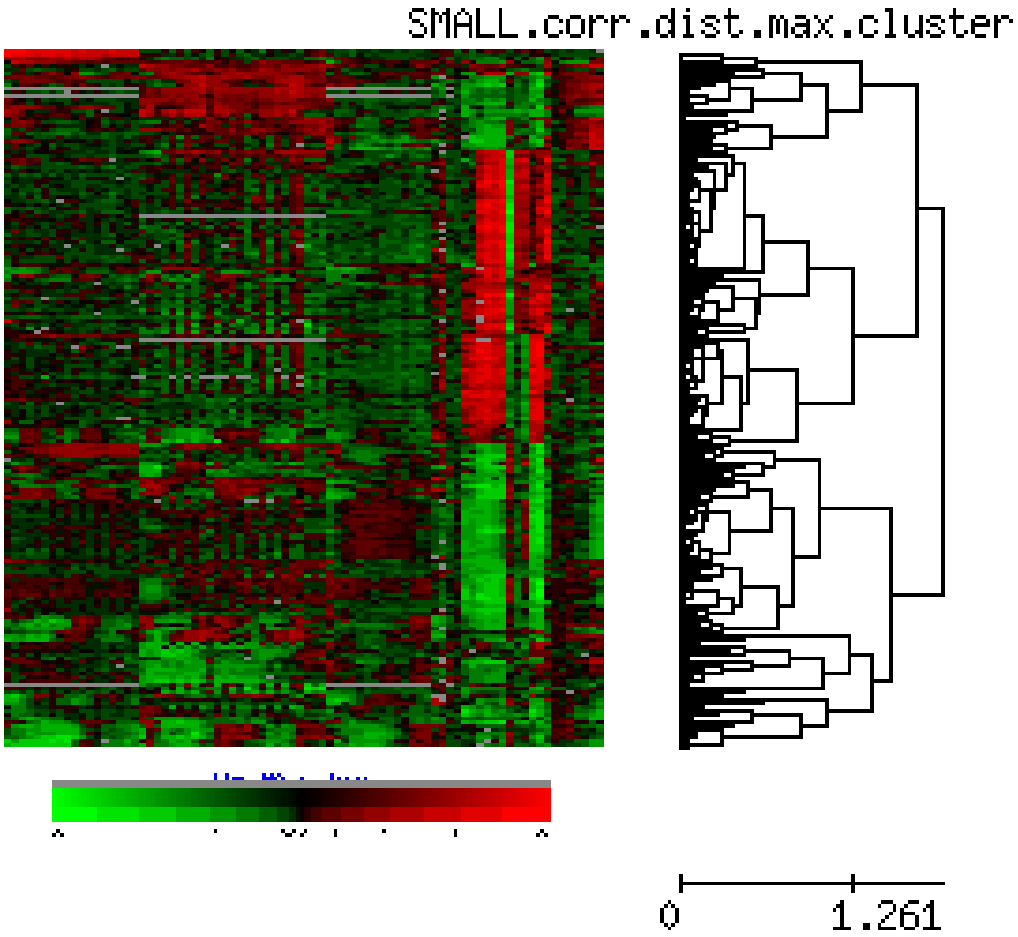


- Used to cluster genes that behave the same in various conditions or differently in the same condition (can also be used for phylogenetic clustering)
- Then within the clusters for inclusion of like sequences.
- Mapping and function are then predicted based on conserved sequences

# Clustering Analysis

- Unsupervised data analysis is expression profile clustering to find groups of co-regulated genes or related samples.
- Supervised approach assumes that for some (or all) profiles we have additional information, such as functional classes for the genes, or diseased/normal states attributed to the samples. We can view this additional information as labels attached to the rows or columns. Having this information, a typical task is to build a classifier able to predict the labels from the expression profile

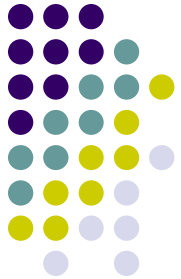
# Clustering



# Time Courses

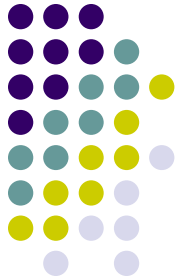
- Control Time Series and ERL (Expectation Ratio Likelihood) are best for time dependent diseases like degenerative pathological conditions.
- However, gene expression is relative to the time point and must be recorded for all time points on the same chip or standardized between successive chips

# Problems with Data



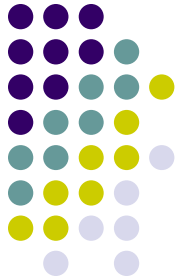
- Sensitivity
- Specificity
- Background Noise to Sample Ratio
  - Is a ratio even an acceptable way to quantify this as of now non-quantifiable amount of signal
- Variation in Hits
- Standardization – Standardizing the same signal on different chips
- Normalization – Normalizing the hits on a single gene chip

# Hit Variation



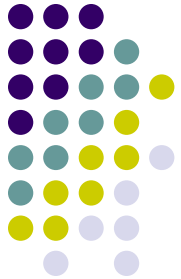
- From blot to blot, what increase in signal represents a real increase in homologous binding?
- Transformations are done
  - Log of the expression values to equalize variability and to normalize the distribution is the most common transformation
  - Taking a standard positive value and dividing all like signals by that standard after the S/N ratio has been calculated is another way
  - Internal controls perhaps

# Normalization Issues



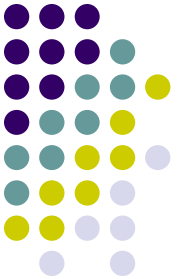
- Unequal RNA/cDNA used (PCR primer issues)
- mRNA does not mean protein concentration
- Different labeling concentrations
- Hybridization techniques leading to different rates of bonding
- Gene expression data have meaning only in the context of the particular biological sample and the exact conditions under which the samples were taken. For instance, if we are interested in finding out how different cell types react to treatments with various chemical compounds, we must record unambiguous information about the cell types and compounds used in the experiments

# Normalization Techniques



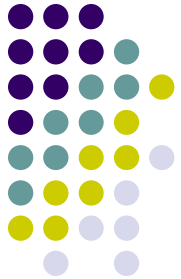
- Use housekeeping genes as standards – too much fluctuation in biological systems
- Median of all signal intensities – good approximation
- Combination of all samples – most accurate

# DNA Arrays – Just the First Step



- Once a high-density chip is used:
  - More focused chips can be used
  - Northern Blots
  - QRT-PCR
  - RDA (Representational Difference Analysis) Subtraction – Subtractive Hybridization

# To recapitulate on why use a gene chip

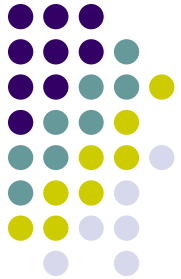


- Cheap way to look a thousands of genes
- Fast way to look a thousands of genes
- Can observe sequences and patterns of expression simultaneously depending on analysis used
- Efficient screening tool for pharmacogenetics

# How to Choose a DNA Array

- The choice of DNA chip depends on several parameters, including cost, density, accuracy, and the type of DNA to be immobilized on the surface.
- The first distinction is whether the chips contain immobilized cDNAs or shorter oligonucleotide sequences. The former must be spotted on the chips as complete molecules, but oligos can either be spotted or synthesized on the surface of a chip.
- The final distinction is whether the user makes or purchases the chip. With homemade systems, researchers are limited to spotting samples.

# Reverse Engineering

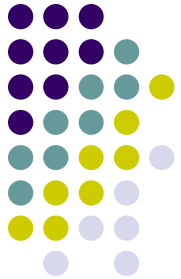


- DNA Arrays are like reverse Engineering
  - We are starting with the final product and working back to the functional components
  - This approach is difficult but easier than proteomics – predicting the function of a protein based on its 3-D structure and amino acid sequence

# Reverse Engineering

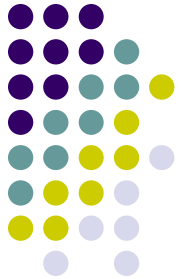
- Reverse engineering of gene regulatory networks is based on the hypothesis that genes that have similar expression profiles (i.e., similar rows in the gene expression matrix) should also have similar regulation mechanisms as there must be a reason why their expression is similar under a variety of conditions.
- If we cluster the genes by similarities in their expression profiles and take sets of promoter sequences from genes in such clusters, some of these sets of sequences may contain a ‘signal’ as a specific sequence pattern such as a particular substring, which is relevant to regulation of these genes yielding putative regulatory elements (data mining)

# Large Questions Left Unanswered

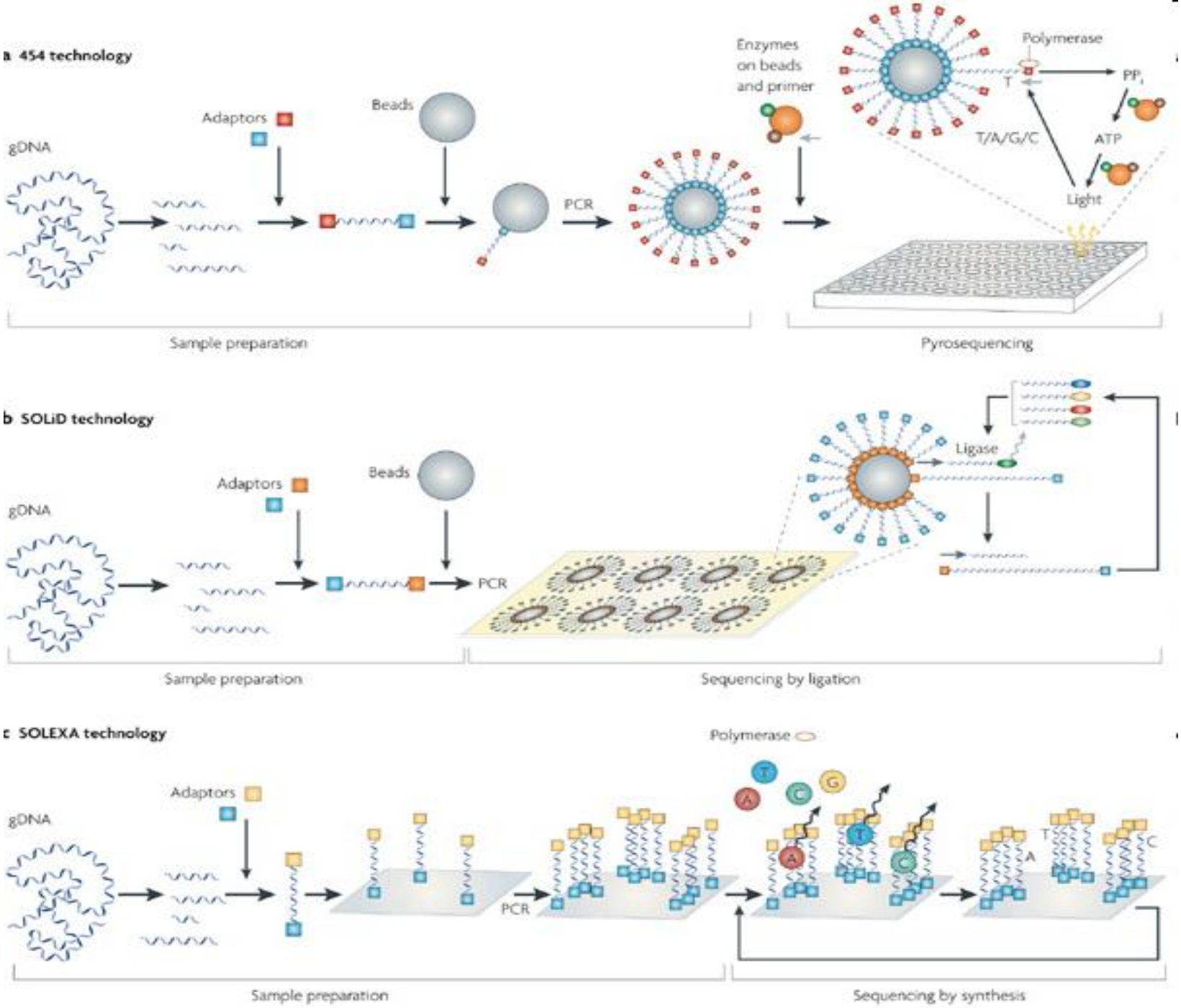


- Little has been published on how to use the reliability of gene expression measurements by combining the information about the spot image in each channel and the replicate images.
- The value of microarray-based gene expression measurements would be considerably higher if reliability and limitations of particular microarray platforms for particular kinds of measurements, as well as cross-platform comparison and normalization, were studied and published.

# Large Questions Left Unanswered



- No established standards for microarray experiments and how the raw data should be processed
- No standard measurement units for gene expression levels
- With the lack of such standards, the information about how exactly the gene expression data matrix was obtained should be kept in the database, if the data are to be properly interpreted later. Many countries are storing the data for future standardization.


a 454 technology
gDNA
Adaptors
Beads
PCR
Enzymes on beads and primer
Polymerase
T
PPi
ATP
T/A/G/C
Light
Sample preparation
Pyrosequencing
b SOLiD technology
gDNA
Adaptors
Beads
PCR
Ligase
Sample preparation
Sequencing by ligation
c SOLEXA technology
gDNA
Adaptors
PCR
Polymerase
T
A
C
G
Sample preparation
Sequencing by synthesis