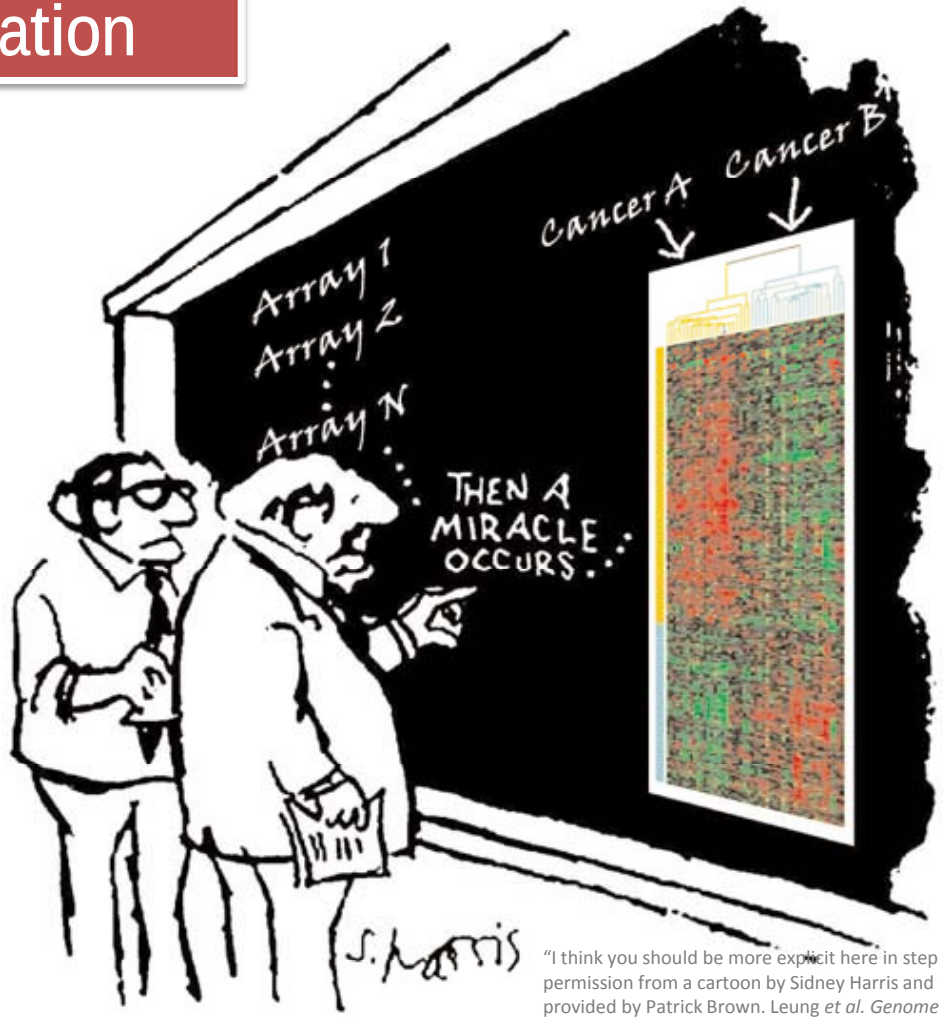
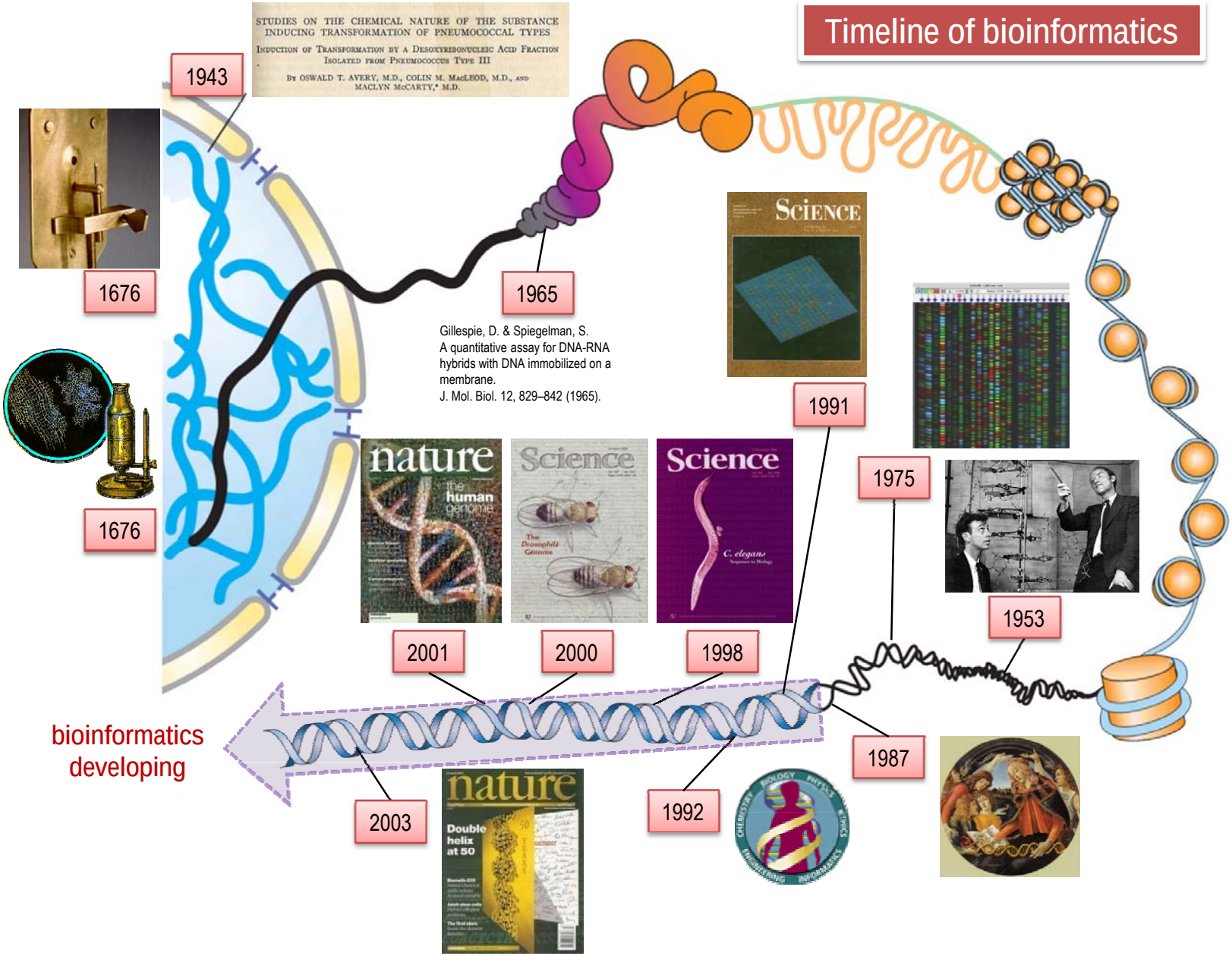
Microarray and its application



"I think you should be more explicit here in step two." Modified with permission from a cartoon by Sidney Harris and from an image provided by Patrick Brown. Leung *et al. Genome Biology* 2001 **2**:reports4021.1 doi:10.1186/gb-2001-2-9-reports4021

Timeline of bioinformatics

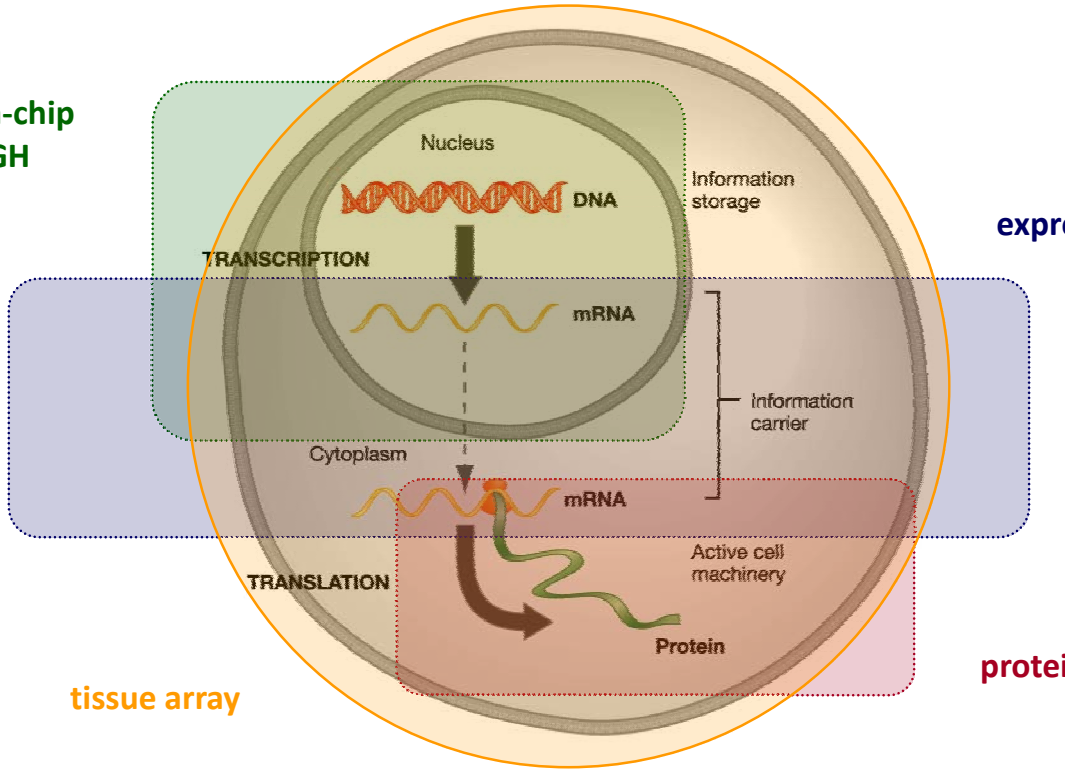


改變二十世紀的大事



The Central Dogma

DNA → Transcription → RNA → Translation → Protein



CHIP-on-chip
array CGH

expression array

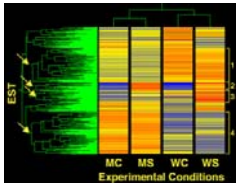
tissue array

protein chip



基因體學

Genomics is the study of *an* organism's entire genome. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts. The field also includes studies of intragenomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome. In contrast, the investigation of single genes, their functions and roles, something very common in today's medical and biological research, and a primary focus of molecular biology, does not fall into the definition of genomics, unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genome's networks

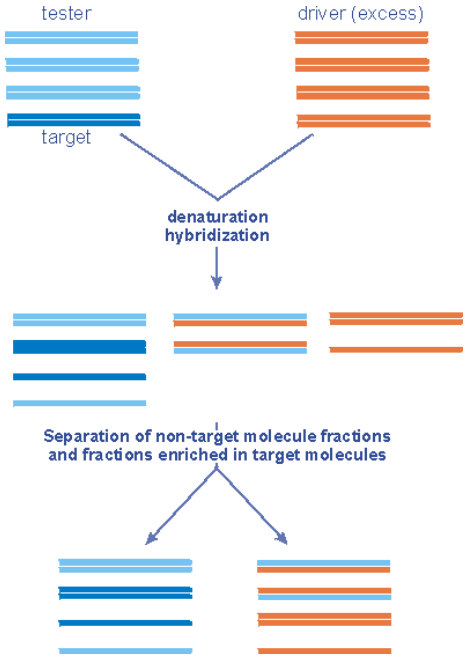


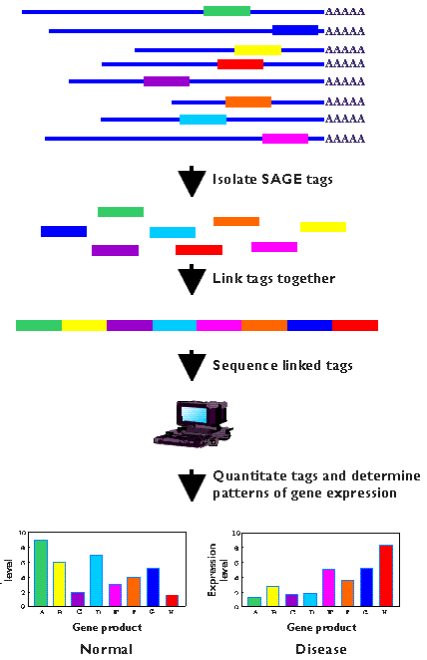
轉錄體學

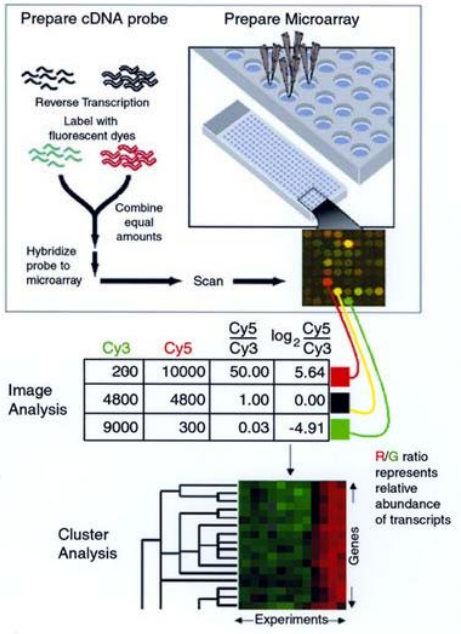
The transcriptome is the set of *all* messenger RNA (mRNA) molecules, or "transcripts", produced in one or a population of cells. The term can be applied to the total set of transcripts in a given organism, or to the specific subset of transcripts present in a particular cell type. Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome can vary with external environmental conditions. Because it includes all mRNA transcripts in the cell, the transcriptome reflects the genes that are being actively expressed at any given time, with the exception of mRNA degradation phenomena such as transcriptional attenuation. The study of transcriptomics examines the expression level of mRNAs in a given cell population, often using high-throughput techniques based on DNA microarray technology.

Evaluate global gene expression

- Differential displaying
- Suppressive subtractive hybridization (SSH)
- Sequencing of expressed sequence tags (EST), serial analysis gene expression (SAGE & Long SAGE)
- Hybridization to microarrays







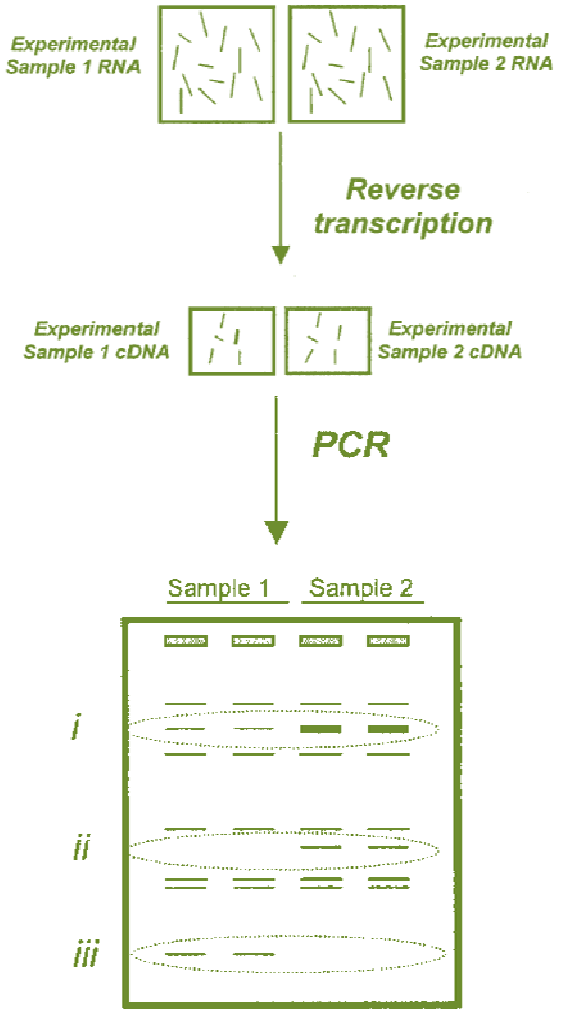
Differential displaying

Several significant limitations of the original protocol were:

- the requirement of large quantities of mRNA
- a bias against the identification of rare transcripts.
- not a quantitative method with a high rate of false positives, or
- gene fragments that seem to be differentially expressed as an artifact

Advantages

- ✓ to compare multiple experimental samples simultaneously
- ✓ to identify genes that are either up or down-regulated in one sample relative to another

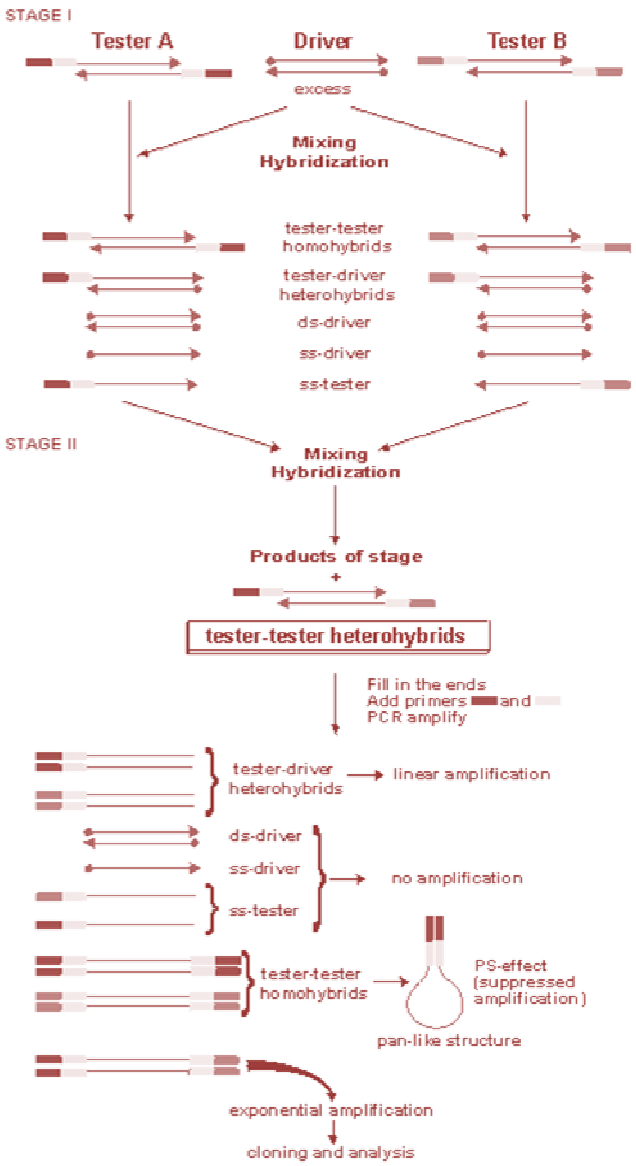


Subtractive hybridization

Subtractive hybridization (or subtraction) is a method for rapid isolation of differentially distributed nucleic acids (differentially expressed, differentially present, or differentially arranged in two or more different sources: different cells, cell populations or cell types, different tissues, disease, or development stages). The **Suppressive subtractive hybridization (SSH)** method is designed to **selectively amplify differentially expressed transcripts** while suppressing the amplification of abundant transcripts, thus eliminating the need to separate single- and double-stranded molecules.

Limitations

- only to pair-wise treatment comparisons
- does not provide a quantitative measure of expression differences

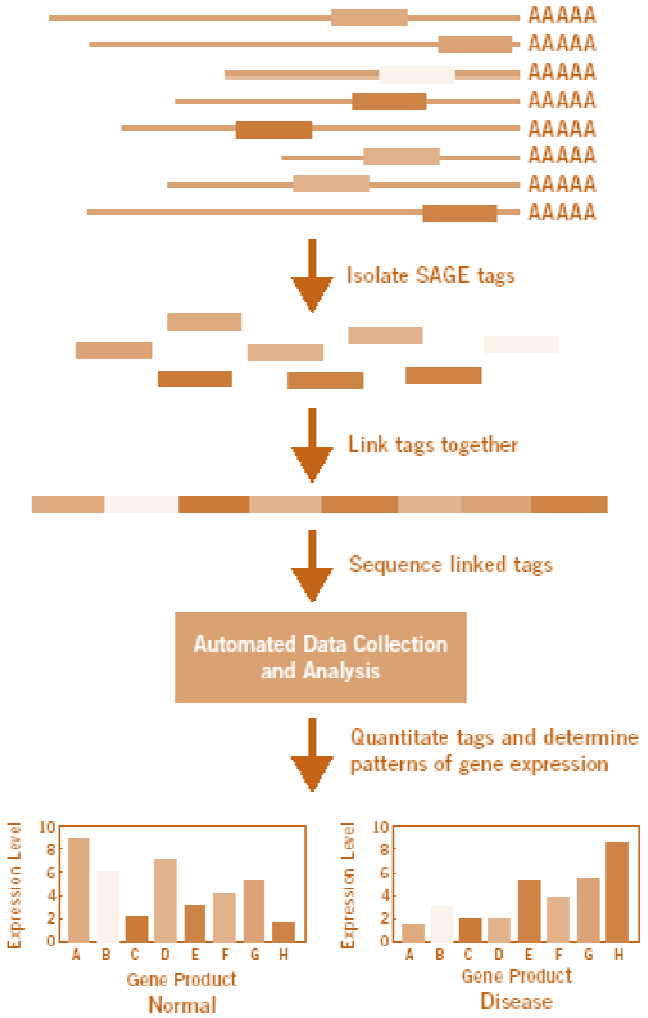


EST and SAGE

- ✓ Expressed sequence tag (EST) sequencing are generated by randomly picking clones from a cDNA library and performing a single sequencing reaction to produce 300 to 500 bps of sequence per clone. Differences in gene expression may be identified by counting the number of times a particular sequence appears in EST libraries of genes from different sources.
- ✓ Serial analysis of gene expression (SAGE) a technique for the analysis of gene expression that is essentially an accelerated version of EST sequencing.
  SAGE: 10-14 bps tags created by type IIS restriction endonuclease Nla III.
  Super SAGE: > 25 bps tag, created by type III restriction enzyme EcoP15I.

Advantages:

- ➢ SAGE data are quantitative and cumulative.
- ➢ Accurate, quantitative transcript profiles describing the abundance of all genes expressed in a cell or tissue are generated by SAGE, if sufficient sequencing is completed.

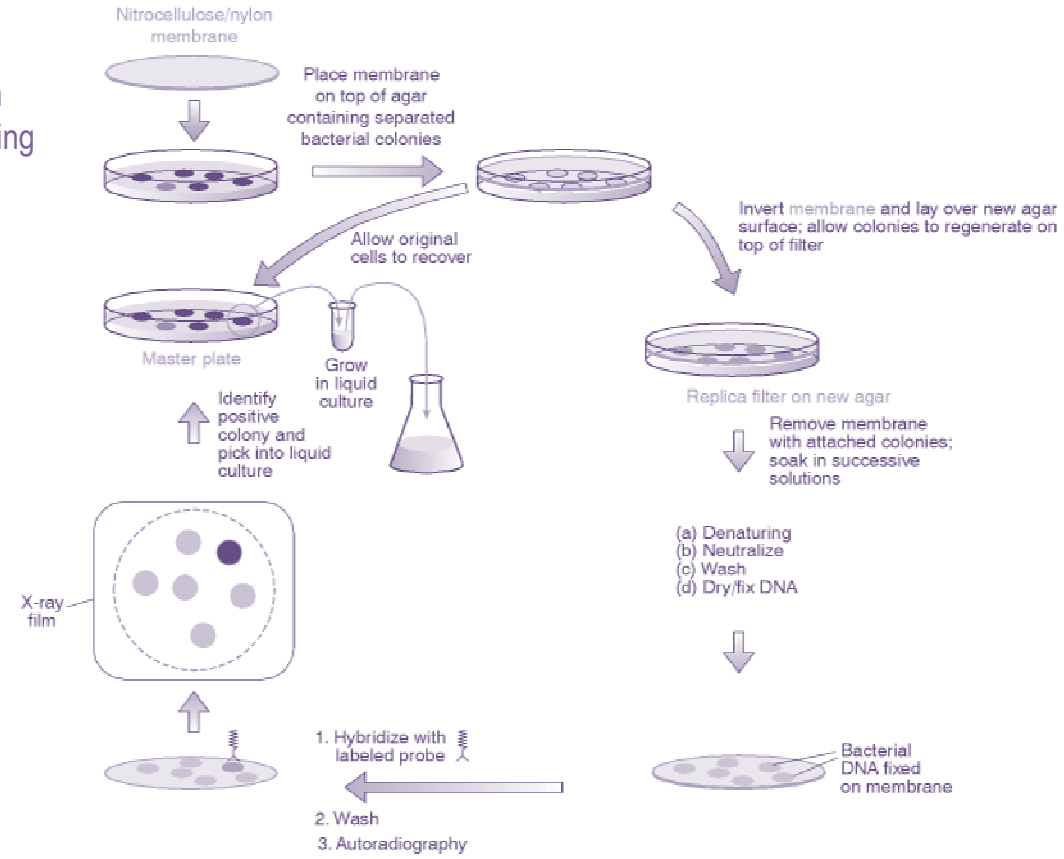


Microarray hybridization

The analysis of array data may be divided into three components:

1) identification and quantification of hybridization intensities,
2) visualization of data, and
3) clustering techniques

Based on the assumption that genes with related functions are coregulated, clustering of microarray data becomes a powerful method to assign putative functional classifications to novel genes.



The Cancer Genome Anatomy Project

CGAP HOW TO | Genes | Chromosomes | Tissues | SAGE Genie | RNAi | Pathways | Tools

CGAP Info

- Educational Resources
- Slide Tour
- Team Members
- References

CGAP Data

- Download

Quick Links:

- ICG
- NCI Home
- NCICB Home
- NCBI Home
- OCG

NATIONAL CANCER INSTITUTE

The CANCER GENOME ANATOMY PROJECT

The goal of the NCI's Cancer Genome Anatomy Project is to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. By collaborating with scientists worldwide, CGAP seeks to increase its scientific expertise and expand its databases for the benefit of all cancer researchers.

The CGAP Web Site

Interconnected modules provide access to all CGAP data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a fraction of the time it once took in the laboratory.

Genes
Gene information, clone resources, SNP500Cancer, GAI, and transcriptome analysis

Chromosomes
FISH-mapped BAC clones, SNP500Cancer, and the Mitelman database of chromosome aberrations

Tissues
cDNA library information, methods, and EST-based gene expression analysis

SAGE Genie
Analysis of gene expression using **long** and **short** SAGE tag data for both **human** and **mouse**

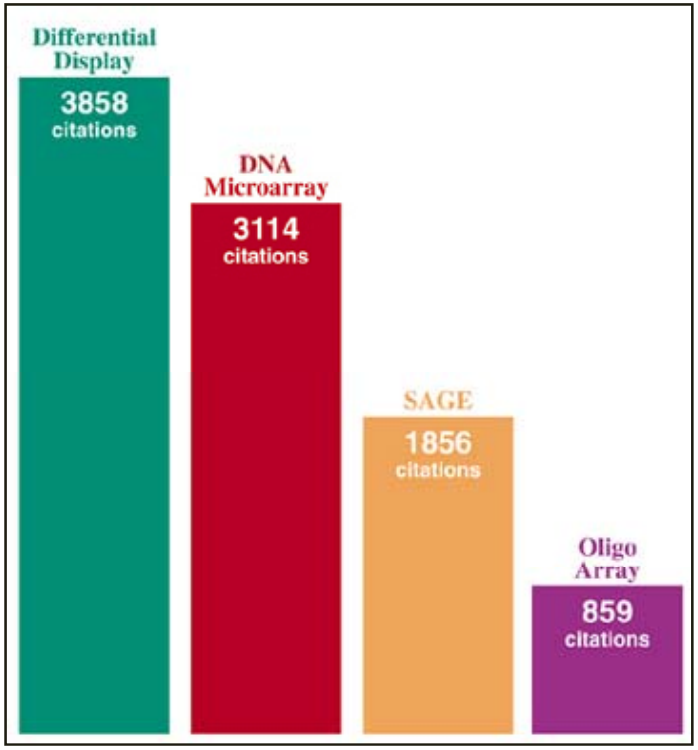
Pathways
Diagrams of biological pathways and protein complexes, with links to genetic resources for each known protein

Tools
Direct access to all analytic and data mining tools developed for the project

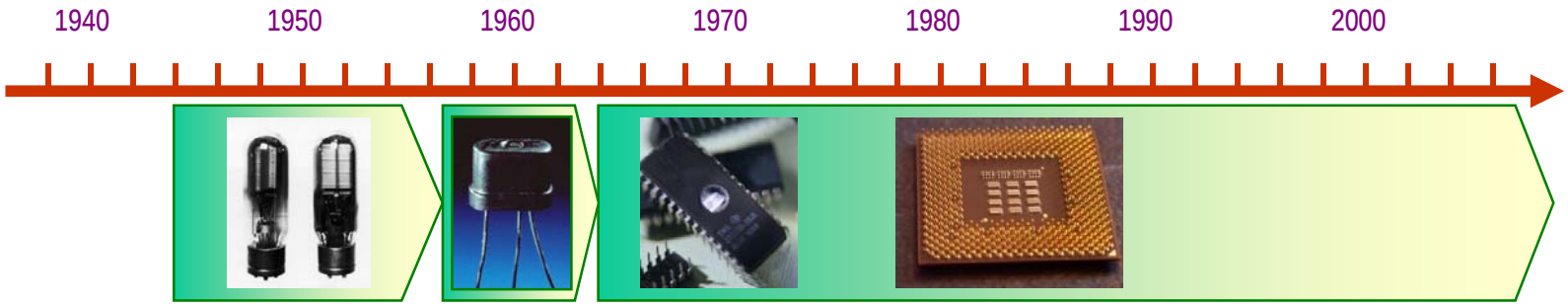
RNAi NEW
RNA-interference constructs, targeted specifically against cancer relevant genes. New addition: Validated set of shRNAs targeted to cancer genes.

http://cgap.nci.nih.gov/

No single technique is optimal for all applications, and the optimal strategy for specific situations may prove to be creative combinations of several techniques.



Generations of Computer



Vacuum tube

Transistor

Small scale integration - 1965 on
- Up to 100 devices on a chip

Medium scale integration - to 1971
- 100-3,000 devices on a chip

Large scale integration - 1971-1977
- 3 – 100K devices on a chip, semi conductor memory

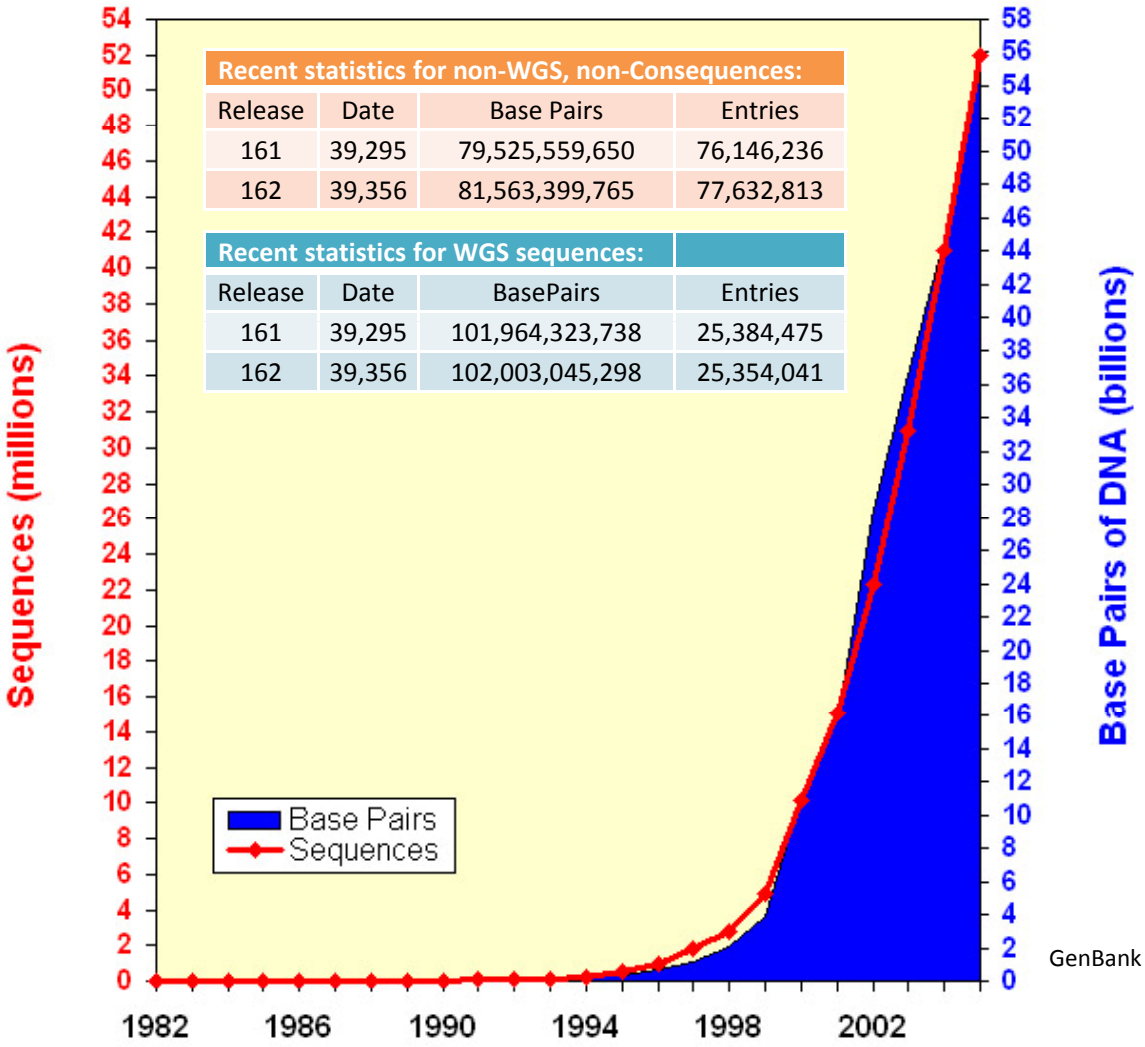
Very large scale integration - 1978 to date
- 100K – 100M devices on a chip

Ultra large scale integration
- Over 100M devices on a chip



Growth of GenBank

(1982 - 2005)



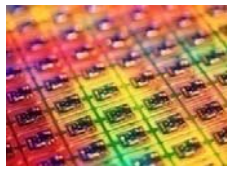
Recent statistics for non-WGS, non-Consequences:

Release	Date	Base Pairs	Entries
161	39,295	79,525,559,650	76,146,236
162	39,356	81,563,399,765	77,632,813

Recent statistics for WGS sequences:

Release	Date	BasePairs	Entries
161	39,295	101,964,323,738	25,384,475
162	39,356	102,003,045,298	25,354,041

GenBank Statistics



陣列

In computer science an array is a data *structure* consisting of a group of *elements* that are accessed by indexing. In most programming languages each element has the same data type and the array occupies a contiguous area of storage. Most programming languages have a built-in array data type.

Some programming languages support array programming (e.g., APL, newer versions of Fortran) which generalises operations and functions to work transparently over arrays as they do with scalars, instead of requiring looping over array members.

Multi-dimensional arrays are accessed using more than one index: one for each dimension.

Arrays can be classified as fixed-sized arrays (sometimes known as static arrays) whose size cannot change once their storage has been allocated, or dynamic arrays, which can be resized.



微

Micro is an English prefix of Greek origin that refers to an object as being *smaller* than an object or scale of focus, in contrast with macro.

微陣列

Different kinds of *biological assays* are called microarrays:

- DNA microarrays, such as cDNA microarrays and oligonucleotide microarrays
- Protein microarrays
- Tissue microarrays
- Transfection microarrays (also called cell microarrays)
- Chemical Compound Microarrays
- Antibody microarrays
- Gene Chip Analysis

大量快速篩選

High-throughput screening (HTS), is a *method* for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry.

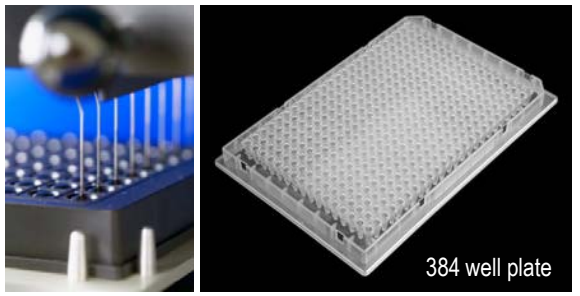












High Throughput Technologies: *The future of Molecular Diagnosis*

High Throughput Technologies (HTTs) are developed to produce huge amount of information from genome projects, but they have clear potential in mass screening and diagnostics of Infectious Diseases. The application of HTTs may revolutionize diagnostic techniques and replacing multiple individual assays.

Genomics
CGH, LOH

Microarray,
EST , SAGE

Proteomics,
3D Modeling

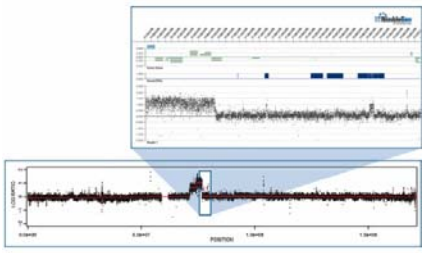


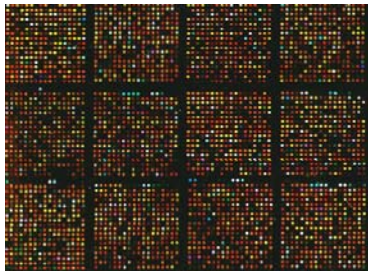
Nucleotide Sequencing

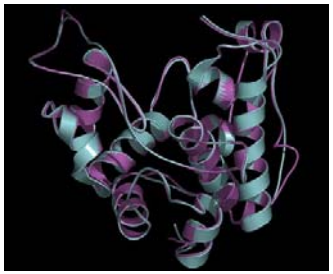


Amino Acid Sequencing

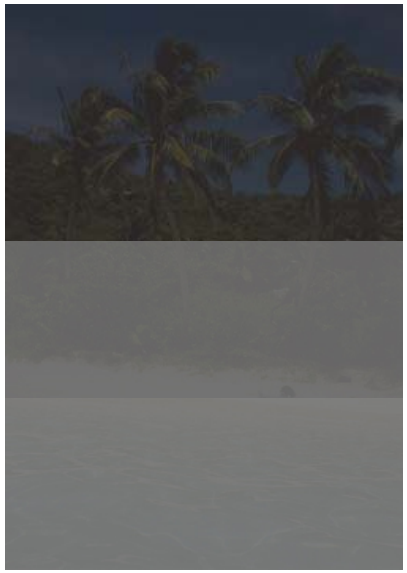




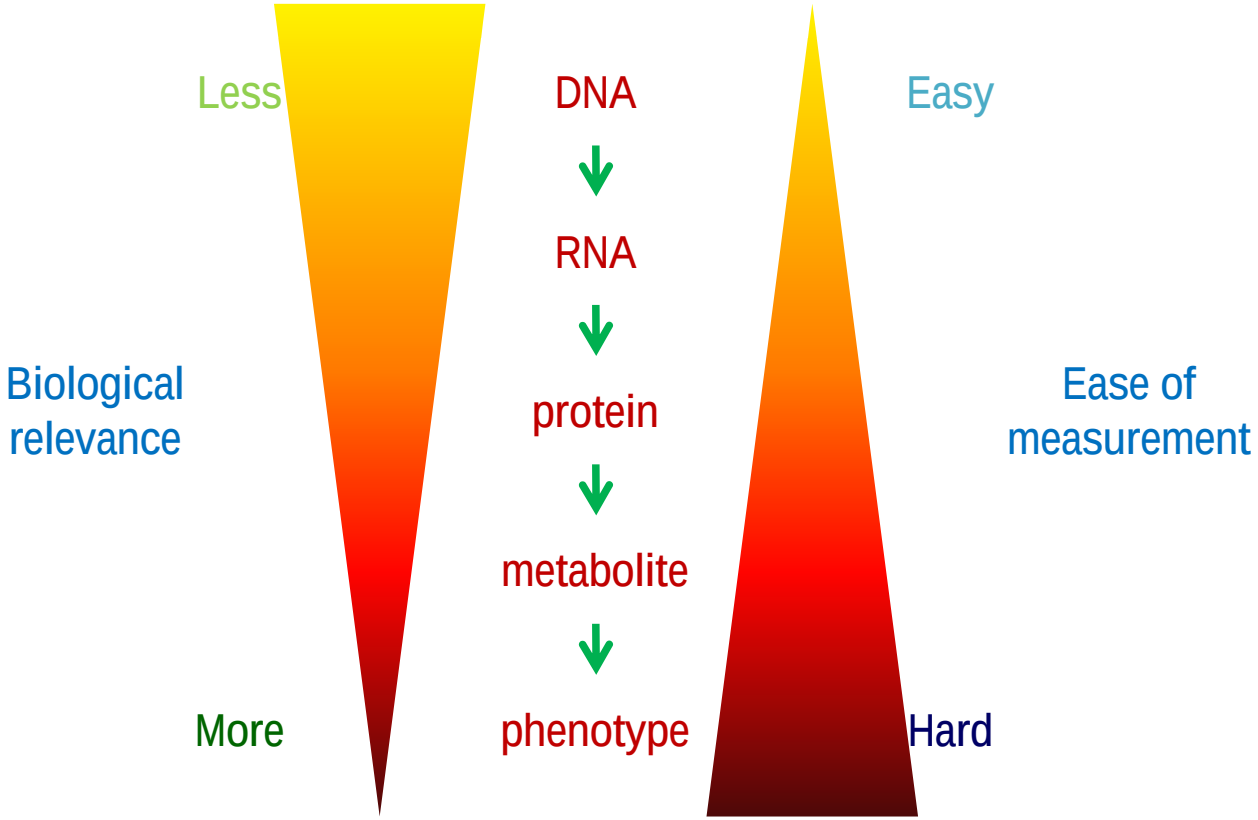




Digital Image Analogy

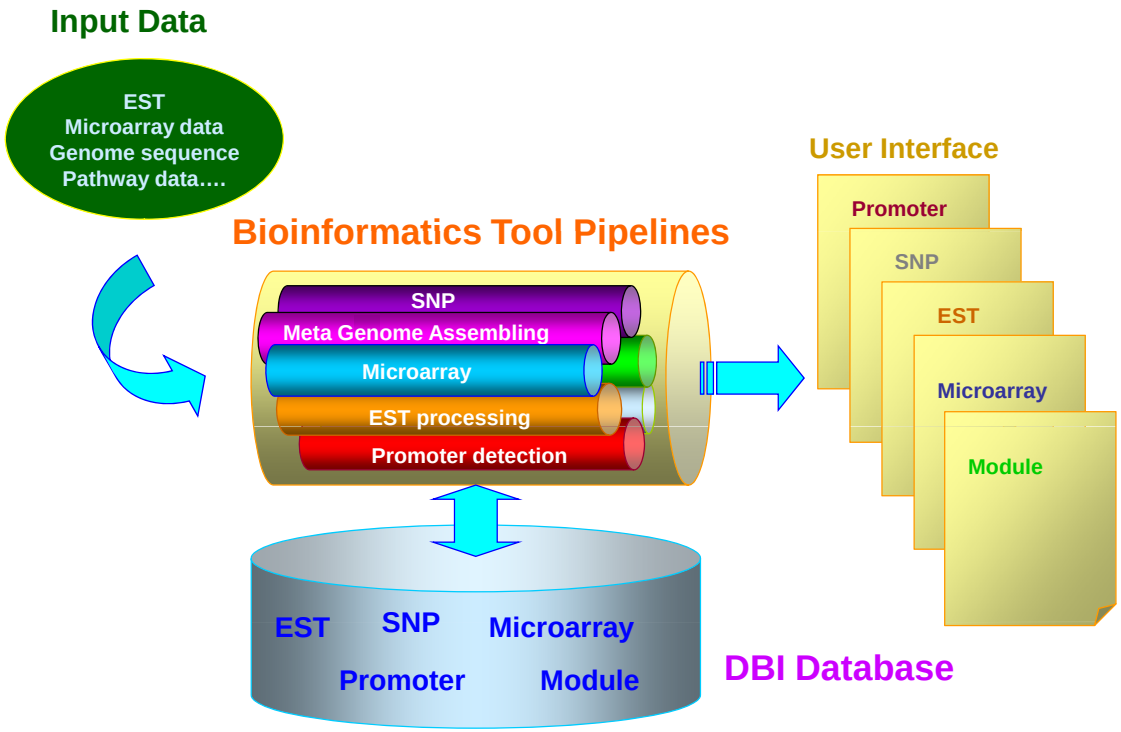
	Northern blot	Customized	Commercial
Expression			
CGH (LOH)			
ChIP, CGR, Exp Tiling, Methylation, …			
	Low density ▪ low content	High density ▪ high content Short probes ▪ limited contrast ▪ limited channels	High density ▪ high content Long, isothermal probes ▪ high contrast ▪ multiple channels

Problems and potentials for high-throughput analysis



Bioinformatics Issues in Microarray Experiment

1. How to manage data ?
2. Which data should be stored ?
3. Validated database or just data storage ?
4. How to integrate in-house data with public repositories ?
5. Which databases ?
6. Local mirror DBs versus web link ?
7. How to analyze and visualize data ?
8. Which analysis should be standardized ?
9. Buy versus develop ?
10. How to integrated microarray data with other omics data ?



Microarray Historical Perspective



1st Sales







1999 | 2000 | 2001 | 2002 | 2003 | 2004











>1000 Array Designs

Xeotron purchased by Invitrogen, Febit bankruptcy, NimbleGen acquires Light Biology

$13M financing in Oct '04, bringing total to >$40M: Investors include Baird, Cargill, Tactics II, Venture Investors, Topspin, SWIB, WARF

Organization → Technology and Applications Development → Commercialization →

Part I: Observing the microarray

Experimental methodology

Data analysis

Part II: The application of microarray

DNA microarrays, such as cDNA and oligonucleotide

Protein microarrays

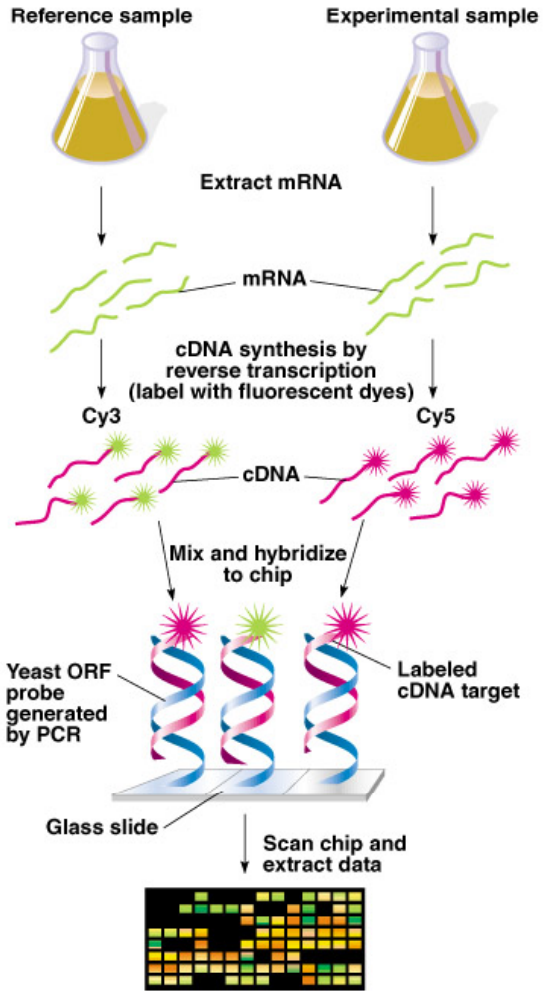
Tissue microarrays

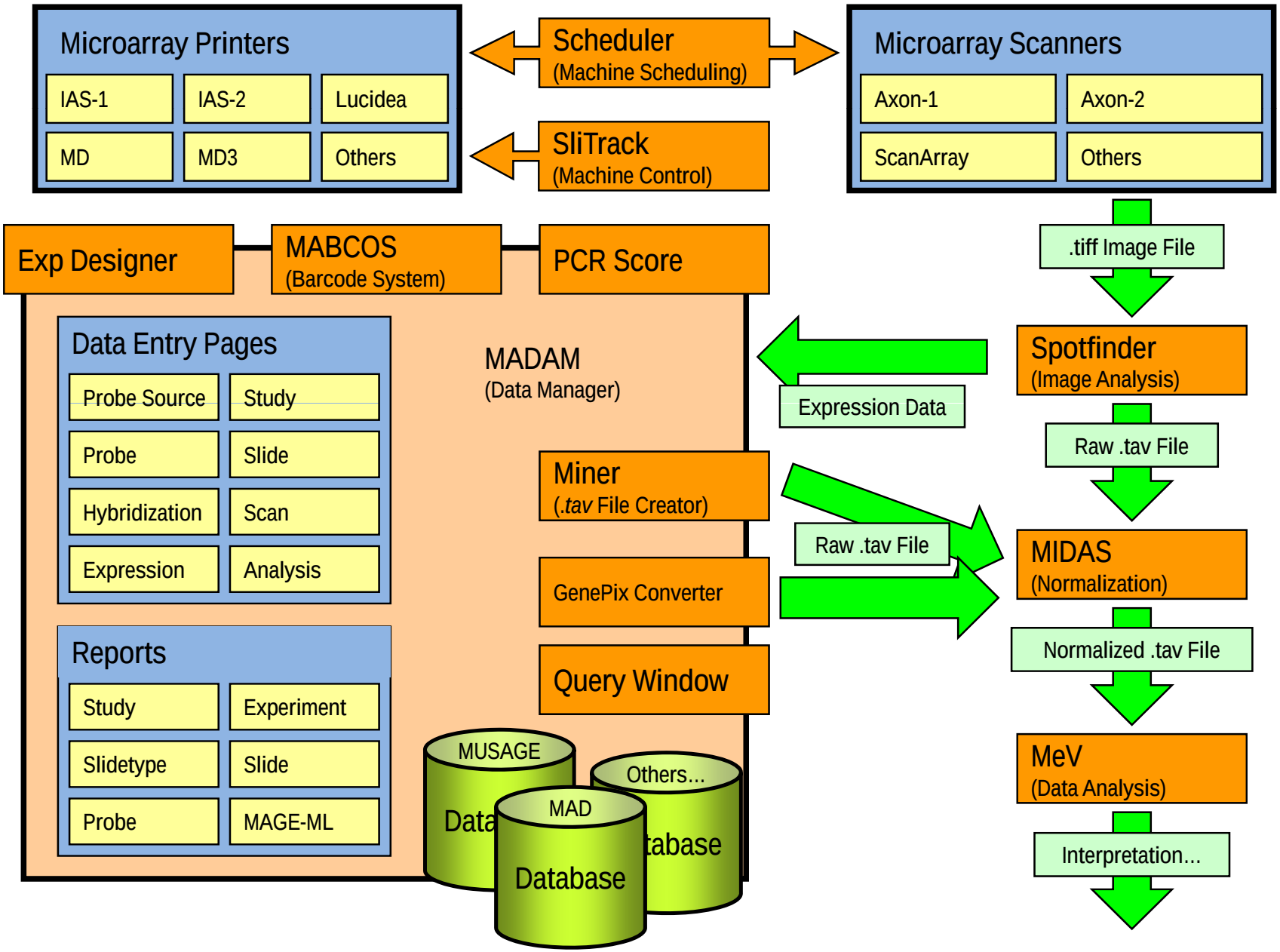
Transfection microarrays (also called cell microarrays)

Chemical Compound Microarrays

Antibody microarrays

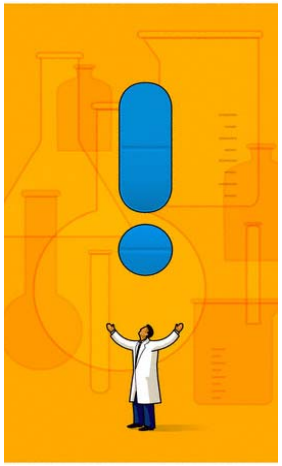
Gene Chip Analysis




Microarray Printers
IAS-1
IAS-2
Lucidea
MD
MD3
Others
Scheduler
(Machine Scheduling)
SliTrack
(Machine Control)
Microarray Scanners
Axon-1
Axon-2
ScanArray
Others
Exp Designer
MABCOS
(Barcode System)
PCR Score
.tiff Image File
Spotfinder
(Image Analysis)
Raw .tav File
Data Entry Pages
Probe Source
Study
Probe
Slide
Hybridization
Scan
Expression
Analysis
MADAM
(Data Manager)
Expression Data
Miner
(.tav File Creator)
Raw .tav File
MIDAS
(Normalization)
GenePix Converter
Normalized .tav File
Reports
Study
Experiment
Slidetype
Slide
Probe
MAGE-ML
Query Window
MeV
(Data Analysis)
MUSAGE
Data
MAD
Database
Others...
tabase
Interpretation...

Experimental methodology

1 **Experimental Design**



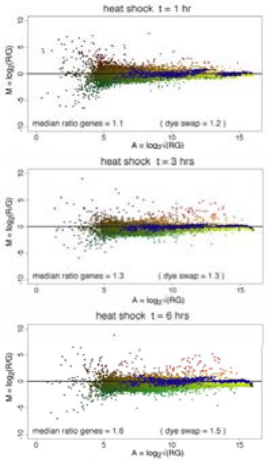
2 **Microarray Experiments**



3 **Image Processing**



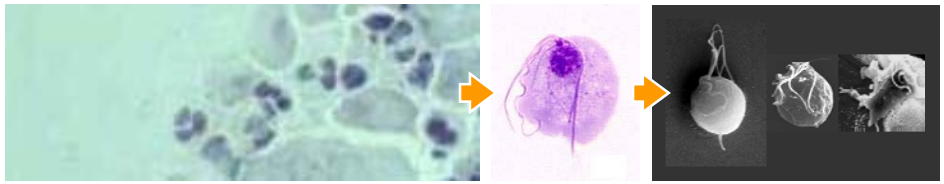
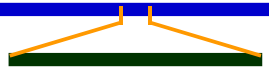
4 **Data Normalization**



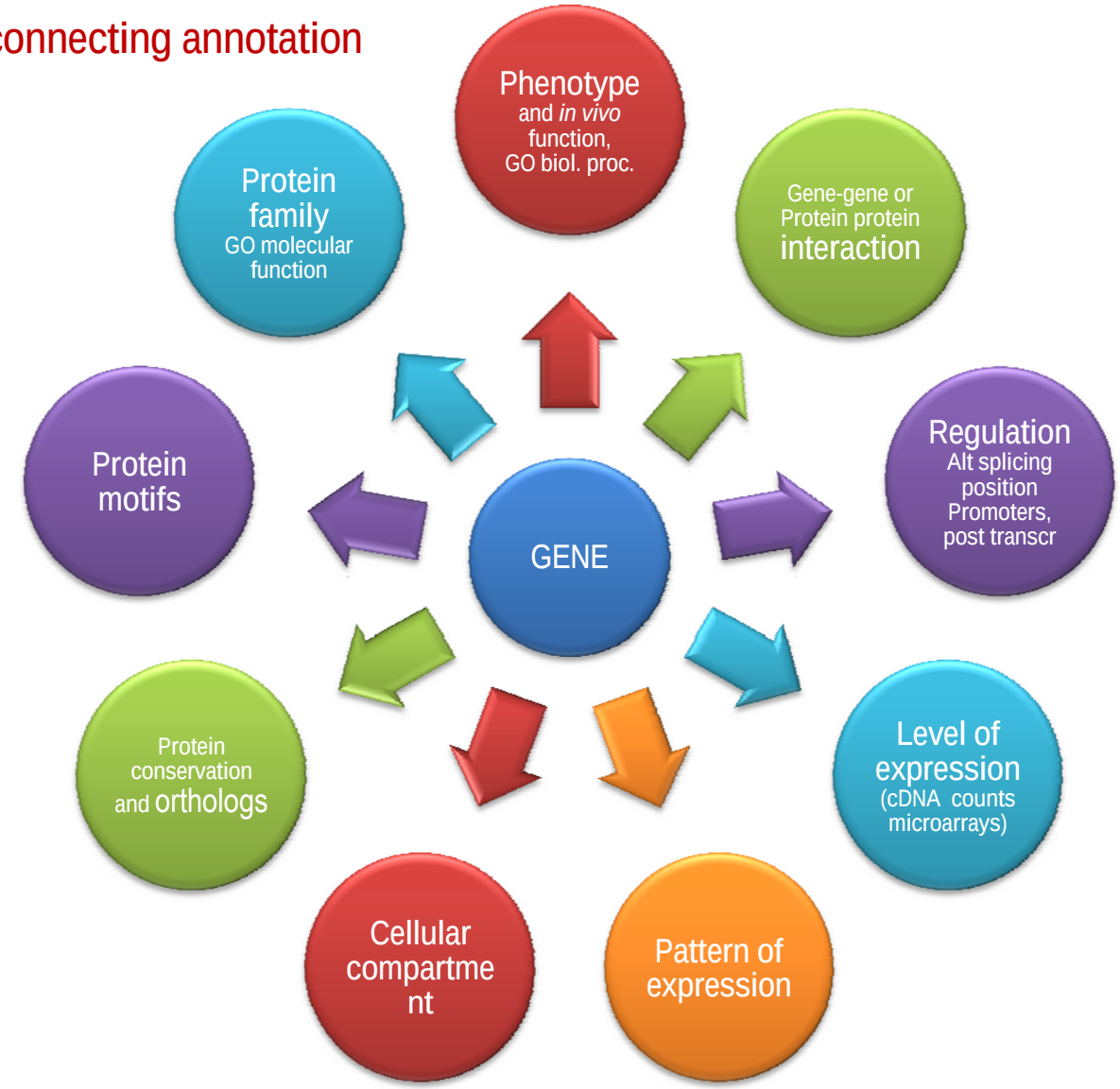
5 **Data Analysis**



Data organization: Basic Model

Research		Resource
Population		*Taxonomy*
Organism		*Genome Project*
Development		*Genome Resources*
Phenotype		
Genetics		*Map Viewer*
Chromosomes		*OMIM, OMIA*
Scaffolds		*Genome*
Gene		*Trace*
mRNA		*Gene* *Nucleotide* *Protein* *RefSeq*
ProProtein		
Mature peptide		*MMDB, Cn3D*
Structure		*CDD*
Domains		
Function	OMIM Online Mendelian Inheritance in Man	*Books*
Disease	OMIA Online Mendelian Inheritance in Animals	*OMIM, OMIA*
Phenotype	Bookshelf PubMed A service of www.pubmed.gov	*PubMed* *PubMed Central*

Multiply connecting annotation



1. Experimental Design
2. Microarray Experiments
3. Image Processing
4. Data Normalization
5. Data Analysis

When you want to determine whether some program or treatment ***causes*** some outcome or outcomes to occur, then you are interested in having strong internal validity. Essentially, you want to assess the proposition:

If X, then Y

or, in more colloquial terms:

If the program is given, then the outcome occurs

Unfortunately, ***it's not enough*** just to show that when the program or treatment occurs the expected outcome also happens. That's because there may be lots of reasons, other than the program, for why you observed the outcome. To really show that there is a causal relationship, you have to simultaneously address the two propositions:

If X, then Y

and

If *not* X, then *not* Y

Or, once again more colloquially:

If the program is given, then the outcome occurs

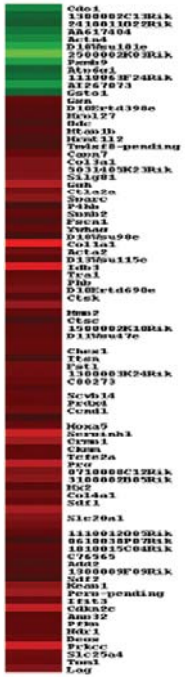
and

If the program is *not* given, then the outcome does *not* occur

Microarray Experiment Design

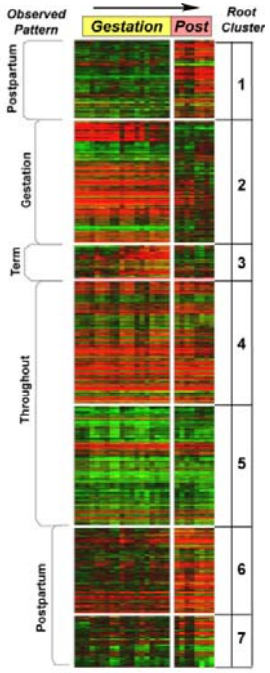
Type I: (n = 2) Differential Expression

- How is this gene expressed in target 1 as compared to target 2?
- Which genes show up/down regulation between the two targets?



Type II: (n > 2) Data Partitioning

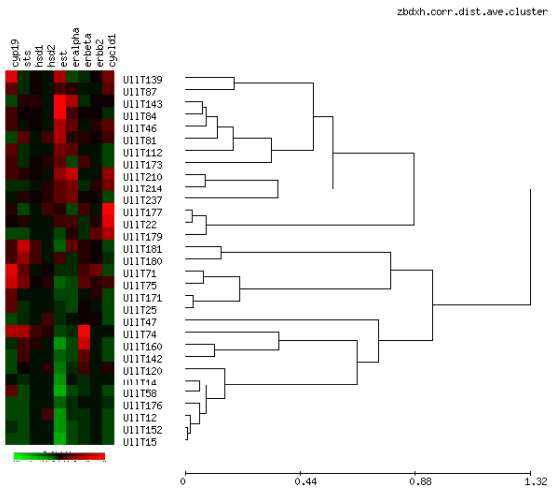
- How does the expression of gene A vary over time, tissues, or treatments?
- Do any of the expression profiles exhibit similar patterns of expression?



Reference design and Balanced block design

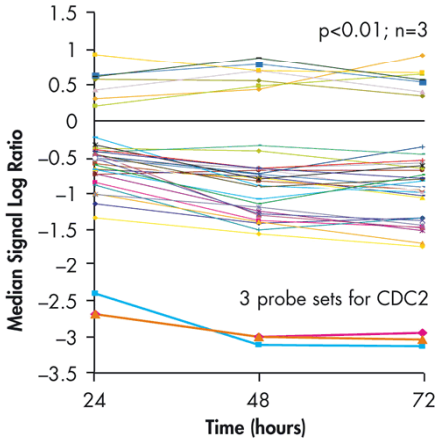
Reference design (flexible)

Reference design is far superior for more complex class comparison as well as for class prediction or class discovery where *clustering analysis is required*. In fact, clustering analysis can not be performed on experiment conducted using balanced block design. As a rule reference design should be used when amount of material is a limiting factor, and balanced block design should be used when number of slides is limited.



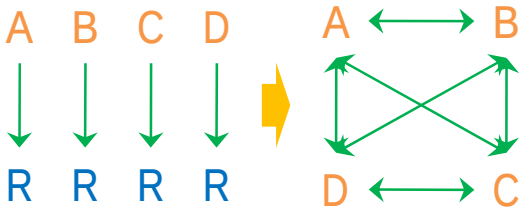
Balanced block design (efficient)

Balanced block design can only be used only for simple class comparison type of analysis (looking for genes that are *upregulated* or *downregulated* between two samples).



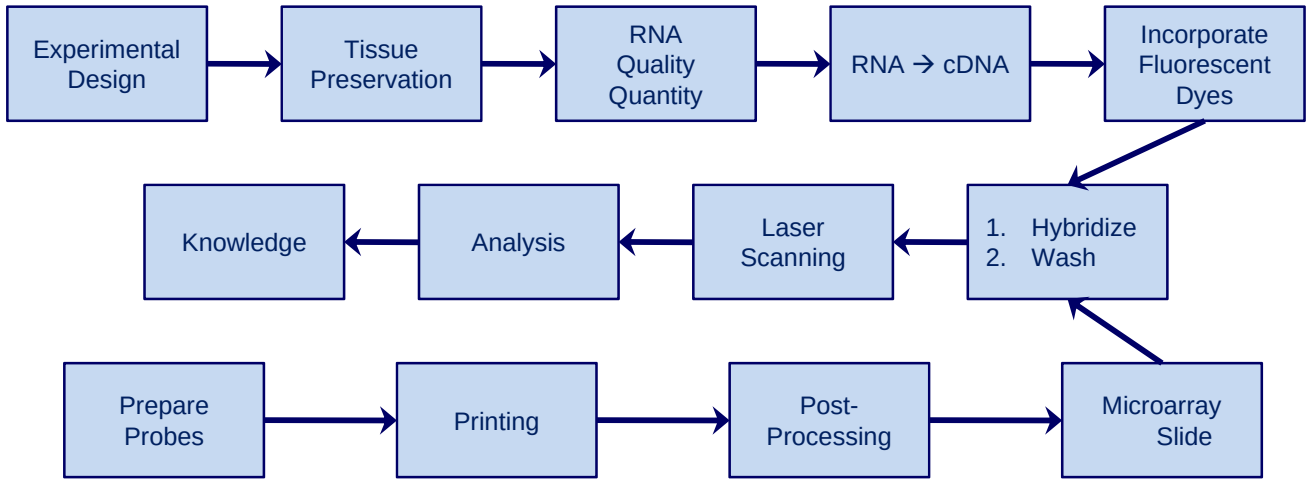
Loop design (elegant)

Reference RNA for reference design experiments used does not necessarily need to be physiologically relevant to your RNA. However, reference design should have at least 80% coverage for the particular slide (i.e. should contain *mRNA representing at least 80% of genes on the slide*). RNA isolated from a cells or tissues usually has 50-60% coverage. Reference RNA is, most often, a pool of RNA from several different cell lines or tissues. Human, mouse, and rat universal reference RNAs are available from several vendors .

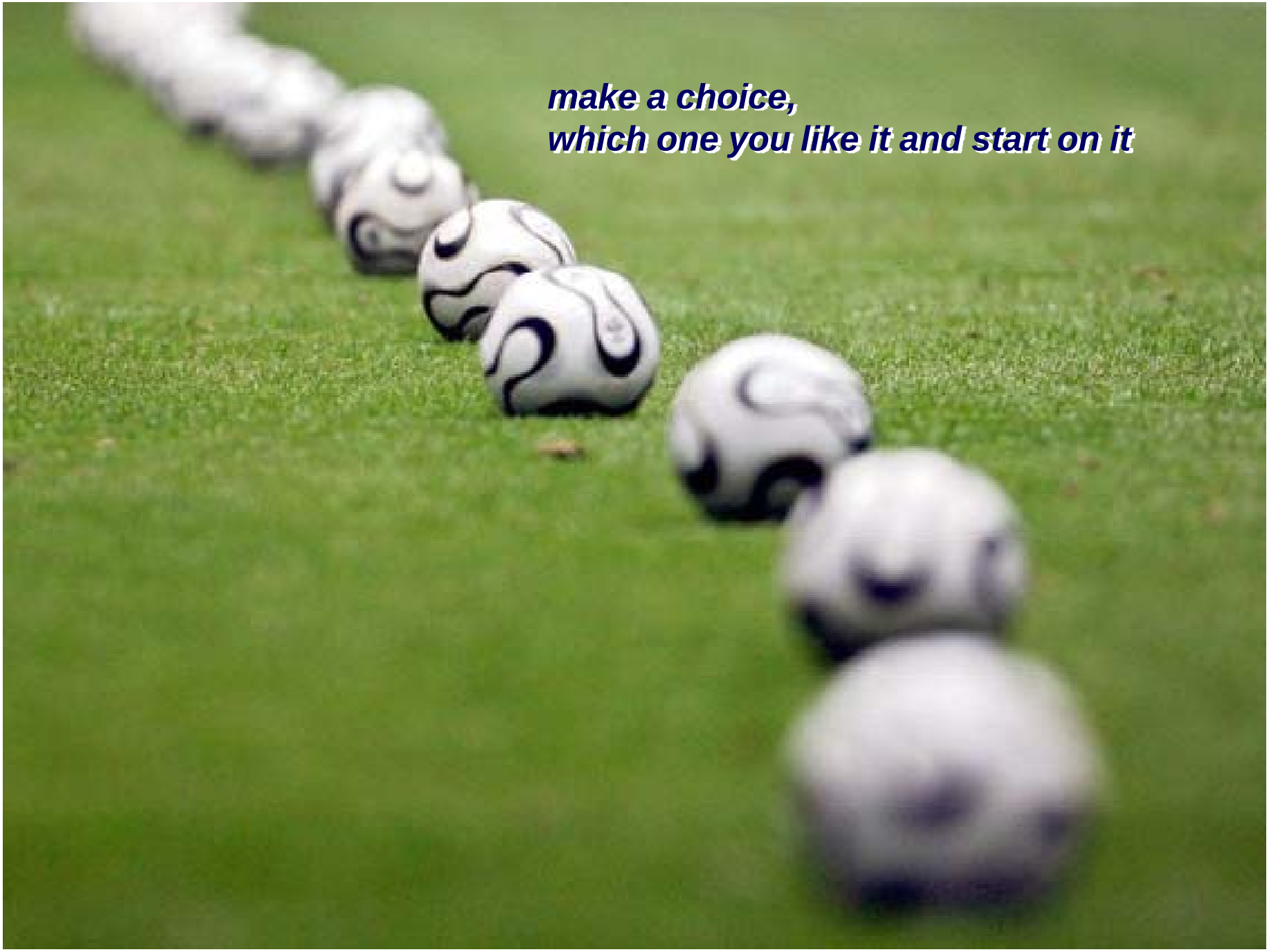


1. Experimental Design
2. **Microarray Experiments**
3. Image Processing
4. Data Normalization
5. Data Analysis

- Microarray technology allow one to identify the genes that are expressed in different cell types, to learn how their expression levels change in different developmental stages or disease states, and to identify the cellular processes in which they participate
- Microarray technology provide clues about how genes and gene products interact and their interaction networks.

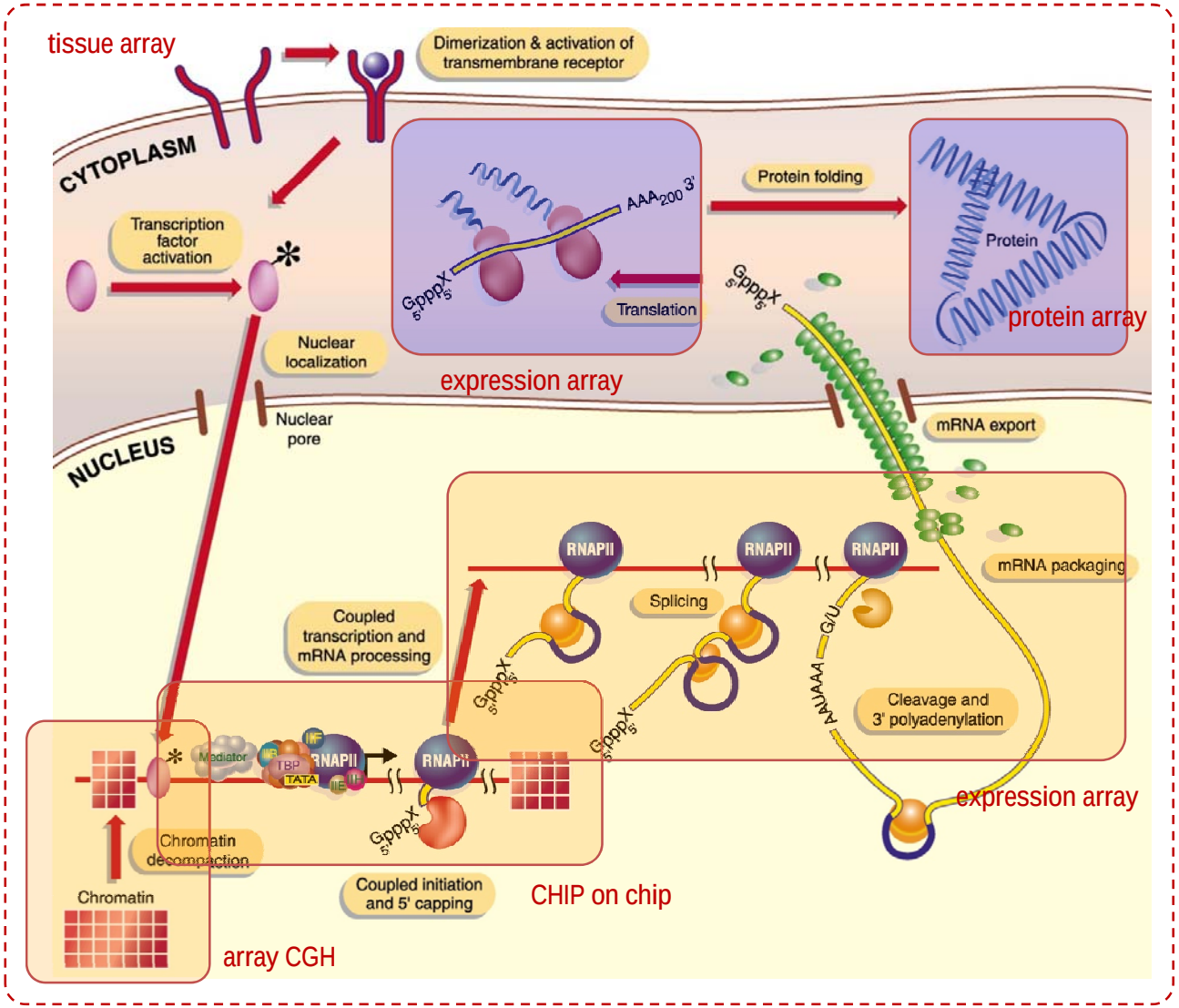


A microarray experiment. The top part of the figure shows steps in the sample preparation process while the bottom part shows steps in the fabrication of the microarray. The samples are hybridized to the microarrays (center). The net result of this complex multi-step procedure is increased understanding of the effects of chemical toxicants.


make a choice,
which one you like it and start on it

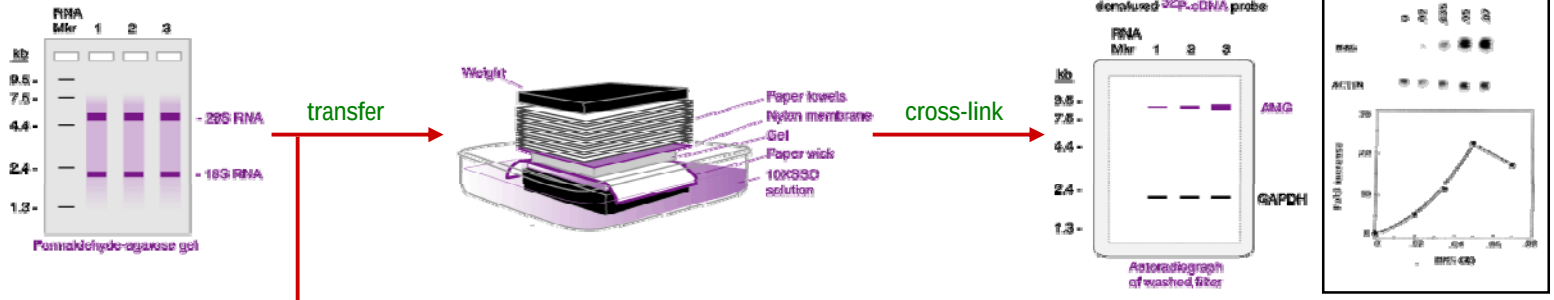
Central Dogma and Application of Microarray

1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array
11. .....

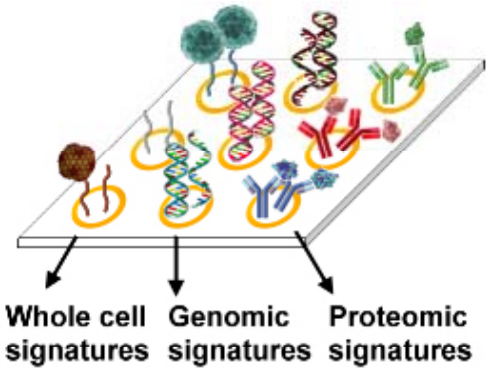


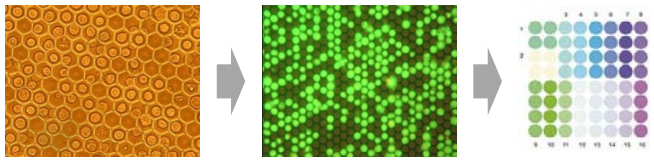
One gene (clone)!

Blotting *versa* Microarray



We could just 35,000 Northern to monitor expression of all genes!!!





signal transfer into digital

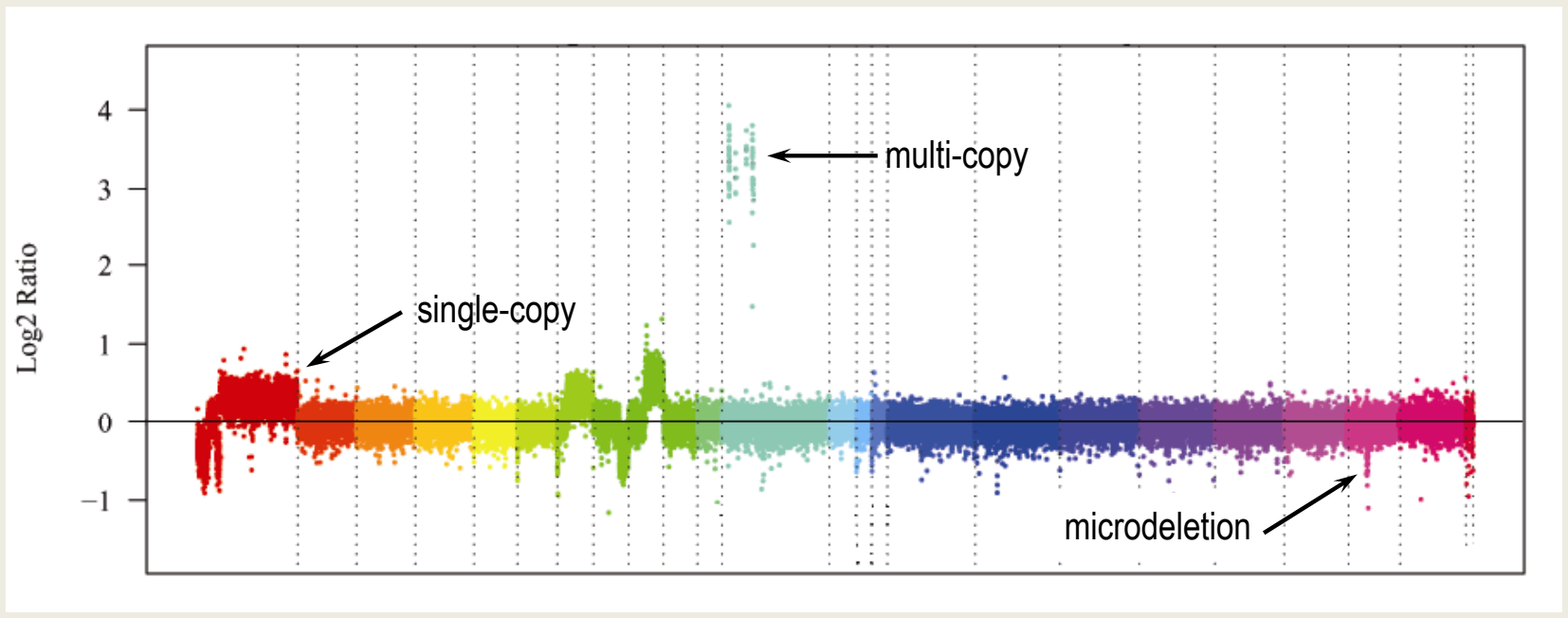
1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

ChIP-on-chip (also known as ChIP-chip) is a technique that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip). Like regular ChIP, ChIP-on-chip is used to investigate interactions between proteins and DNA *in vivo*.



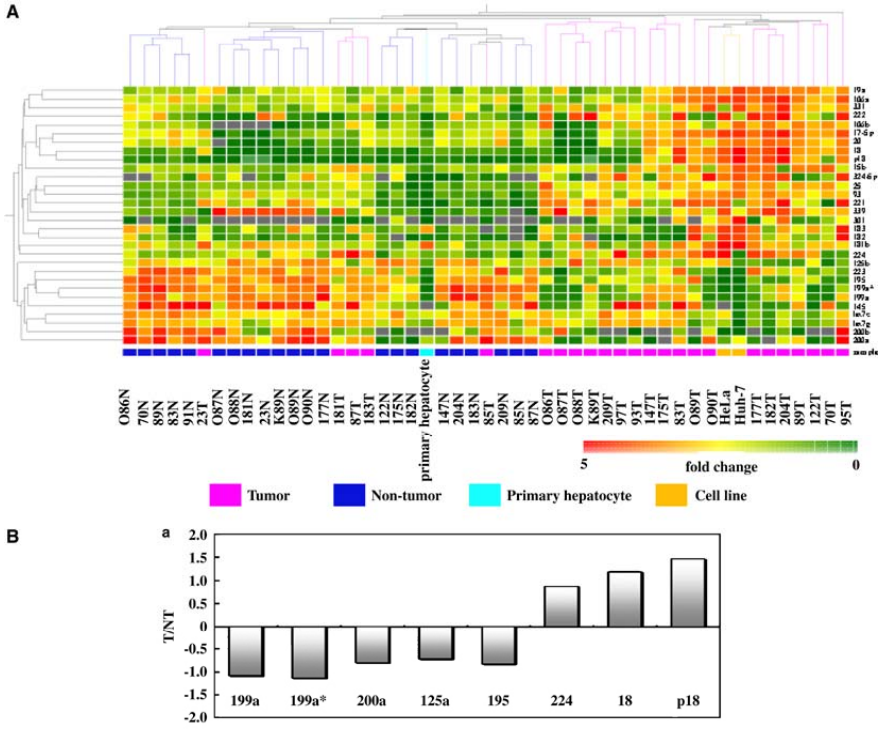
1. ChIP-on chip array
2. **array CGH**
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

Array comparative genomic hybridization (also CMA, Chromosomal Microarray Analysis, Microarray-based comparative genomic hybridization, array CGH, a-CGH, or aCGH) **detects chromosome copy number changes** at a higher resolution level than conventional chromosome-based comparative genomic hybridization (CGH).



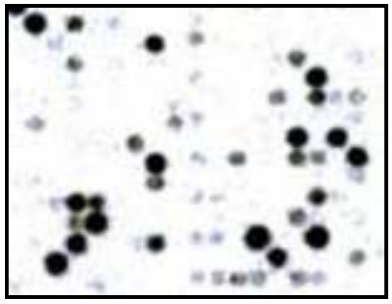
1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

MicroRNAs (miRNAs) are believed to play important roles in developmental and other cellular processes by hybridizing to complementary target mRNA transcripts. This results in either cleavage of the hybridized transcript or negative regulation of translation. MicroRNA array is by using a high-density oligonucleotide array to determine the pattern of global miRNA expression levels in mammals and other higher eukaryotes.

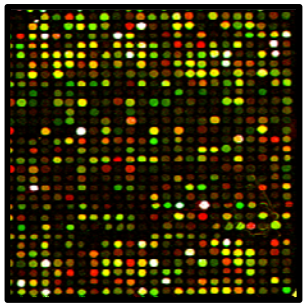


1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

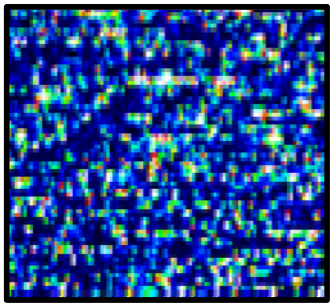
A DNA microarray (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface by covalent attachment to a chemical matrix. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system. Qualitative or quantitative measurements with DNA microarrays utilize the selective nature of DNA-DNA or DNA-RNA hybridization under high-stringency conditions and fluorophore-based detection. DNA arrays are commonly used for expression profiling, i.e., monitoring expression levels of thousands of genes simultaneously, or for comparative genomic hybridization.



Isotope
Nylon – cDNA (300-900 nt)



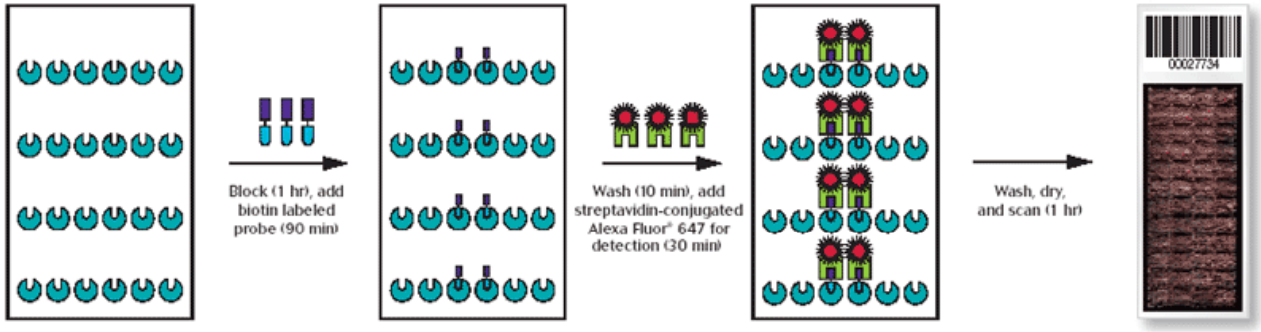
Two-colour
cDNA or Oligo (80 nt)
500 – 11,000 elements



Affymetrix
Silicone – oligo (20 nt)
22 ,000 elements

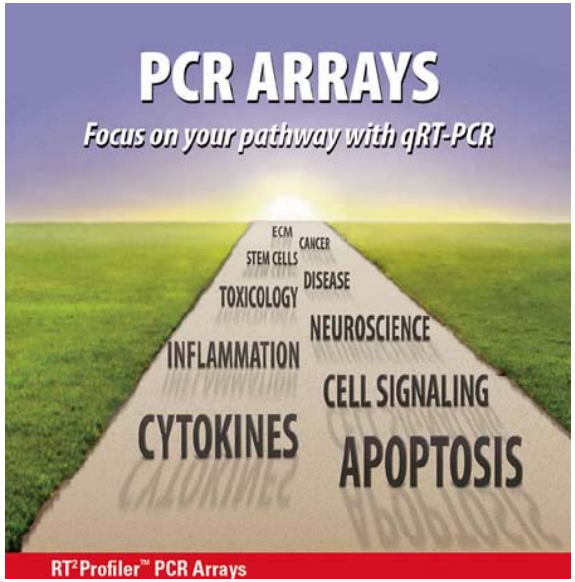
1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

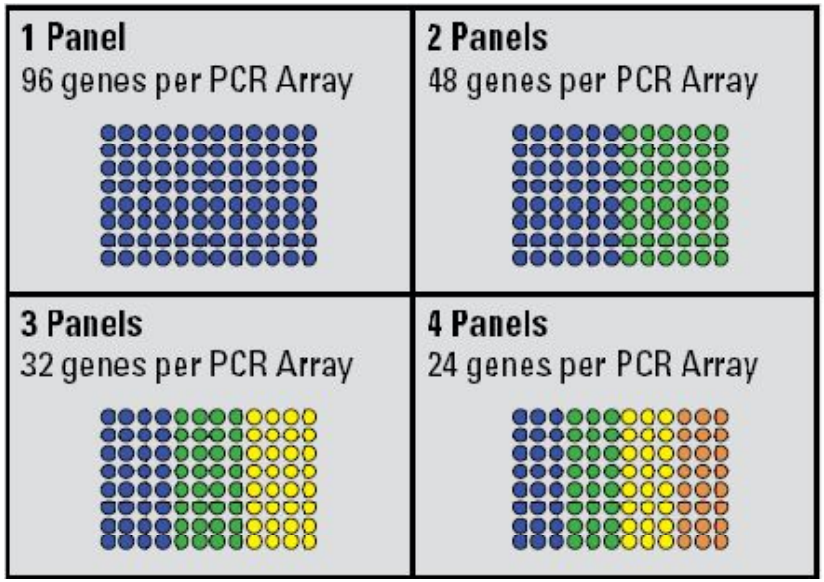
A protein microarray is a piece of glass on which different molecules of protein have been affixed at separate locations in an ordered manner thus forming a microscopic array. These are used to identify protein-protein interactions, to identify the substrates of protein kinases, or to identify the targets of biologically active small molecules. The most common protein microarray is the antibody microarray, where antibodies are spotted onto the protein chip and are used as capture molecules to detect proteins from cell lysate solutions. Related microarray technologies also include DNA microarrays, Antibody microarrays, Tissue microarrays and Chemical Compound Microarrays.



1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

Quantitative polymerase chain reaction (qPCR) is a modification of the polymerase chain reaction used to rapidly measure the quantity of DNA, complementary DNA or ribonucleic acid present in a sample. Like other forms of polymerase chain reaction, the process is used to amplify DNA samples, via the temperature-mediated enzyme DNA polymerase. qPCR Arrays are designed to profile the expression of a panel of genes relevant to a specific pathway or disease state.

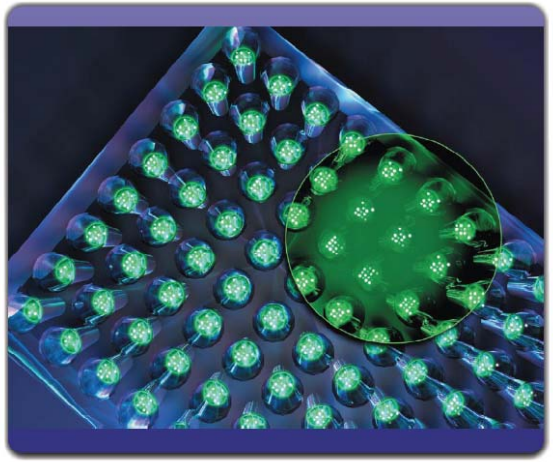




1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. **ELISA array**
8. cell array
9. tissue array
10. lab-on-a-chip array

Enzyme-Linked ImmunoSorbent Assay, or ELISA, is a biochemical technique used mainly in immunology to detect the presence of an antibody or an antigen in a sample. The ELISA has been used as a diagnostic tool in medicine and plant pathology, as well as a quality control check in various industries. ELISA array is a new technology capable of simultaneously identifying multiple biomarkers to generate a biochemical profile. Once the ELISA protocol is completed the array is imaged using either a CCD imaging system or X-ray film to capture the chemiluminescent signal. The pixel intensity of each spot is directly correlated to the concentration.





Q-Plex™ Cytokine Arrays

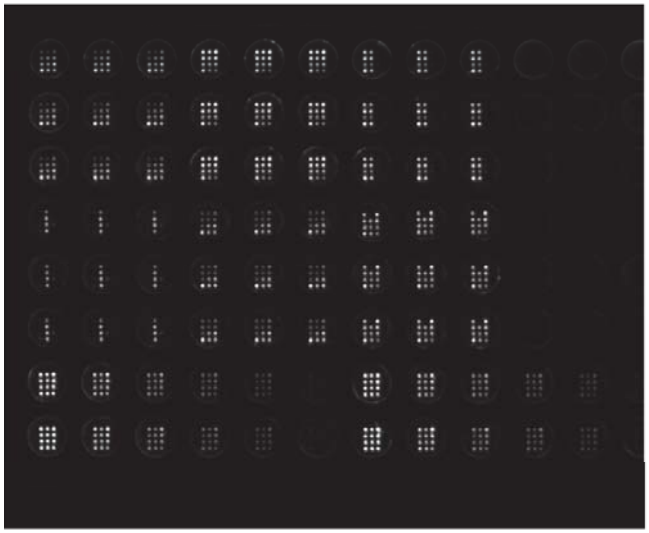
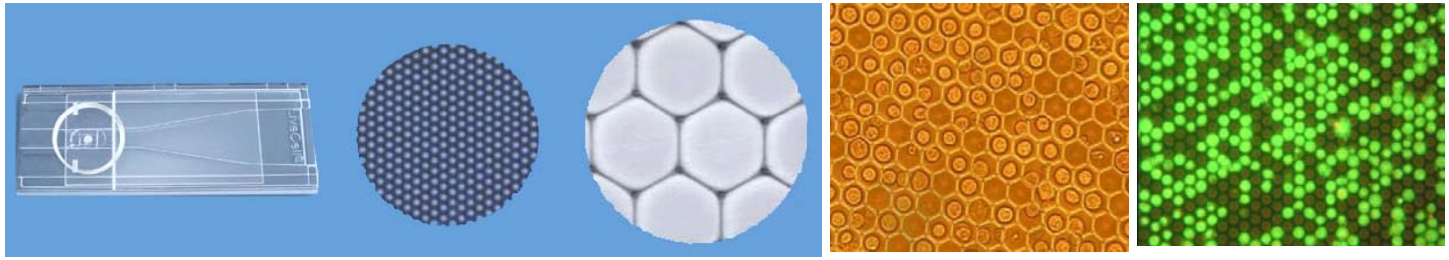


Plate Example

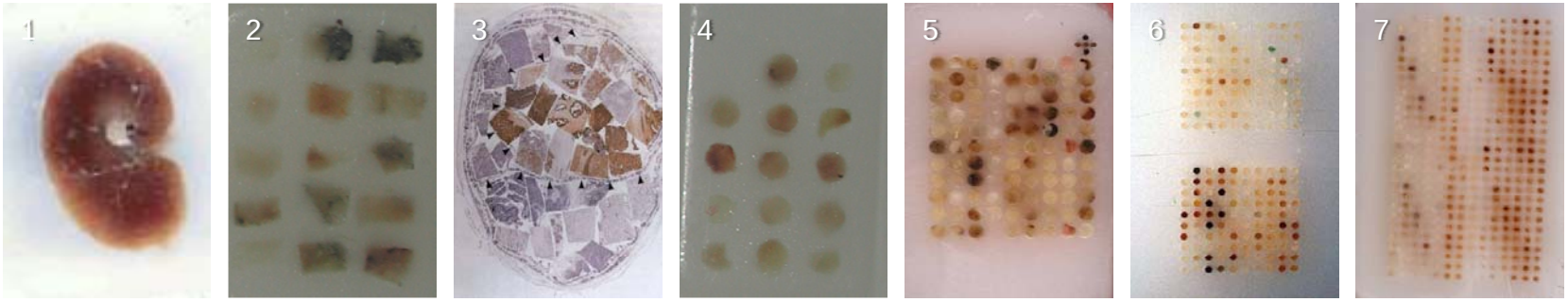
1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. **cell array**
9. tissue array
10. lab-on-a-chip array

Cell array is a common term for different techniques, which are used in genomic-level cell biological testing. The live cell array, a microscope slide-based high content analysis tool, enables **multi-parametric imaging-based assays on thousands of intact individual cells**, including non-adhering blood and bone marrow cells. Cells can be observed in their own identified location, tracking individual, real-time responses to intervention. Multiple functional assays can be performed on a living cell, followed by post-fixation studies on the same cell to maximize usage of cell samples



1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. **tissue array**
10. lab-on-a-chip array

Tissue microarrays are produced by a method of re-locating tissue from conventional histologic paraffin blocks so that tissue from multiple patients or blocks can be seen on the same slide. This is done by using a needle to biopsy a standard histologic sections and placing the core into an array on a recipient paraffin block.



whole tissue

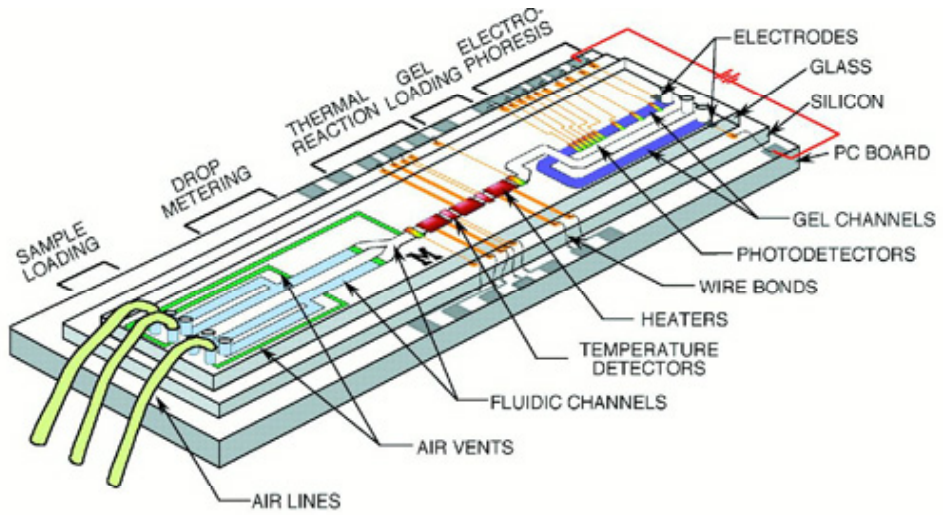
tissue cocktail

tissue sausage

tissue array

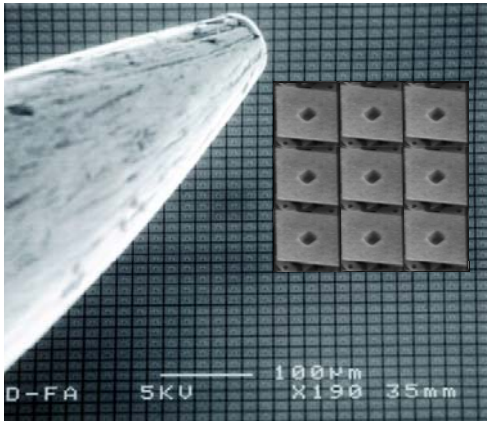
1. ChIP-on chip array
2. array CGH
3. miRNA array
4. DNA array
5. protein array
6. qPCR array
7. ELISA array
8. cell array
9. tissue array
10. lab-on-a-chip array

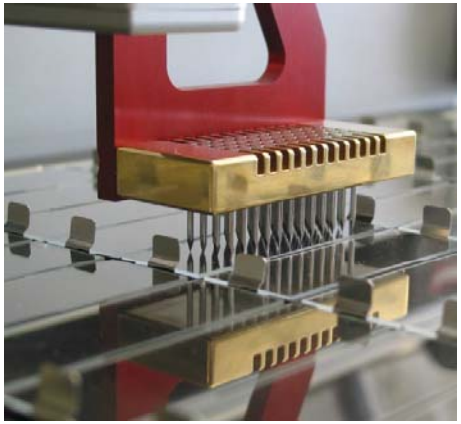
Lab-on-a-chip (LOC) is a term for devices that integrate (multiple) laboratory functions on a single chip of only millimeters to a few square centimeters in size and that are capable of handling extremely small fluid volumes down to less than pico liters

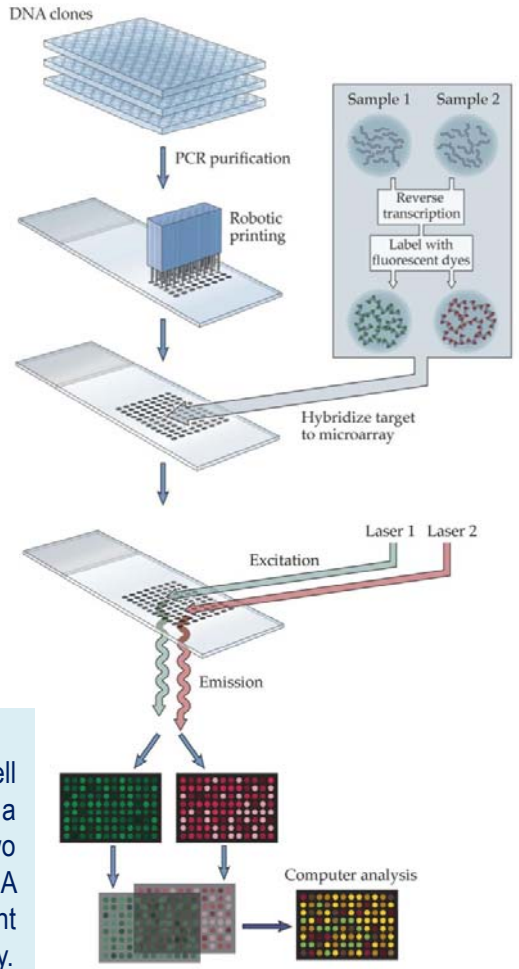


What are microarrays and how do they work ?

- ~10,000 spots on an array
- each spot contains ~10^7 of identical DNA of lengths from 10s to 100s of bp
- spots are either printed on the microarrays by a robot or jet, or synthesized by photolithography or by inkjet printing

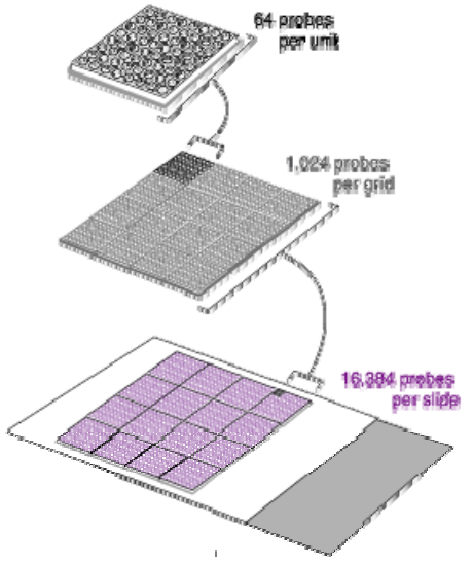


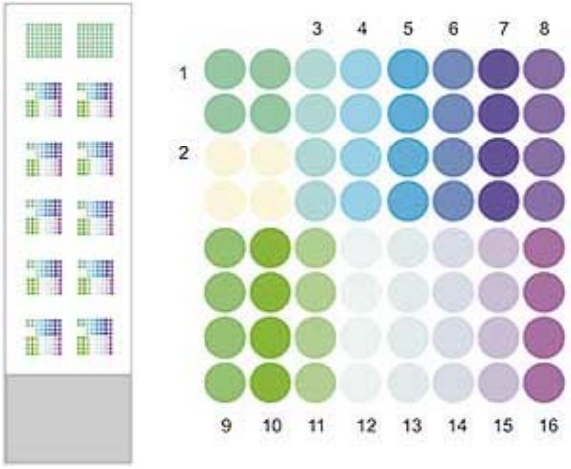


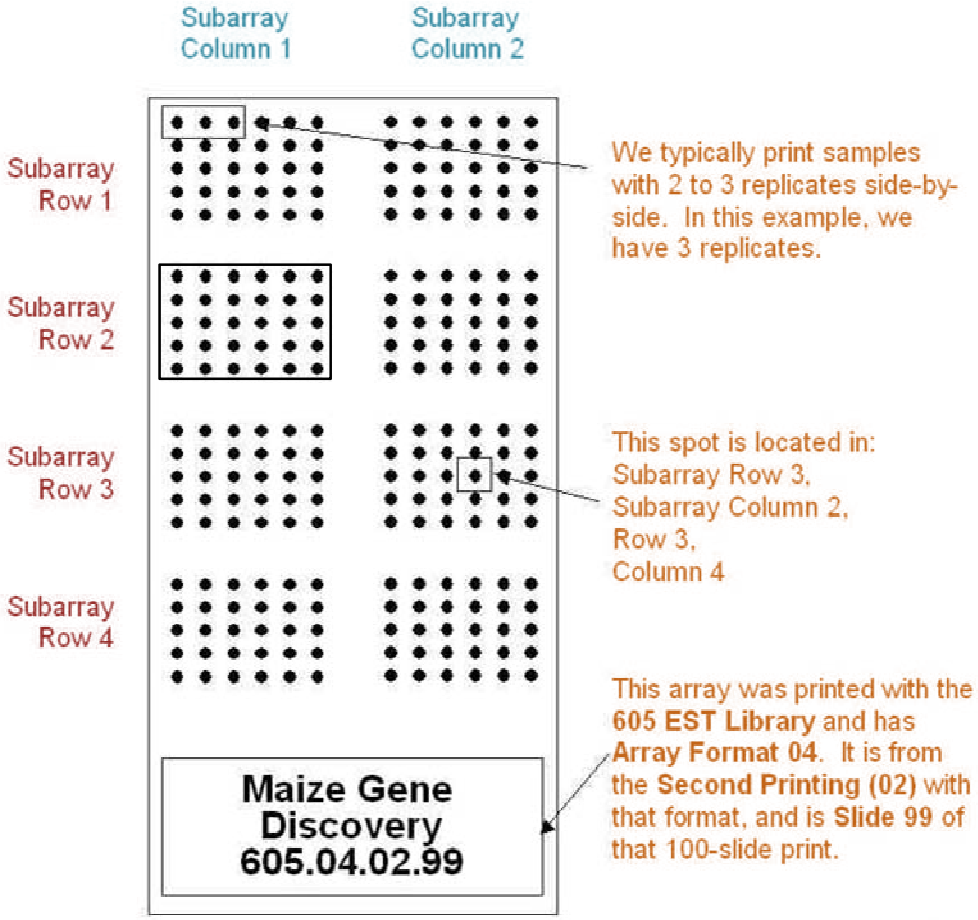


Principle of cDNA microarrays
EST fragments arrayed in 96- or 384-well plates are spotted at high density onto a glass microarray slide. Subsequently, two different fluorescently labeled cDNA populations derived from independent mRNA samples are hybridized to the array.

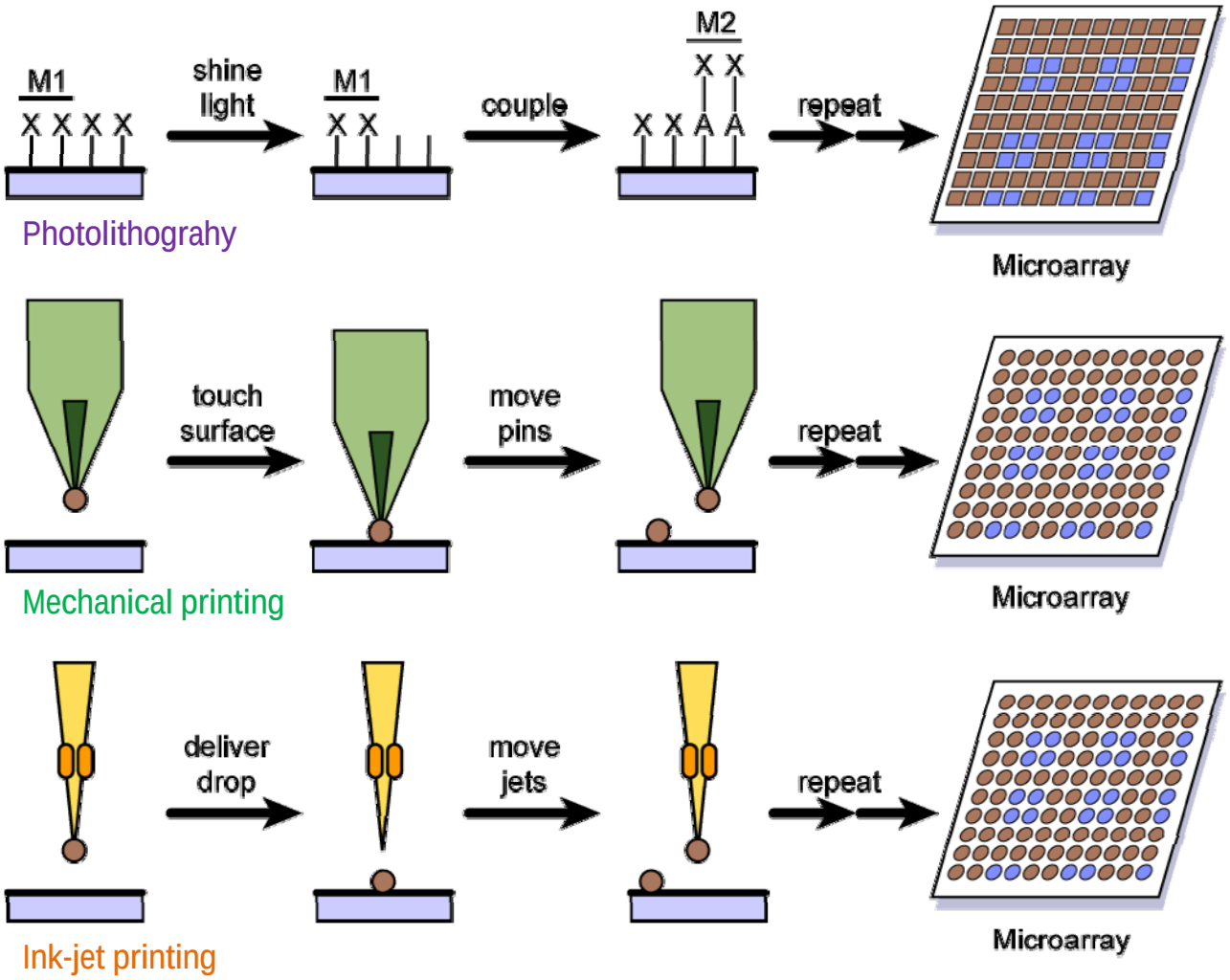
- A microarray is typically a glass or polymer slide. Smooth surface enables even deposition of surface chemistries and perfect spot morphology
- DNA molecules are attached at fixed locations called *spots* or *features*.





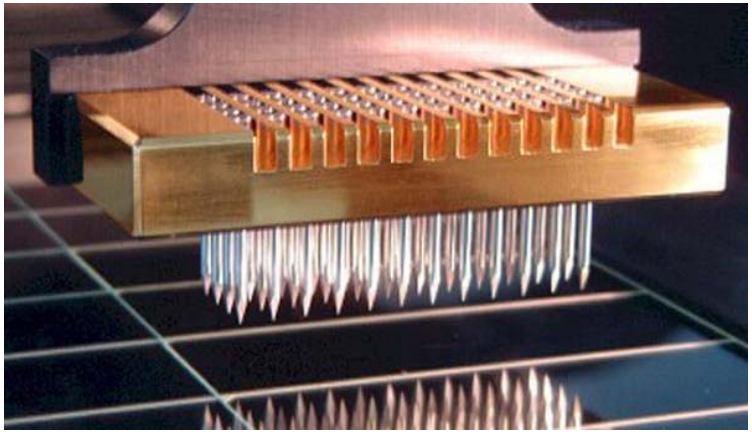


Comparisons of microarray printers



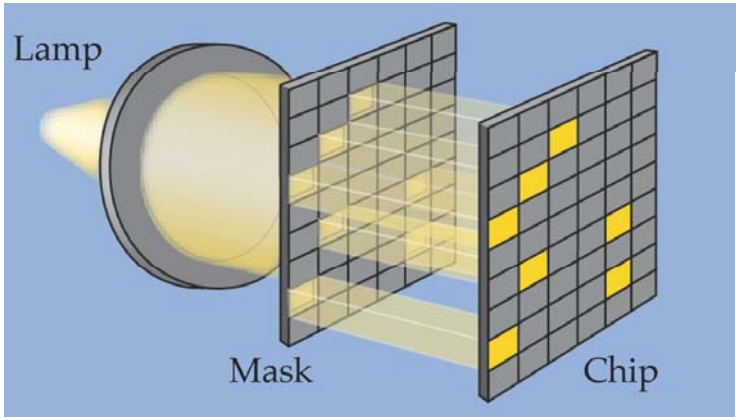
Ink-jet printer microarrays

- ✓ Ink-jet printhead draws up DNA Ink-jet printer microarrays
- ✓ Printhead moves to specific location on solid support
- ✓ DNA ejected through small hole
- ✓ Used to spot DNA or synthesize oligonucleotides directly on glass slide
- ✓ Use pioneered by Agilent Technologies, Inc.

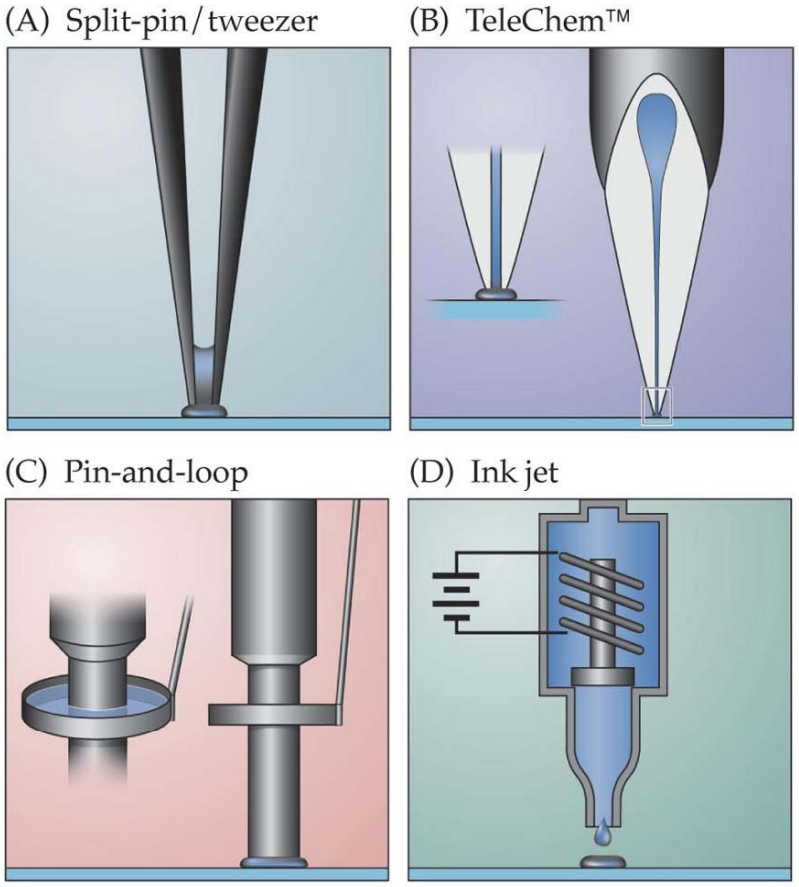


Photolithography

- ✓ Light-activated chemical reaction
  - For addition of bases to growing oligonucleotide
- ✓ Custom masks
  - Prevent light from reaching spots where bases not wanted
- ✓ Mirrors also used
  - NimbleGen™ uses this approach



Types of printing pins



(A) Tweezers or split-pin designs transfer low nanoliter (10-9 liter) amounts of DNA to the array by capillary action as the tip strikes the solid surface.

(B) TeleChem™ tips and pins apply small droplets by contact between the pin and substrate.

(C) Pin-and-loop design picks up the DNA in a small loop, and a pin stamps solution on a slide at a uniform density.

(D) Ink jets spray picoliter (10-12 liter) droplets of liquid under pressure.

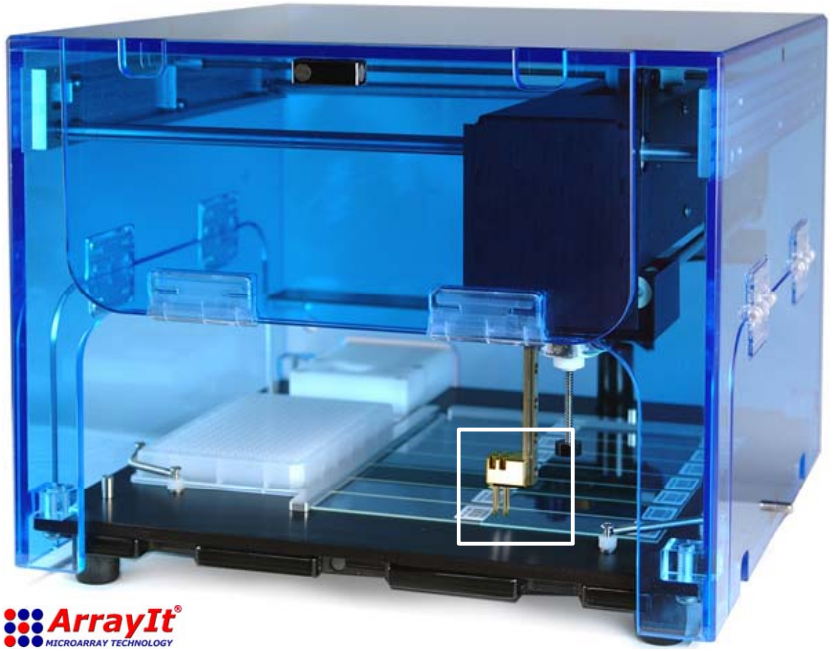
The spacing between spot centers is specified form 120-250 μm according to the density required. The entire microarray usually covers an area 2.5x5.0 cm, though shorter grids can be printed when fewer clones are to be represented.

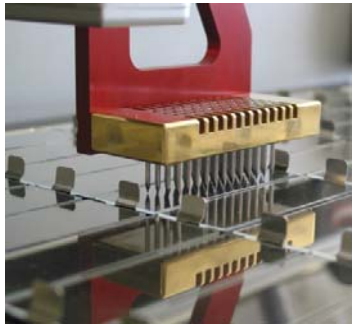
DNA spotting

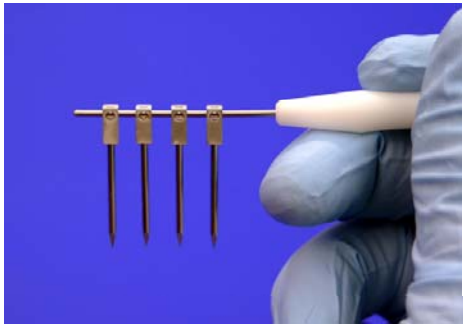
- DNA spotting usually uses multiple pins
- DNA in microtiter plate
- DNA usually PCR amplified
- Oligonucleotides can also be spotted



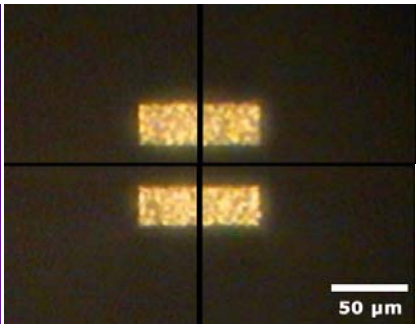








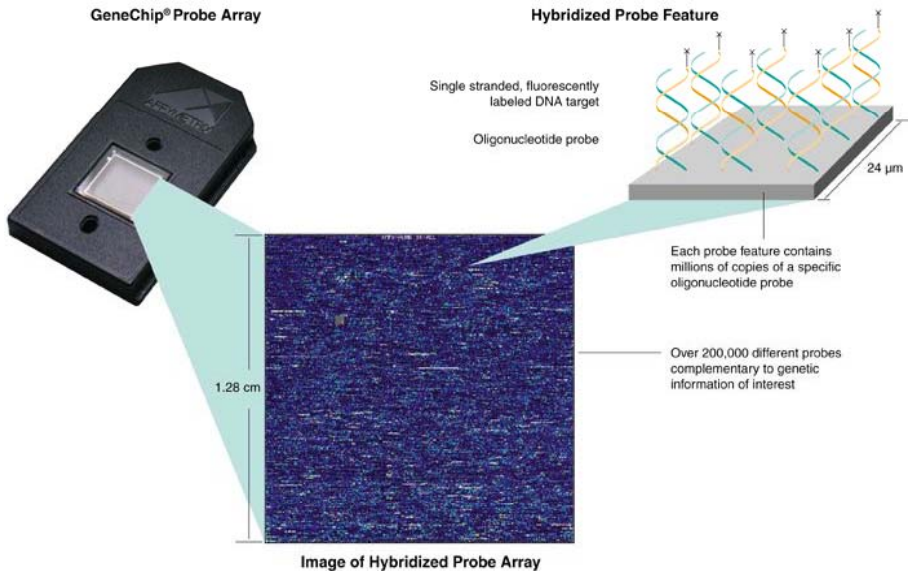




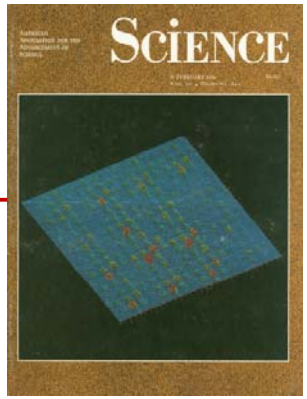
Oligonucleotide microarrays – pioneered by Affymetrix

- Oligonucleotides
  - usually at least 20–25 bases in length, optimal with 45~60 bps long
  - 10–20 different oligonucleotides for each gene
- Oligonucleotides for each gene selected by computer program to be the following:
  - unique in genome ($4^{(20\ to\ 25)} = 2^{(40\ to\ 50)} >> 3*10^9 = 2^{30}$), not likely to appear twice
  - non-overlapping (if the sequence length is too short then specificity is low, whereas if the length is too long, self-hybridization could happen)
- Composition based on design rules
  - empirically derived rules

  (ratio of G-C pairs vs. A-T pairs which could affect the melting temperature of the seq., ie. $T_m = 64.9+0.41*(GC\%) -675/L$)



Affymetrix History



Light-directed, spatially addressable parallel chemical synthesis
SP Fodor, JL Read, MC Pirrung, L Stryer, AT Lu, and D Solas
Science 15 February 1991: 767-773.

High Contrast film-based Photomask
1989



Photocamera
1989



First Microarray Scanner
1989



First Commercial Scanner
1994



Affymetrix Microarray Scanner by Hewlett-Packard
1997



Affymetrix GeneChip® Scanner 3000
2003



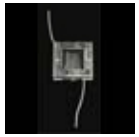




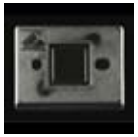
1989
First Microarray Prototype



1989
Microarray Fabrication Cell



1993
First Microarray Containing over 1 Million DNA Sequences



1994
First Commercial DNA Microarray Prototype



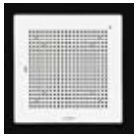
1989
500 Micron Manufacturing Mask:



1993
16-Format Manufacturing Mask



1995
49-Format Manufacturing Mask



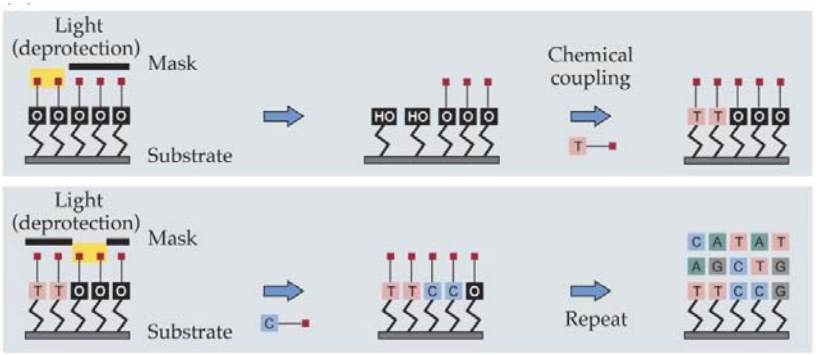
1999
400-Format Manufacturing Mask

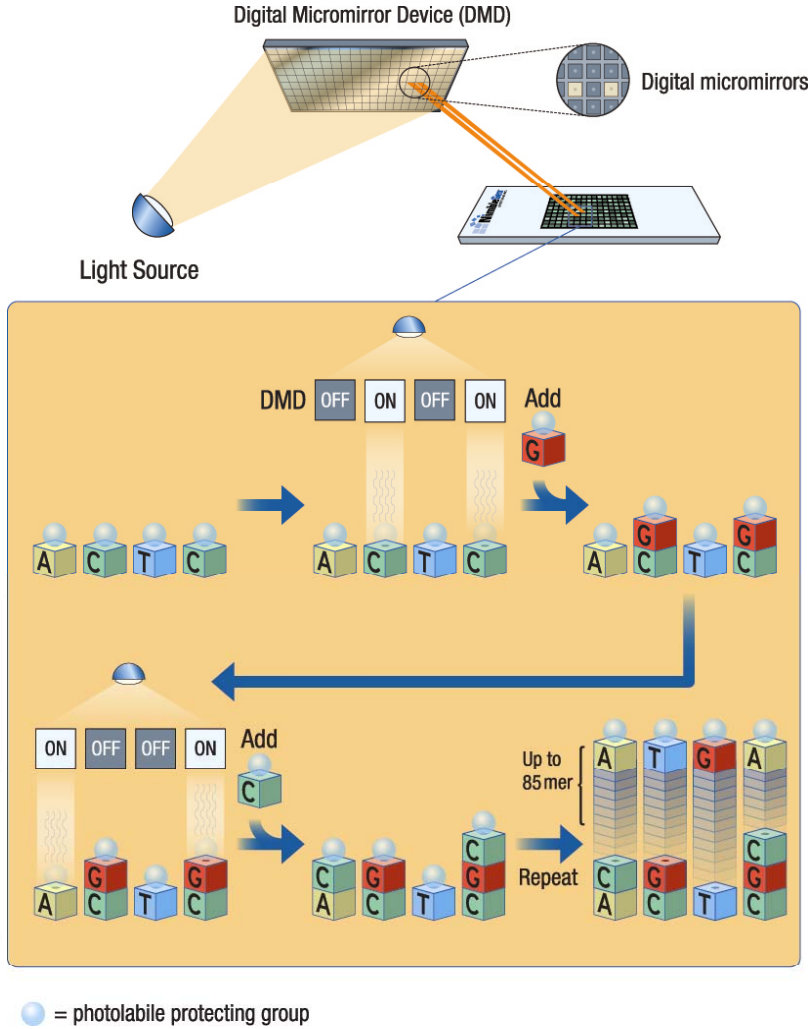


2002
GeneChip® Human Genome U133A array

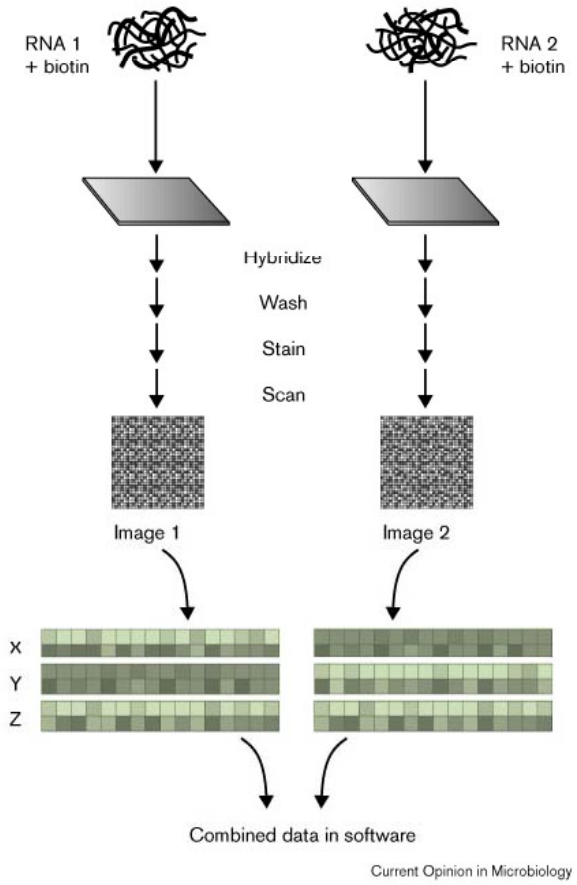
Construction of oligonucleotide arrays

- The light flash is produced by photolithography using a mask to allow light to strike only the required features on the surface of the chip.
- Oligonucleotide are synthesized *in situ* in the silicon chip. In each step, a flash of light "deprotects" the oligonucleotides at the desired location on the chip; then "protected" nucleotides of one of the four types (A, C, G or T) are added so that a single nucleotide can add to the desired chains. There are four types of masks according to the added nucleotide.





1. Experimental Design
2. Microarray Experiments
3. **Image Processing**
4. Data Normalization
5. Data Analysis



Samples

Platform

Raw Data *(image, .tif file)*

(1) image acquisition
(2) spot location

Preprocessed Data *(.gal file)*

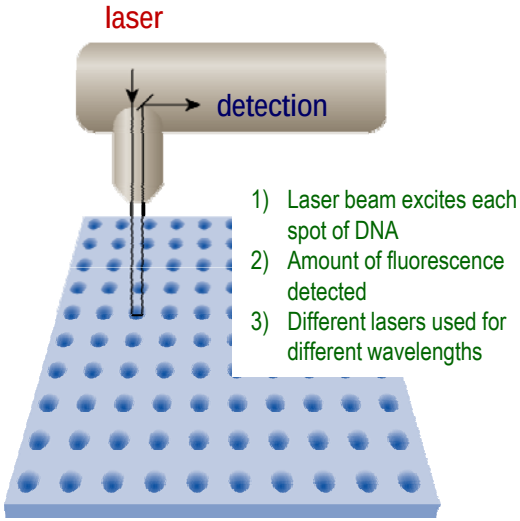
(3) computation of spot intensities

Data Analysis (different levels)

(4) data reporting

Image Processing

- Resolution
  - standard 10μm (100,000 atoms wide)
  - 100μm spot on chip = 10 pixels in diameter
- Image format
  - TIFF 16 bit (64K grey levels)
  - 1cm x 1cm image at 16 bit = 2Mb (uncompressed)
- Separate image for each fluorescent sample
  - channel 1, channel 2



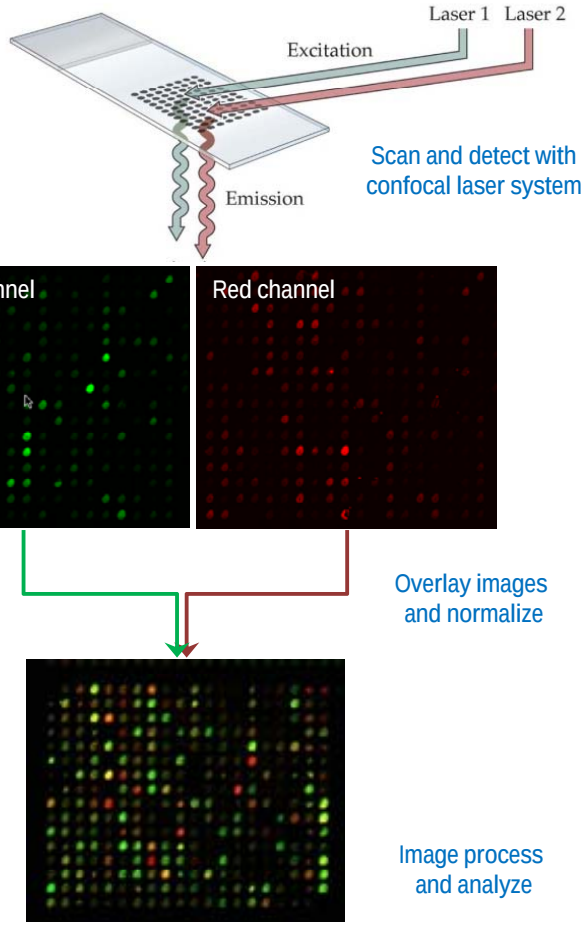
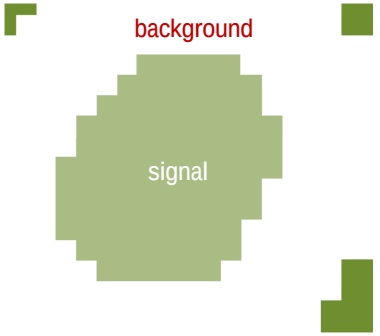
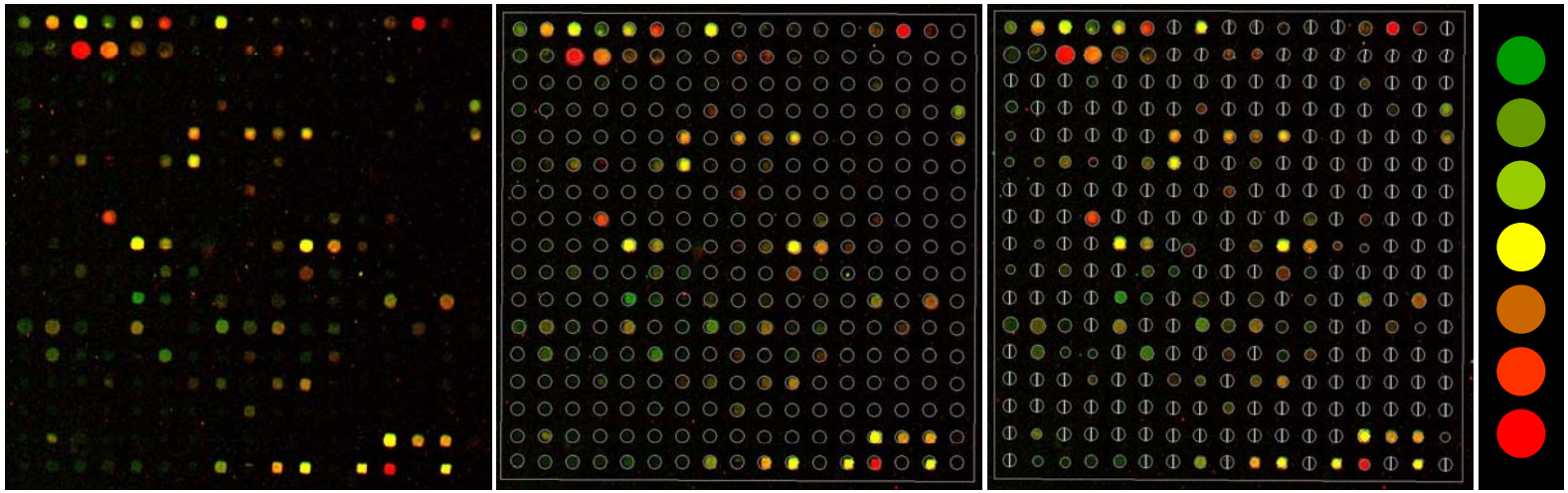


Image Processing

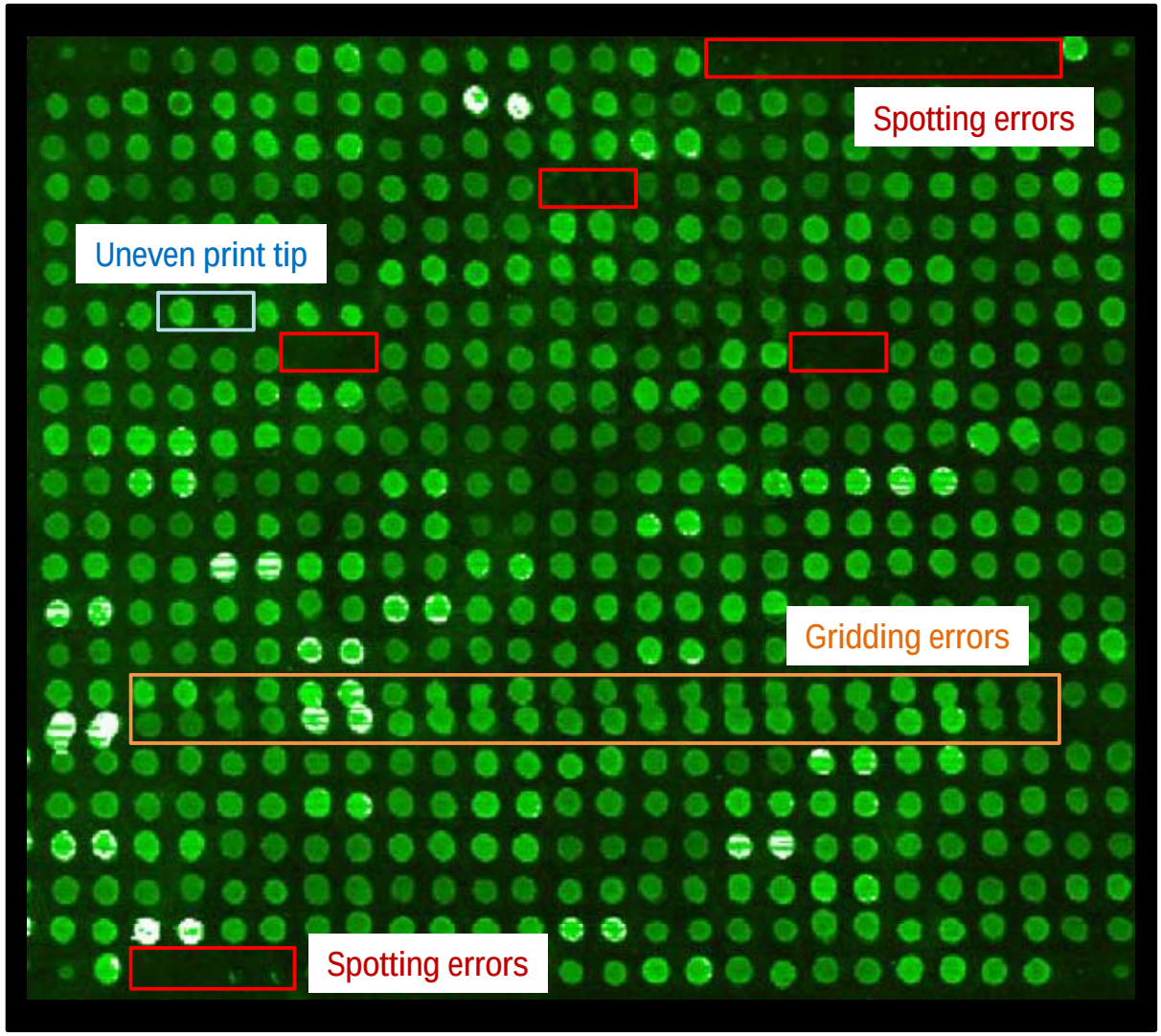
- Addressing or gridding
  - locate centers, assigning coordinates to each of the spots
- Segmentation or spot picking
  - classifying pixels either as foreground or as background
- Intensity extraction (for each spot)
  - Foreground fluorescence intensity pairs (R, G)
  - Background intensities
  - Quality measures
- Information extraction
  - for each spot of the array, calculates signal intensity pairs, background and quality measures





Raw (combined) image | Gridded | Spots picked & flagged | Intensity

Gridding Errors



Quality problems

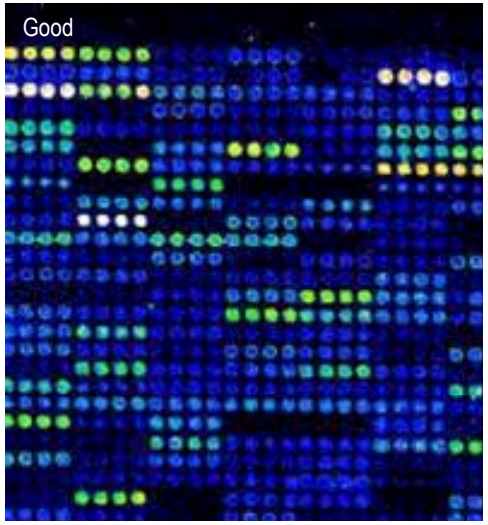
too much rehydration during post –processing

likely caused by insufficiently rapid immersion of the slides in the succinic anhydride blocking solution.

did not wash well and phase change insufficient

insufficient blocking and precipitation of the labeled probe

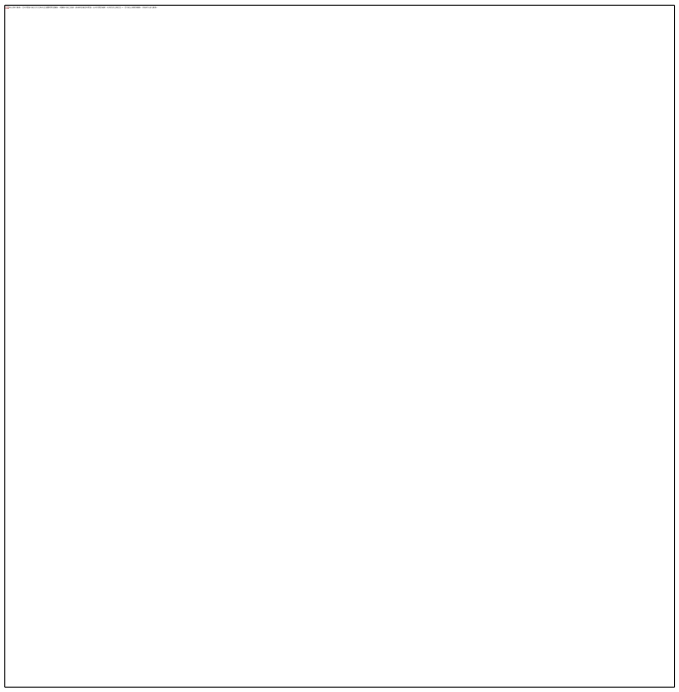
slide expired, wrong Tm, and labelling insufficient

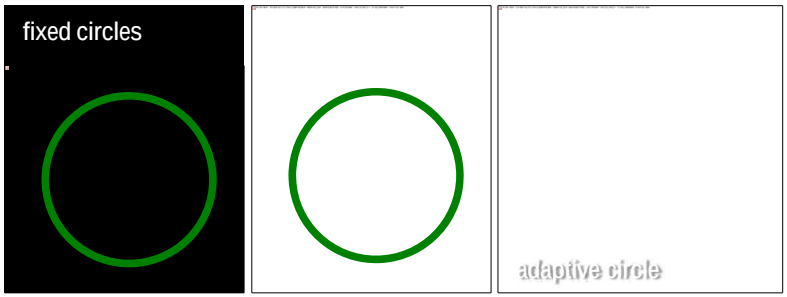


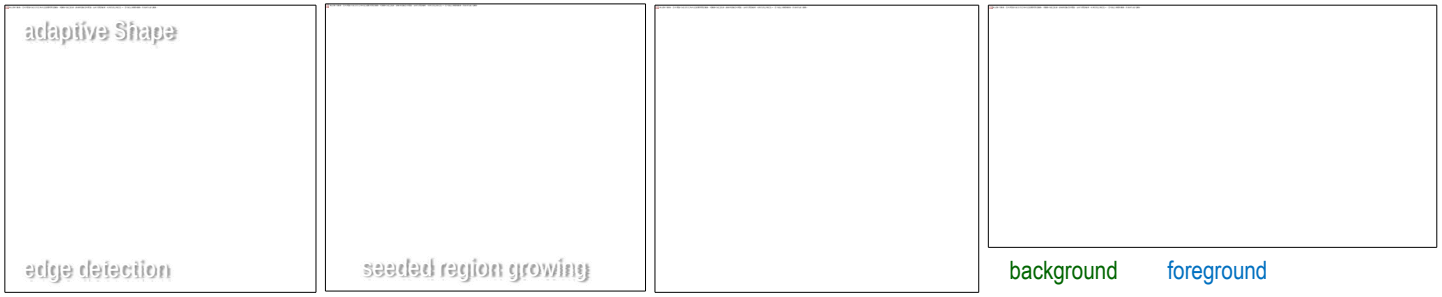
Segmentation methods

1. Fixed circles
2. Adaptive Circle: circle diameter is estimated separately for each spot
3. Adaptive Shape
   - edge detection
   - seeded region growing
4. Histogram Segmentation
   - adaptive threshold

Spot morphology does not affect dynamic range

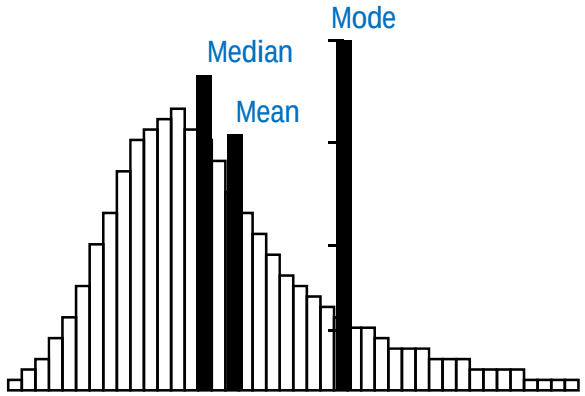


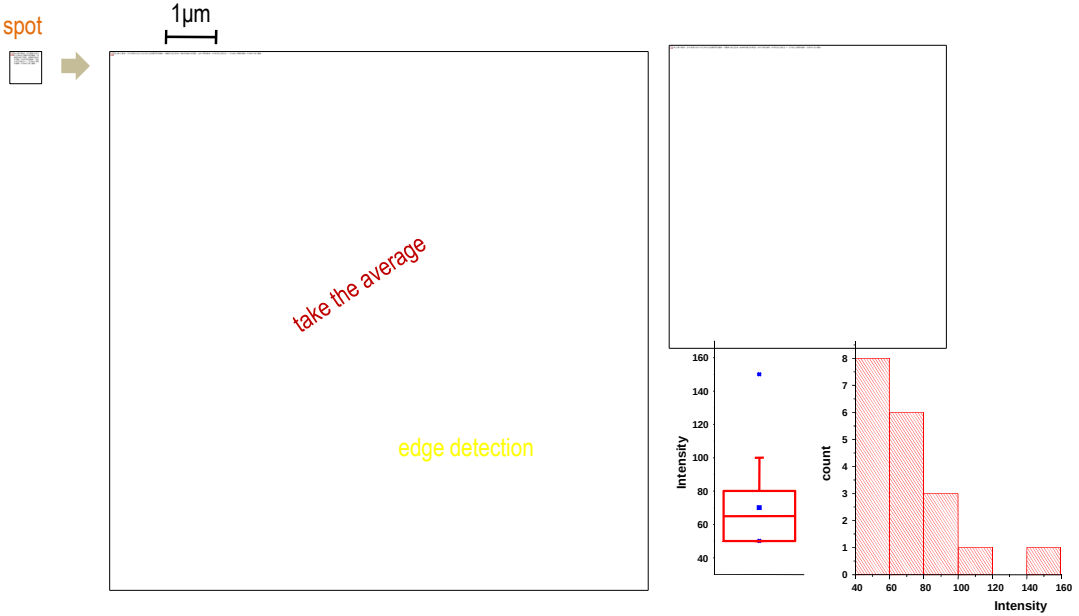




Information Extraction

- Spot Intensities
  - mean (pixel intensities)
  - median (pixel intensities)
- Background values
  - Local
  - Morphological opening
  - Constant (global)
  - None
- Quality Information

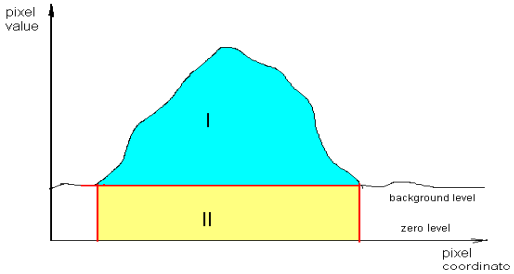




- The total amount of hybridization for a spot is proportional to the total fluorescence at the spot
- Spot intensity = pixel intensities within a spot
- Later calculations are based on ratios between Cy5 and Cy3, so we tally in some way the intensity of the spot
- Can use ratios of medians, means or even modes (if binned)
- Non-specific hybridization subtracted (area outside the spot)

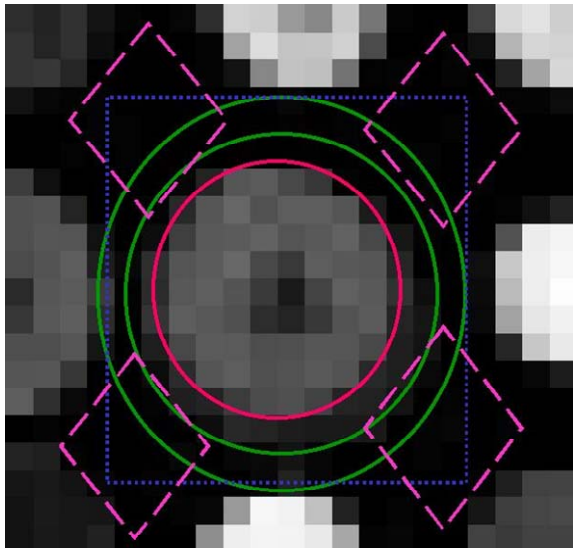
Background Intensity

- A spot's measured intensity includes a contribution of non-specific hybridization and other chemicals on the glass.
- Fluorescence intensity from regions not occupied by DNA can be different from regions occupied by DNA .

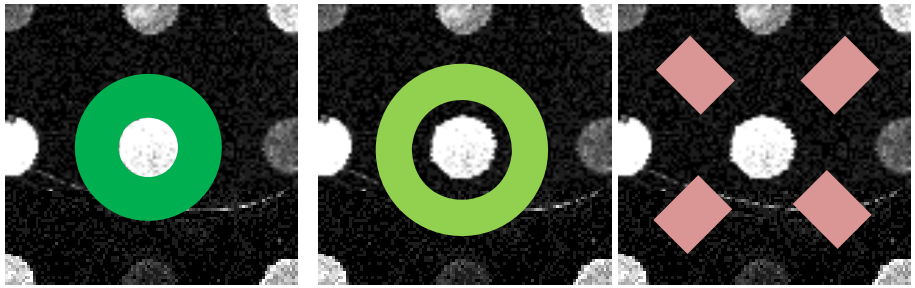


I-integral corrected for background
II- background integral
I+II - total spot integral

Spot intensities =
mean intensity for each pixel within the circle surrounding a spot – mean (median) intensity of the background pixels immediately surrounding the spot



- Focuses on small regions around spot mask.
- Determine median/mean pixel values in this region.



ScanAlyze ImaGene GenePix

SNR (signal-to-noise ratio) = Median signal intensity / Median noise value
, where signal = foreground, noise = background

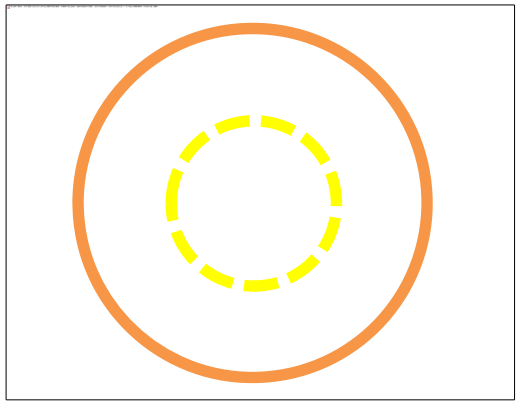
Quality Measurements

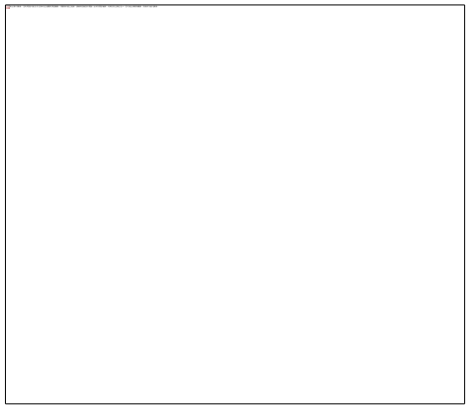
- Array
  - Correlation between duplicate spot intensities
  - Percentage of spots with negative signals
  - Distribution of actual spot signal area vs. idealized
  - Inter-array consistency
- Spot
  - Signal / Noise ratio
  - Variation in pixel intensities within spots

$$\text{SNR (signal-to-noise ratio)} = \frac{\text{Median signal intensity*}}{\text{Median noise value}}$$

, where signal = foreground, noise = background

$$\text{* Mean signal intensity} = \frac{\text{Sum of intensity value}}{\text{Total number of pixels}}$$



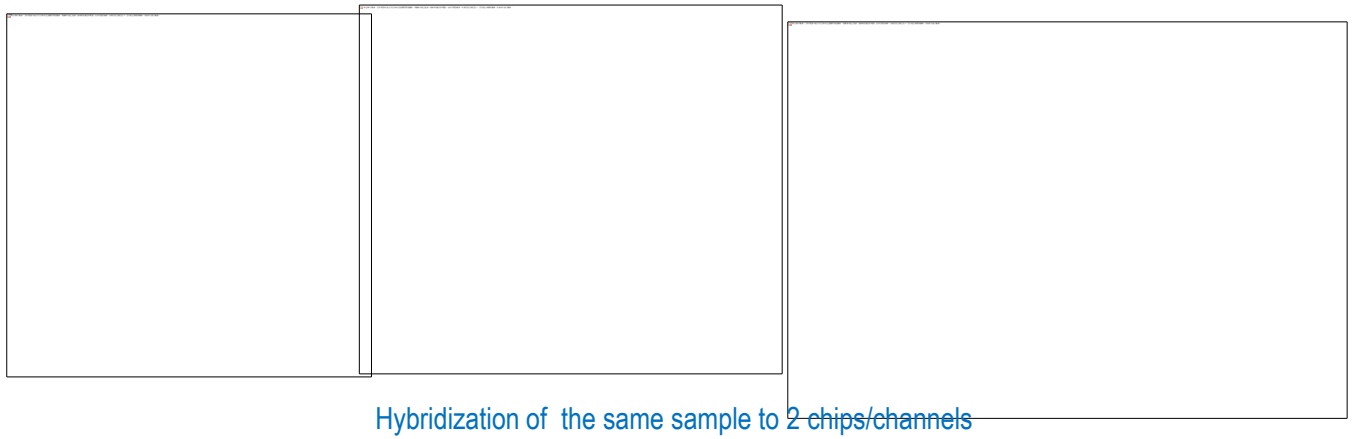




Sources of Systematic Errors (*bias*)

- Different incorporation efficiency of dyes
- Different amounts of mRNA
- Experimenter/protocol issues (comparing chips processed by different labs)
- Different scanning parameters
- Batch bias

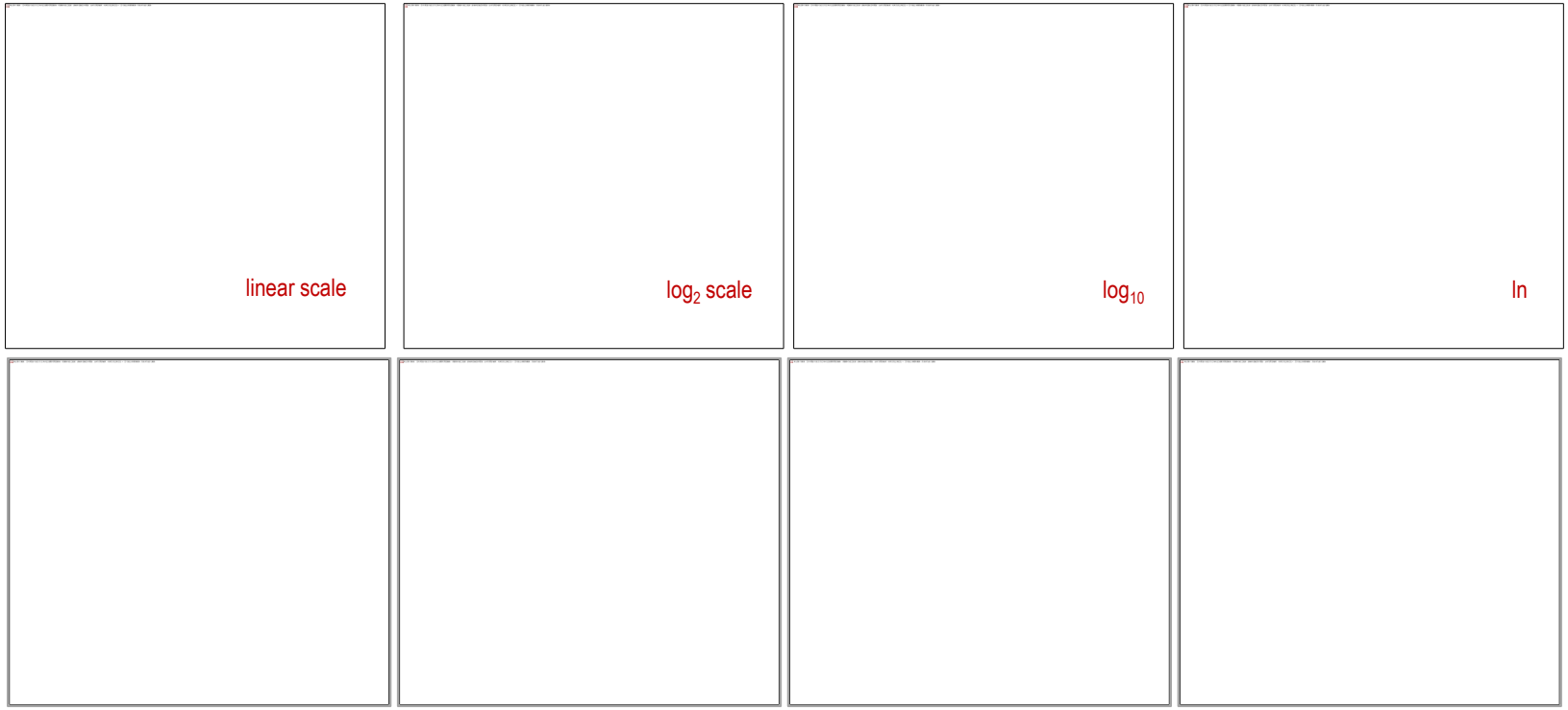
- Ideally: scatter plot coincides with the x=y diagonal
  - Due to Random errors: we expect to see a 'cloud' around the x=y diagonal.
- In practice: Both Random and Systematic measurement errors (Bias)
  - Due to Biases scatter plots are not centered around the x-y diagonal

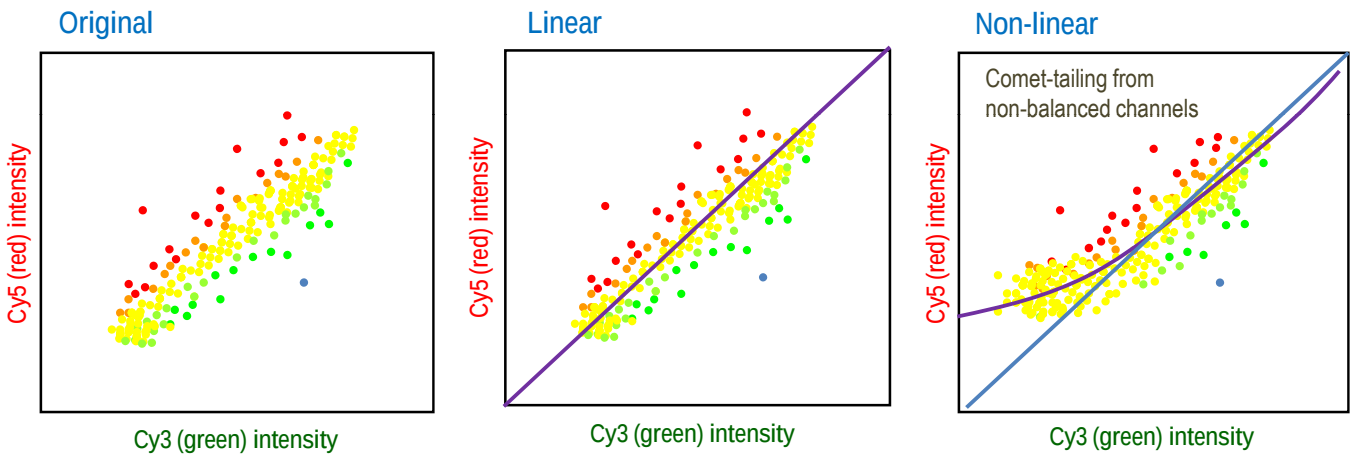


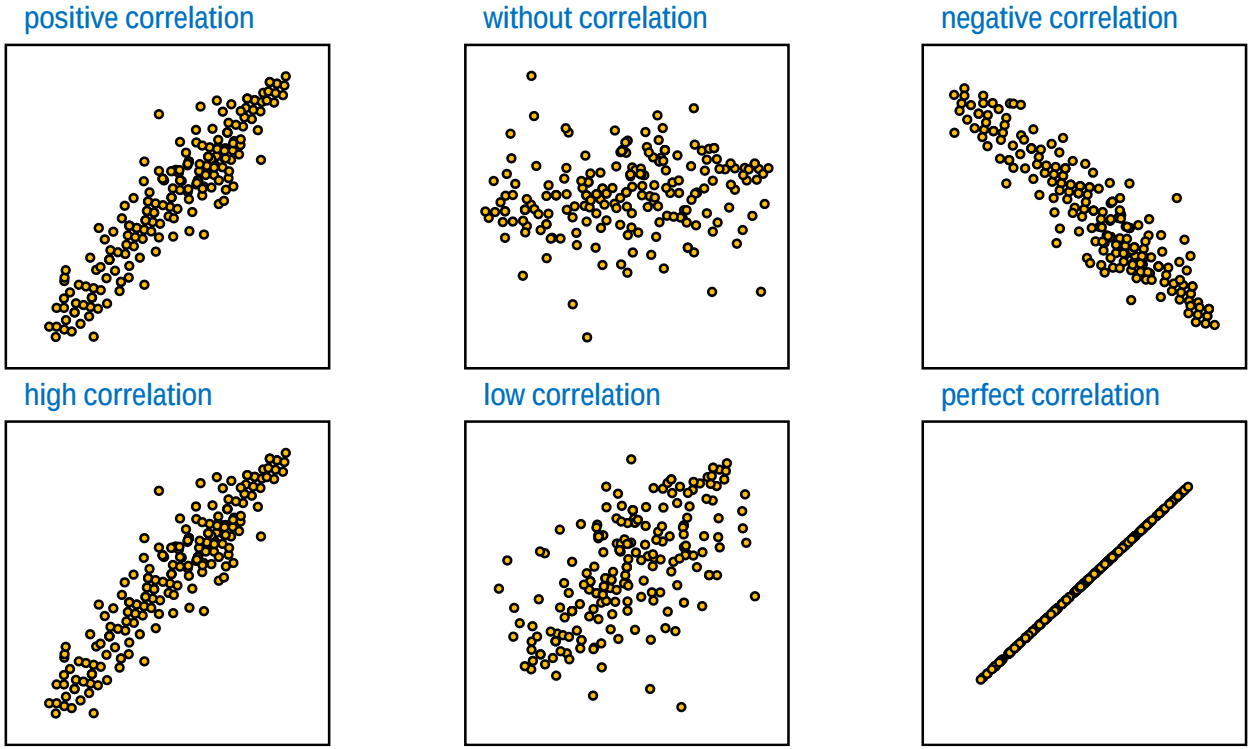
Hybridization of the same sample to 2 chips/channels

Visualizing the expression data – *log* transformation

- Makes variation of intensities and ratios of intensities more independent of absolute magnitude
- Evens out highly skewed distributions
- Gives more realistic sense of variation
- Approximates normal distribution
- Treats up- and down- regulated genes symmetrically

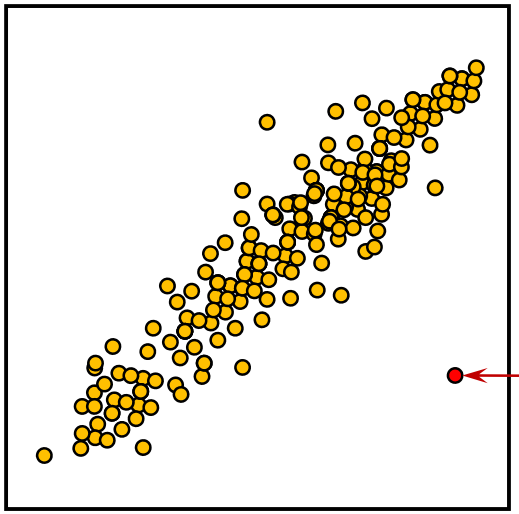



Original
Cy5 (red) intensity
Cy3 (green) intensity
Linear
Cy5 (red) intensity
Cy3 (green) intensity
Non-linear
Comet-tailing from
non-balanced channels
Cy5 (red) intensity
Cy3 (green) intensity


positive correlation
without correlation
negative correlation
high correlation
low correlation
perfect correlation

Correlation and Outliers

A single “bad” point can affect a good correlation, and the problem with microarrays is that we are expecting “bad” points



- Sigma (z) Score - the deviation from the mean divided by the standard deviation.

$$z_i = (x_i - \mu) / \sigma$$

- T Score (T) - a variation of the z score which produces a standard score with a range of 20 to 80 and a mean of 50.

$$T_i = 50 + 10 \times z_i$$

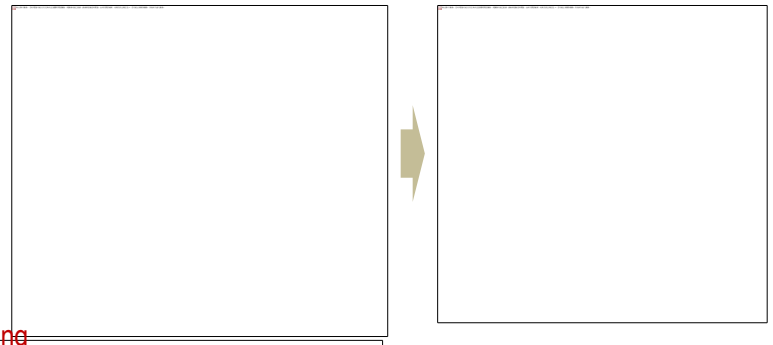


Global Mean method -cross slide normalization

- A single normalization factor (k) is computed for balancing chips/channels:

$$X_i^{norm} = k \times X_i \qquad k = \Sigma R_i / \Sigma G_i$$

- Multiplying intensities by this factor equalizes the mean (median) intensity among compared chips
  - Calculate mean intensity of all spots in channels 1 & 2
    e.g. μch1 = 20,000
    μch2 = 25,000
    μch2/μch1 = 1.25
  - On average, spots in ch2 are 1.25X brighter than spots in ch1
  - To normalize, multiply spots in ch1 by 1.25



Before Normalization

After Scaling

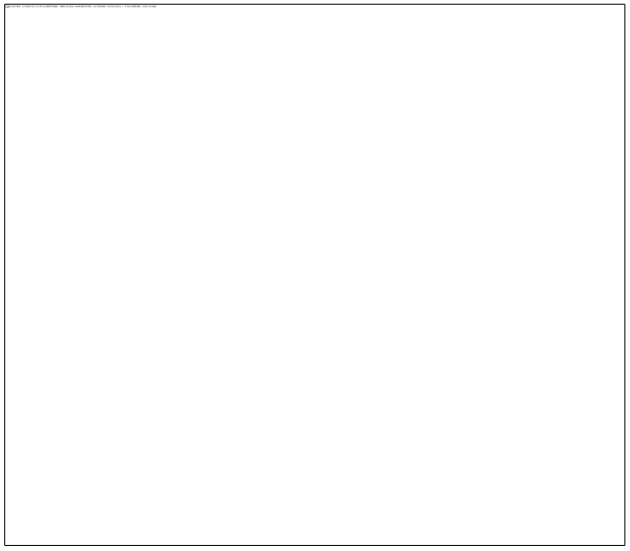


Global (*LOWESS*) Normalization -within slide normalization

$(R, G) \leftrightarrow (M, A)$ Transformation

"Observed" data $\{(R, G)\}_{n=1..n}$

R = red channel signal
G = green channel signal
(background corrected or not)



Transformed data $\{(M, A)\}_{n=1..n}$

$M = \log_2(R/G)$ (ratio),
$A = \log_2(R \times G)^{1/2} = 1/2 \times \log_2(R \times G)$ (intensity)

$\Leftrightarrow R=(2^{2A+M})^{1/2}$, $G=(2^{2A-M})^{1/2}$

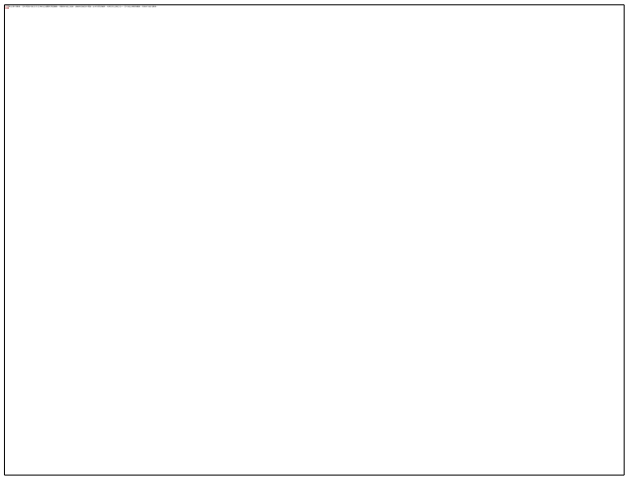


LOWESS Normalization

$\log_2 R'/G' = \log_2 R/G - \text{lowess}_i(A)$

Non-normalized data $\{(M, A)\}_{n=1..n}$

$M = \log_2(R/G)$



Global normalized data $\{(M, A)\}_{n=1..n}$

$M^{norm} = M_i - c_i(A) = \log_2 R/(k_i(A)G)$

$X_i^{norm} = k(A)*Xi$; $c(A) = \log_2(k(A))$

$c(A)$ is an *intensity dependent* function.
or c(A) is the (lowess fit) to the MA-plot.



Print-tip Normalization

Print-tip normalized data $\{(M, A)\}_{n=1..n}$

$$M_{p,\,norm} = M_p - c_p(A)$$

p = print tip (1-16)

where $c_p(A)$ is an *intensity dependent* function for print tip p.

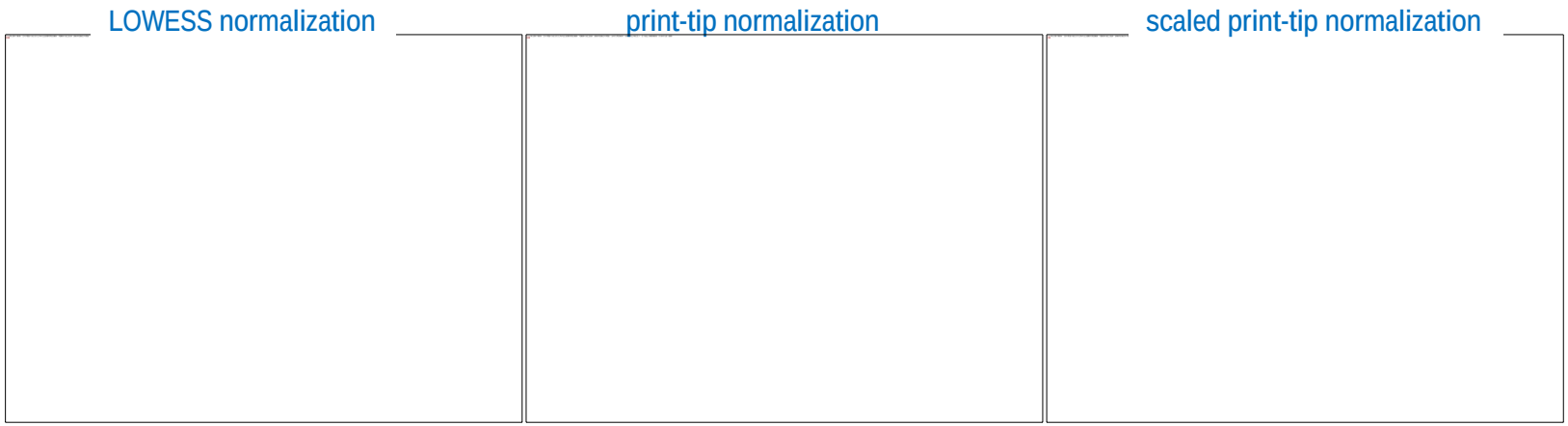
Scaled Print-tip Normalization

Scaled print-tip normalized data $\{(M,A)\}_{n=1..n}$

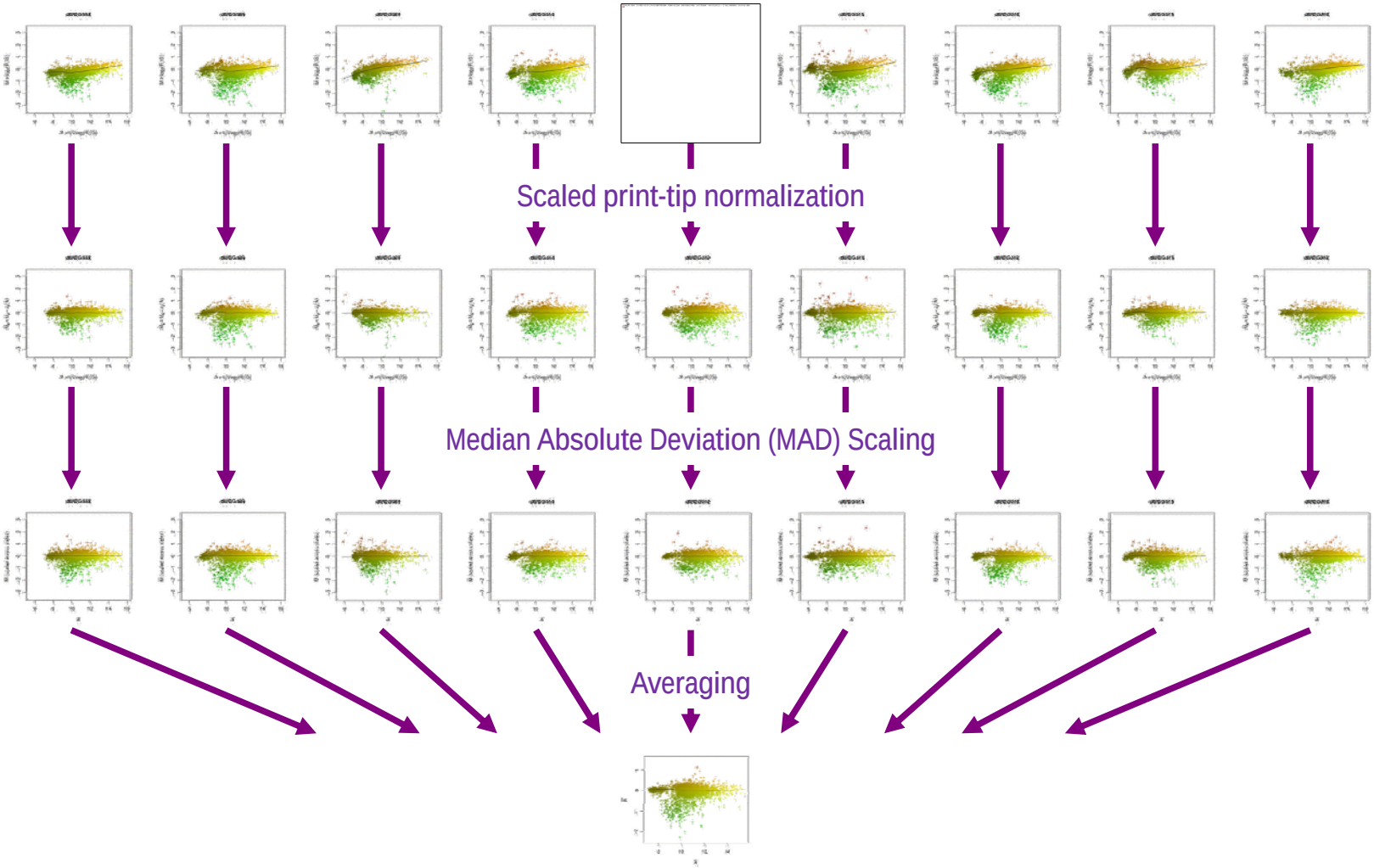
$$M_{p,\,norm} = s_p \times (M_p - c_p(A))$$

p = print tip (1-16)

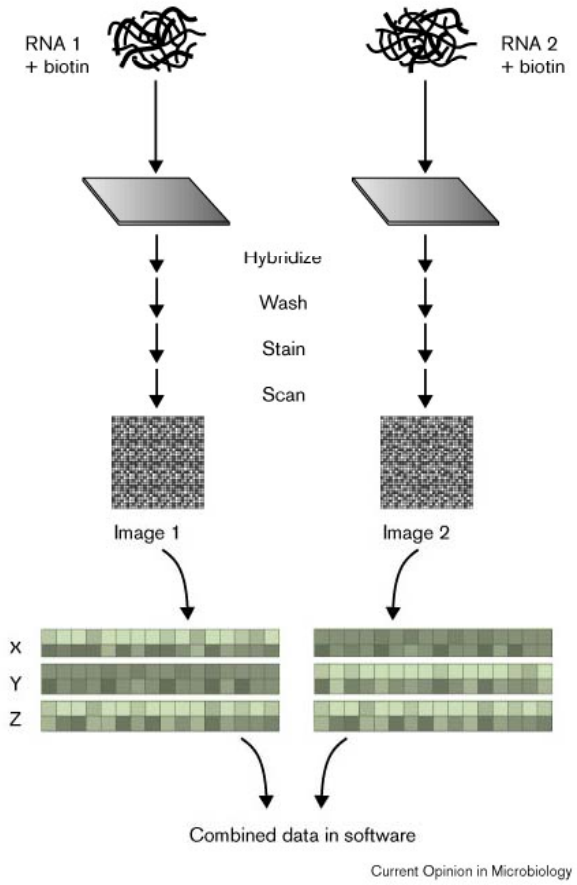
where s_p is a *scale factor* for print tip p (Median Absolute Deviation).



Across Slides Normalization



1. Experimental Design
2. Microarray Experiments
3. Image Processing
4. Data Normalization
5. **Data Analysis**



Samples

Platform

Raw Data *(image, .tif file)*

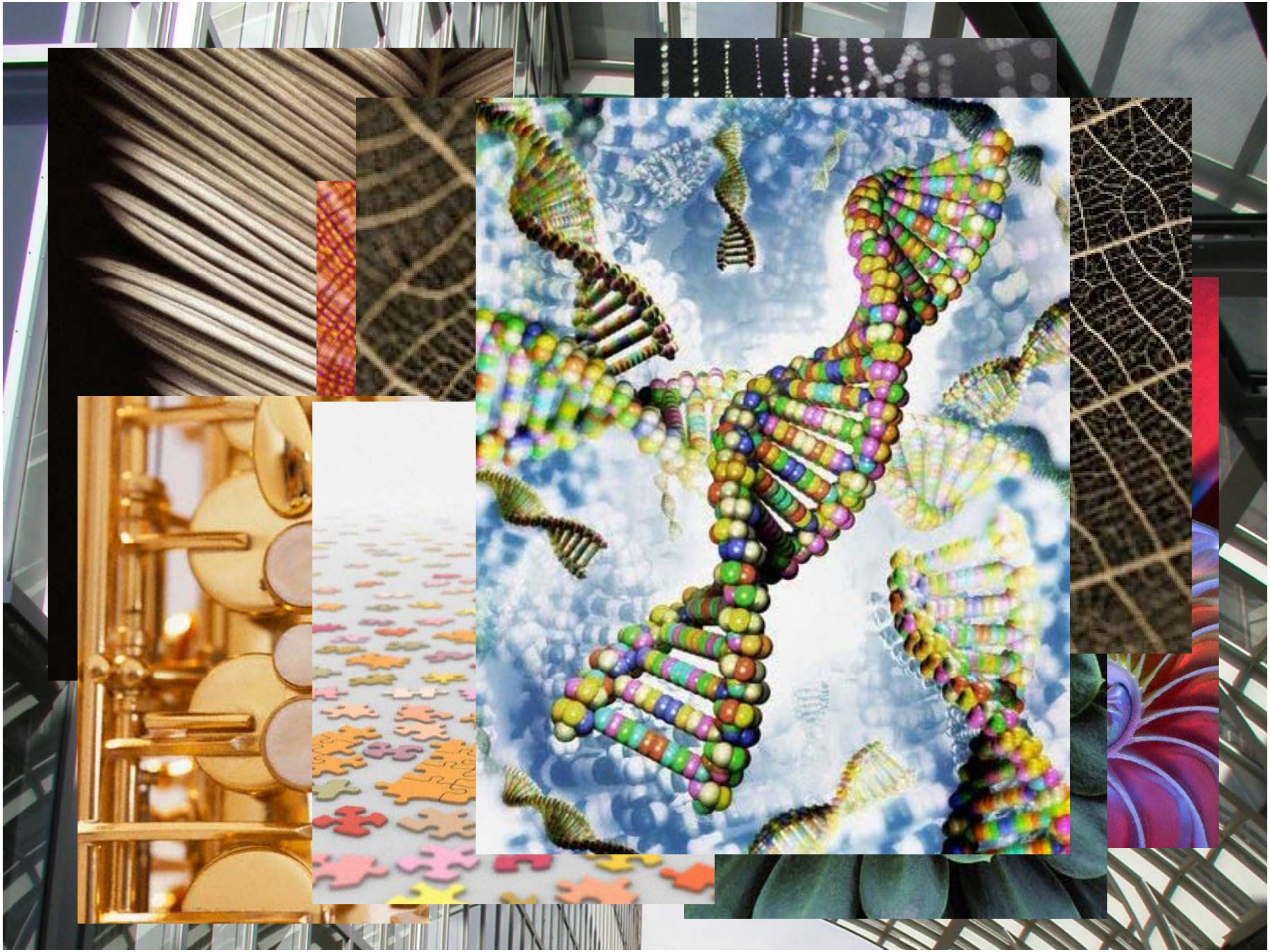
(1) image acquisition
(2) spot location

Preprocessed Data *(.gal file)*

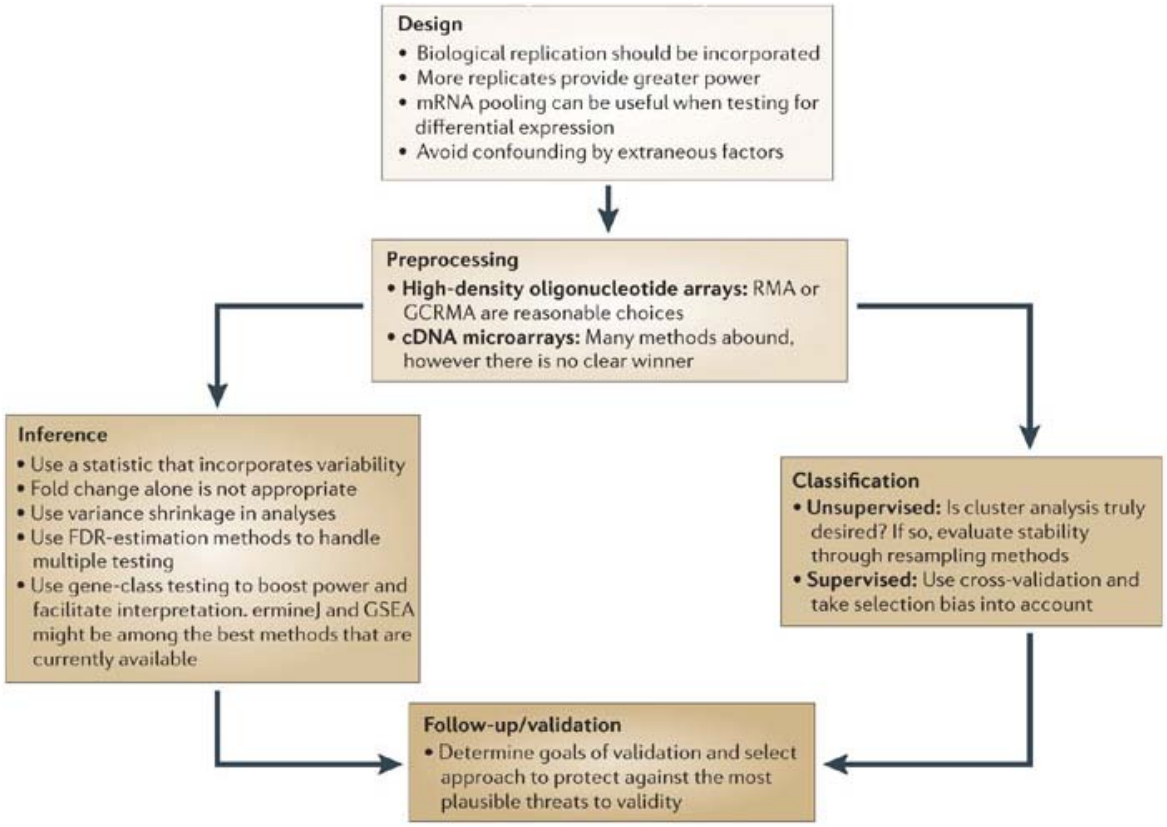
(3) computation of spot intensities

Data Analysis (different levels)

(4) data reporting



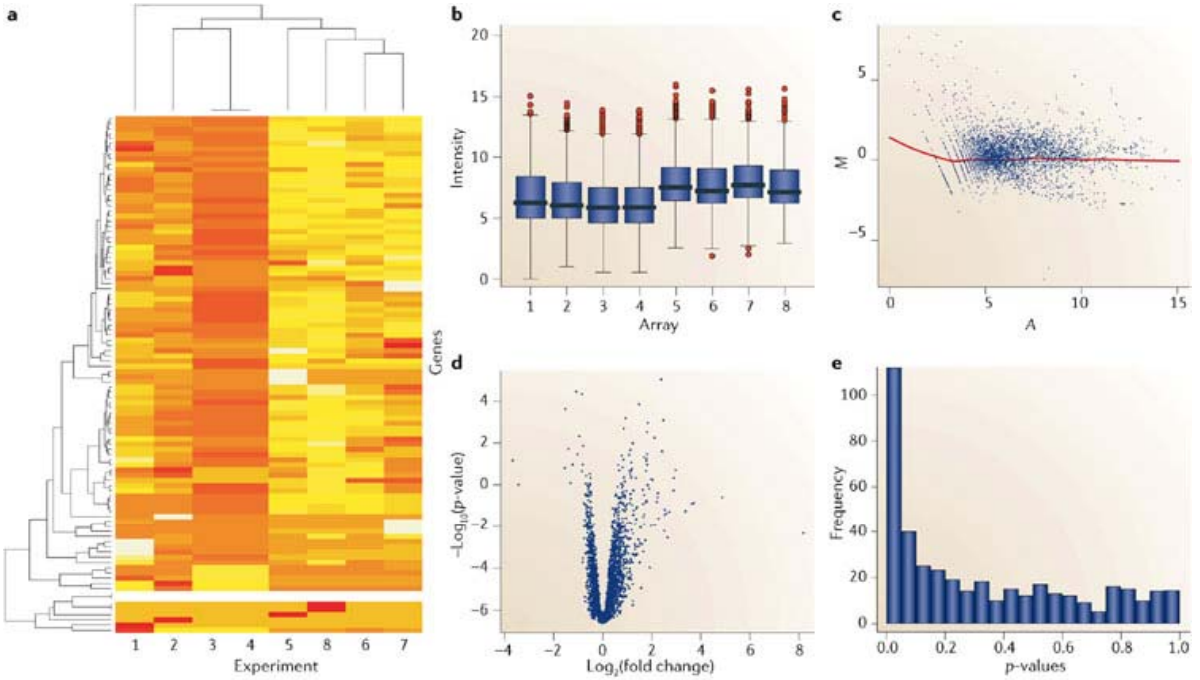
Guidelines for the statistical analysis of microarray experiments.



Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

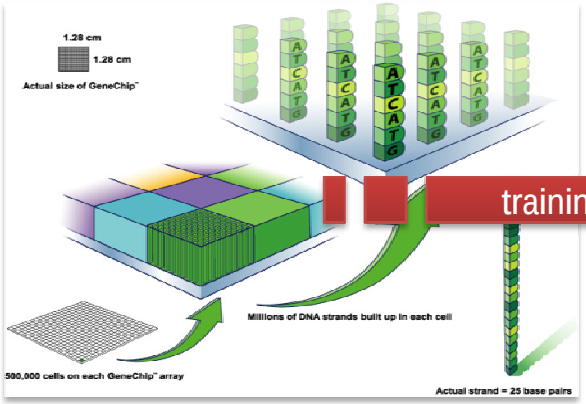
Allison DB *et al.* (2005) Microarray data analysis: from disarray to consolidation and consensus
Nat Rev gene. 7: 55–65 doi:10.1038/nri1749

1. Find the genes that change expression between experimental and control samples
2. Classify samples based on a gene expression profile
3. Find patterns: Groups of biologically related genes that change expression together across samples and treatments

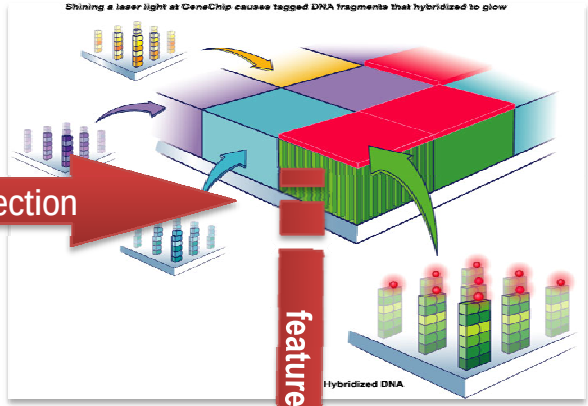


Allison DB *et al.* (2005) Microarray data analysis: from disarray to consolidation and consensus *Nat Rev gene.* 7: 55–65 doi:10.1038/nri1749

Single-Test Platform of Microarray & Knowledge Discovery



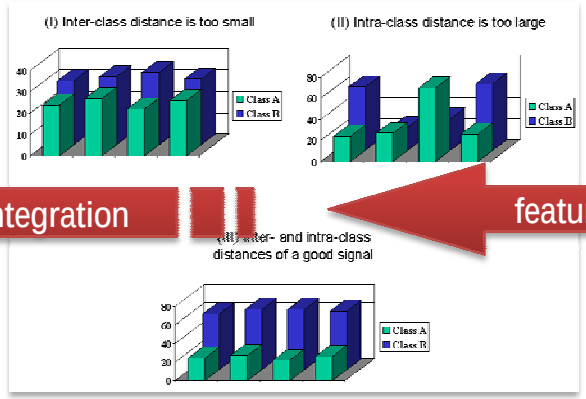
training data collection



feature generation

	00-0586-U	00-0586-U	00	U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative			Avg Diff	Abs Call	
AFFX-MurI	5	2			297.5	A	M16762 Mouse int
AFFX-MurI	3	2			554.2	A	M37897 Mouse int
AFFX-MurI	4	2			308.6	A	M25892 Mus musc
AFFX-MurI	1	3		19	141	A	M83649 Mus musc
AFFX-BioE	13	1		19	9340.6	P	J04423 E coli bioB
AFFX-BioB	15	0		19	12862.4	P	J04423 E coli bioB
AFFX-BioB	14	0		19	8716.5	P	J04423 E coli bioB
AFFX-BioC	15	0		19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0		20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0		19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0		20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0		20	280036.6	P	X03453 Bacterioph
AFFX-CreX	20	0		20	401741.8	P	X03453 Bacterioph
AFFX-BioE	7	5		18	-483	A	J04423 E coli bioB
AFFX-BioE	5	4		18	313.7	A	J04423 E coli bioB
AFFX-BioE	7	6		20	-1016.2	A	J04423 E coli bioB

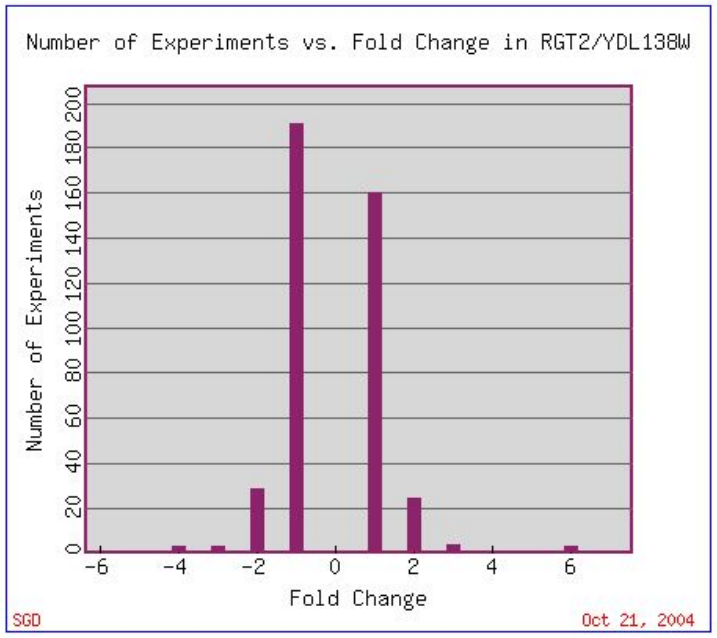
feature selection

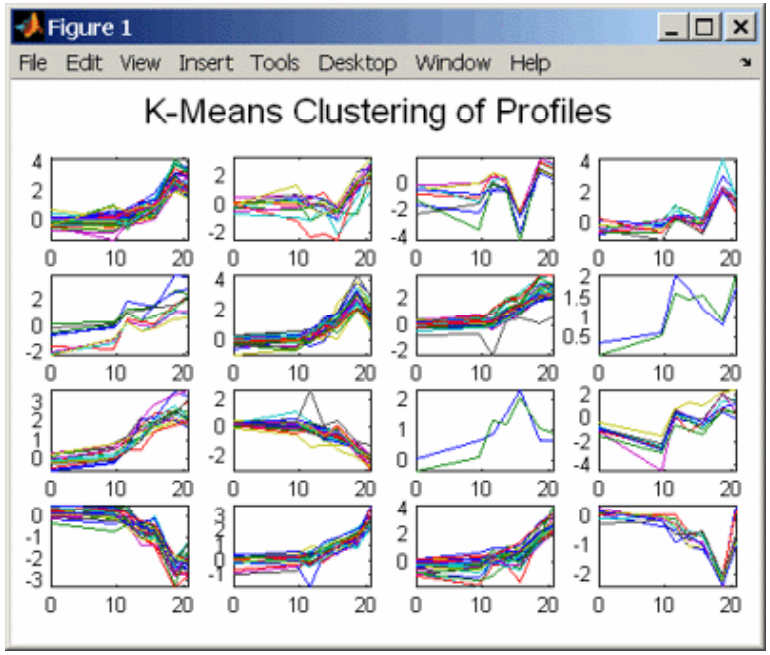


feature integration

Basic Data Analysis

- Fold change (relative increase or decrease in intensity for each gene)
- Set cutoff filter for low values (background +noise)
- Cluster genes by similar changes - only really meaningful across multiple treatments or time points
- Cluster samples by similar gene expression profiles





From : A Beginner's Guide Microarray Gene Expression Data Analysis

Table 4.2 Gene expression data matrix of absolute expression measurements for samples C1, C2, C3 and C4, given in abstract measurement units.

	C1	C2	C3	C4
G1	2	3	4	2
G2	2	3	4	20
G3	20	30	40	20
G4	200	300	400	200
G5	20	30	40	2
G6	20	30	40	1
G7	1	4	16	4
G8	4	5	4	4

Table 4.3 Gene expression data matrix containing relative expression measurements comparing conditions C1, C2 and C3 with the control condition C4. Compare with Table 4.2. Note that rather different absolute measurements, such as 4/2, 40/20 or 400/200, all give the same ratios.

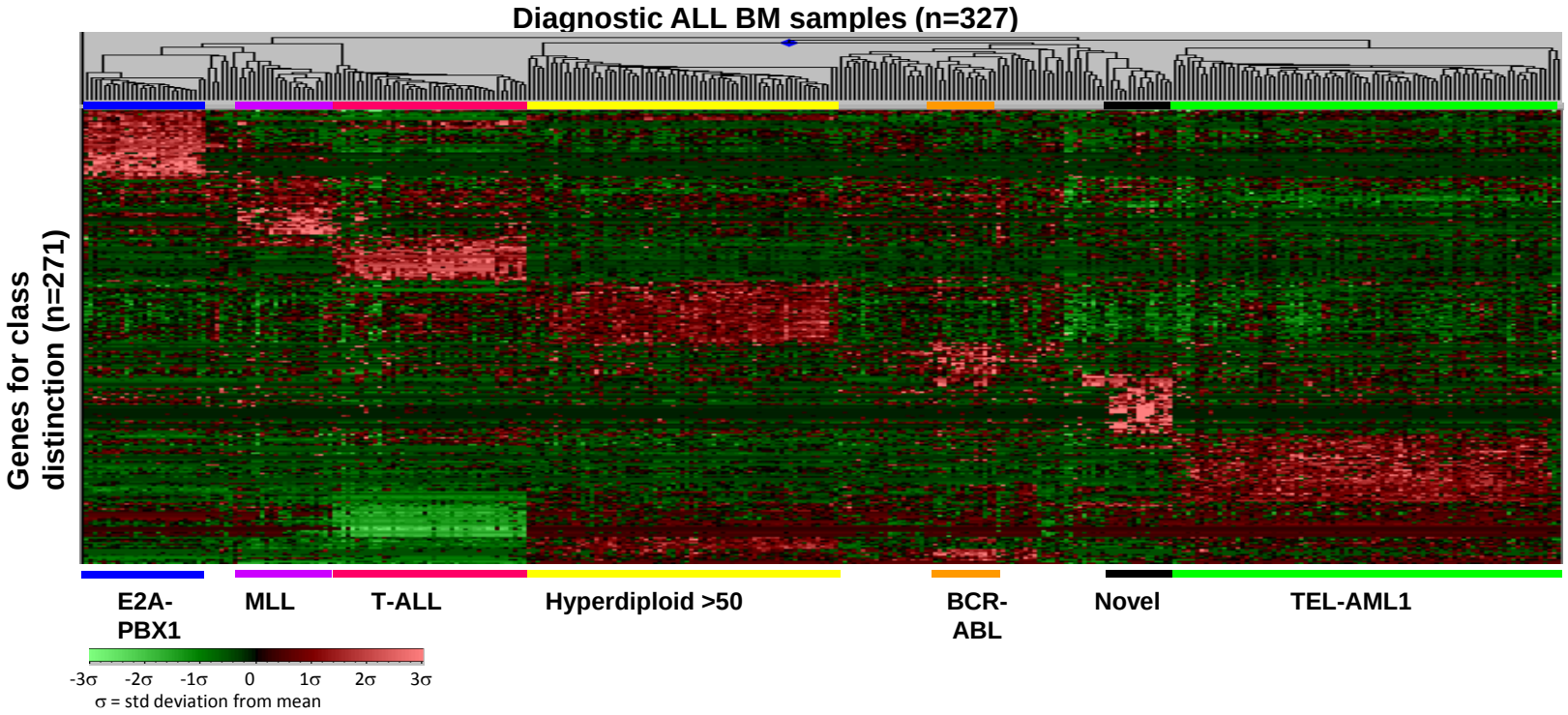
	r1 = C1/C4	r2 = C2/C4	r3 = C3/C4
G1	1	1.5	2
G2	0.1	0.15	0.2
G3	1	1.5	2
G4	1	1.5	2
G5	10	15	20
G6	20	30	40
G7	0.25	1	4
G8	1	1.25	1

Table 4.4 Gene expression data matrix containing log ratios (base 2) from Table 4.3. Compare with Table 4.3. Note that four-fold down- and up-regulation for gene G7 reflects as 0.25 and 4 in the ratio table, while in the log ratio table the values are symmetric, –2 and +2.

	log r1	log r2	log r3
G1	0	0.58	1
G2	–3.32	–2.74	–2.32
G3	0	0.58	1
G4	0	0.58	1
G5	3.32	3.91	4.32
G6	4.32	4.91	5.32
G7	–2	0	2
G8	0	0.32	0

Higher Level Microarray data analysis

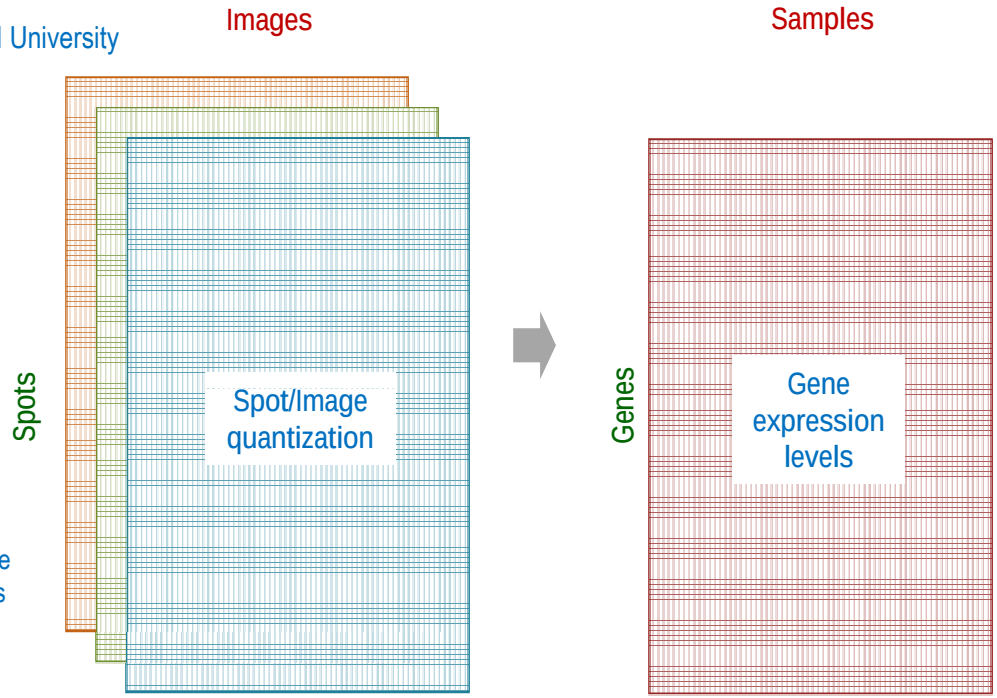
- Clustering and pattern detection
- Data mining and visualization
- Controls and normalization of results
- Statistical validatation
- Linkage between gene expression data and gene sequence/function/metabolic pathways databases
- Discovery of common sequences in co-regulated genes
- Meta-studies using data from multiple experiments



MIAME (Minimum Information About a Microarray Experiment)

MIAME is a standard for reporting microarray experiments. It is intended to specify all the information necessary to interpret the results of the experiment unambiguously and to potentially to reproduce the experiment. While the standard defines the content required for compliant reports, it does not specify the format in which this data should be presented. There are a number of file formats for representing this data, and both public and subscription-based repositories for such experiments. (http://www.mged.org/Workgroups/MIAME/miame_1.1.html)

- A need to establish a public repository or repositories for microarray gene expression data became apparent in 1998, which requires data standards
- MGED-1 (Microarray Gene Expression Database) Group: November 14-15, 1999, Cambridge, UK
  - Established *five* working groups, including the microarray data annotation group (MIAME)
- MGED-2: May 25 - 27, 2000, Heidelberg, Germany
  - Endorsed a MIAME draft
- MGED-3: March 29-31, 2001, Stanford University
  - Adopted MIAME 1.0
- MGED-4: Feb. 13-16, 2002, Boston
  - Adopted MIAME 1.1

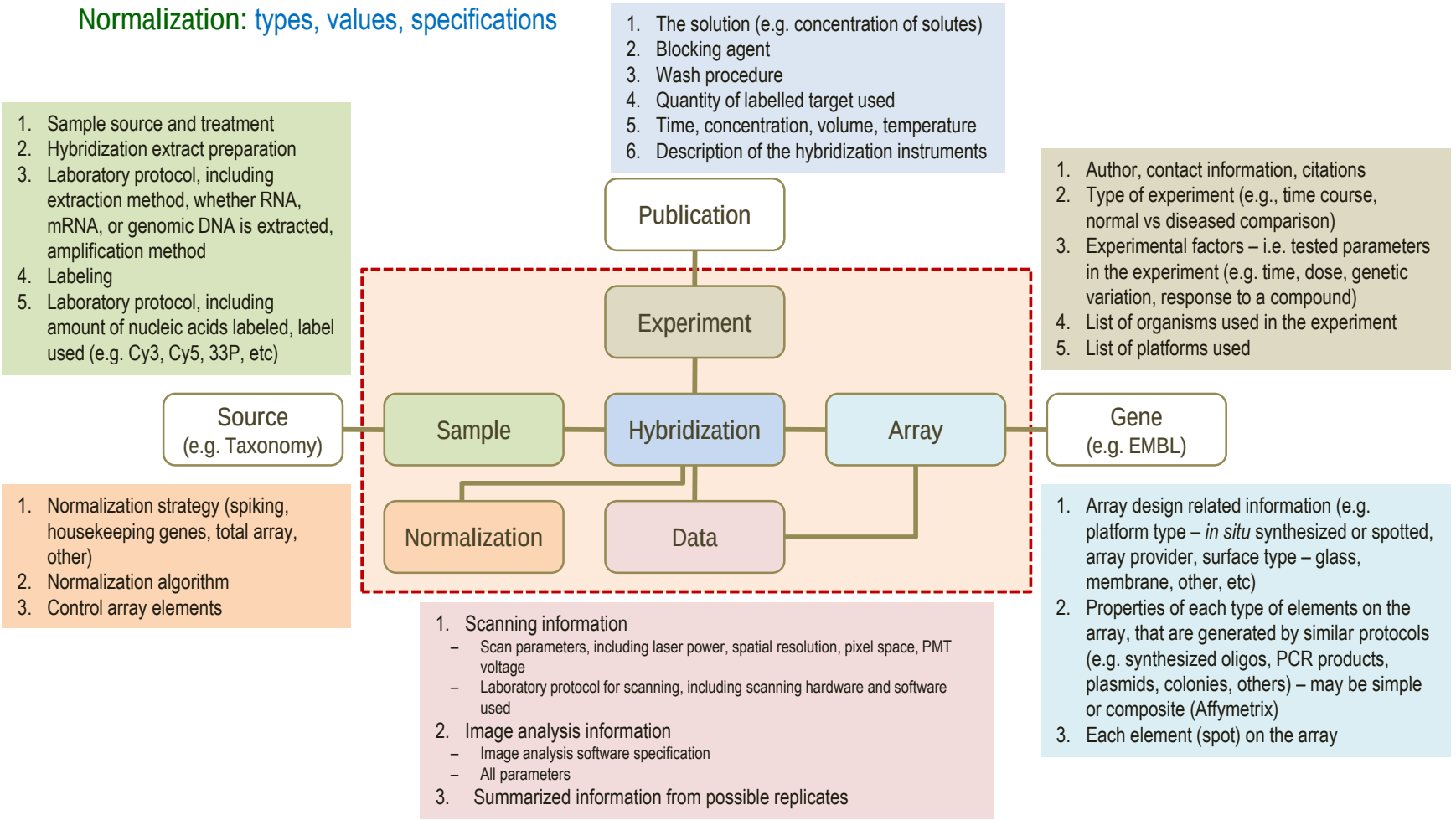


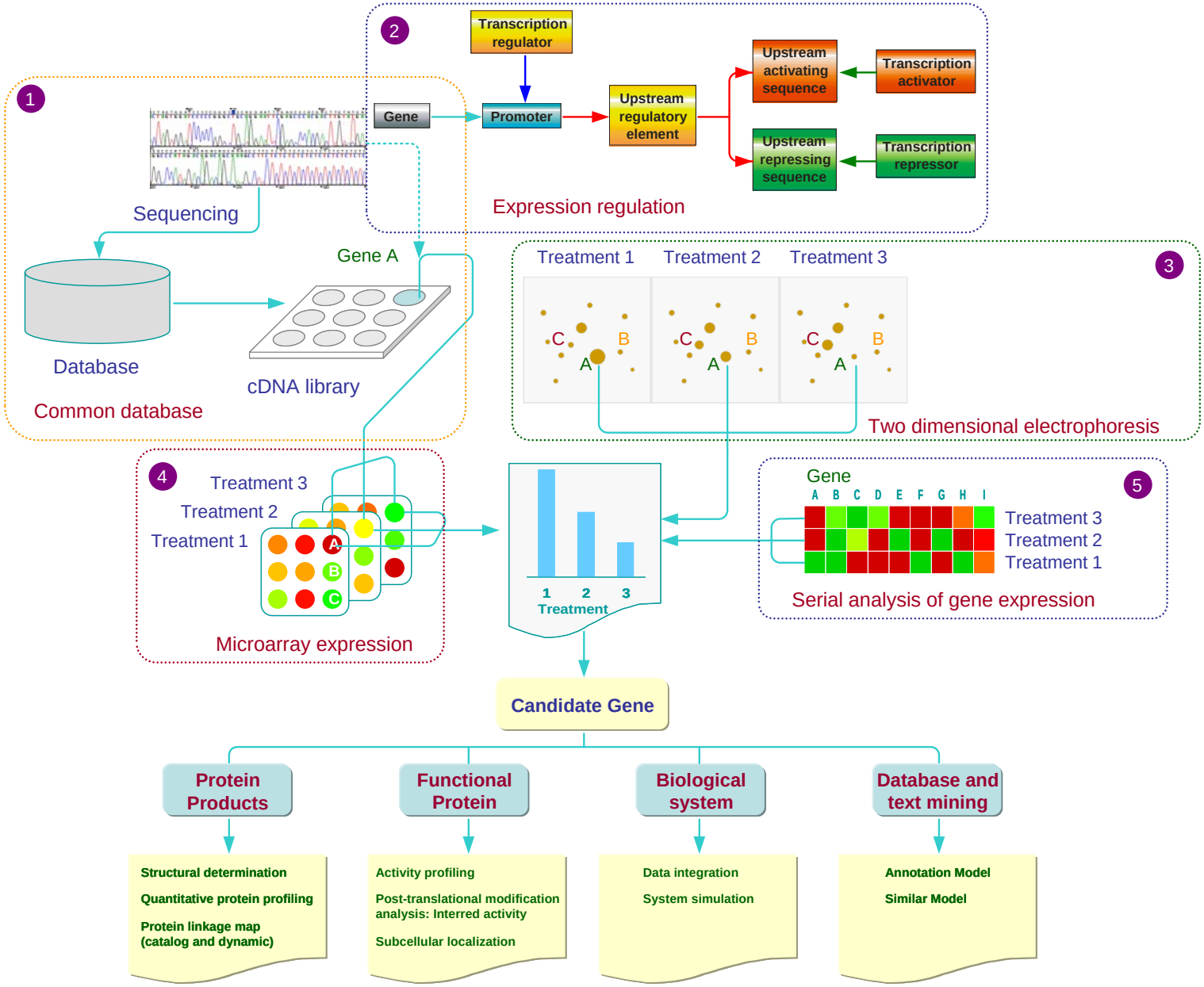
Gene Expression Matrix

The final gene expression matrix (on the right) is needed for higher level analysis and mining.

MIAME: six parts

Experimental Design: the set of the hybridization experiments as a whole
Array Design: each array used and each element (spot) on the array
Samples: samples used, the extract preparation and labeling
Hybridizations: procedures and parameters
Measurements: images, quantitation, specifications
Normalization: types, values, specifications



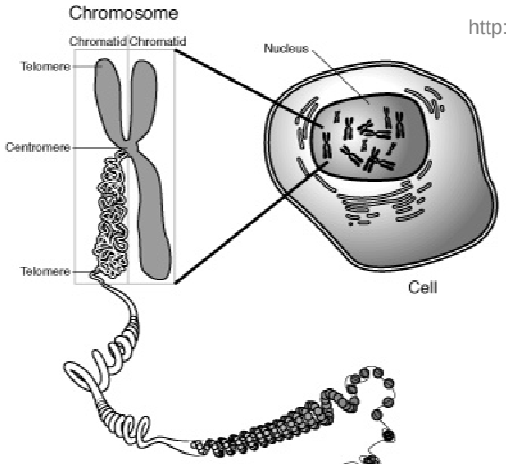

1
2
3
4
5
Transcription regulator
Gene
Promoter
Upstream regulatory element
Upstream activating sequence
Transcription activator
Upstream repressing sequence
Transcription repressor
Expression regulation
Sequencing
Gene A
Database
cDNA library
Common database
Treatment 1
Treatment 2
Treatment 3
Two dimensional electrophoresis
Treatment 3
Treatment 2
Treatment 1
Microarray expression
1 2 3
Treatment
Gene
A B C D E F G H I
Treatment 3
Treatment 2
Treatment 1
Serial analysis of gene expression
Candidate Gene
Protein Products
Functional Protein
Biological system
Database and text mining
Structural determination
Quantitative protein profiling
Protein linkage map (catalog and dynamic)
Activity profiling
Post-translational modification analysis: Interred activity
Subcellular localization
Data integration
System simulation
Annotation Model
Similar Model

Data Analysis Strategy

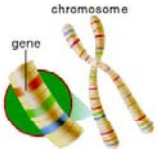
3 Pathway



http://cgap.nci.nih.gov/Pathways



2 Cytoband



4 GO term



- Molecular function
- Biological process
- Cellular component

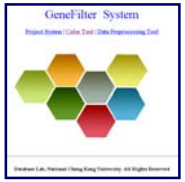
http://www.ebi.ac.uk/GO/index.html

5 Text mining



http://140.116.247.48:8080/IGIS/index.jsp

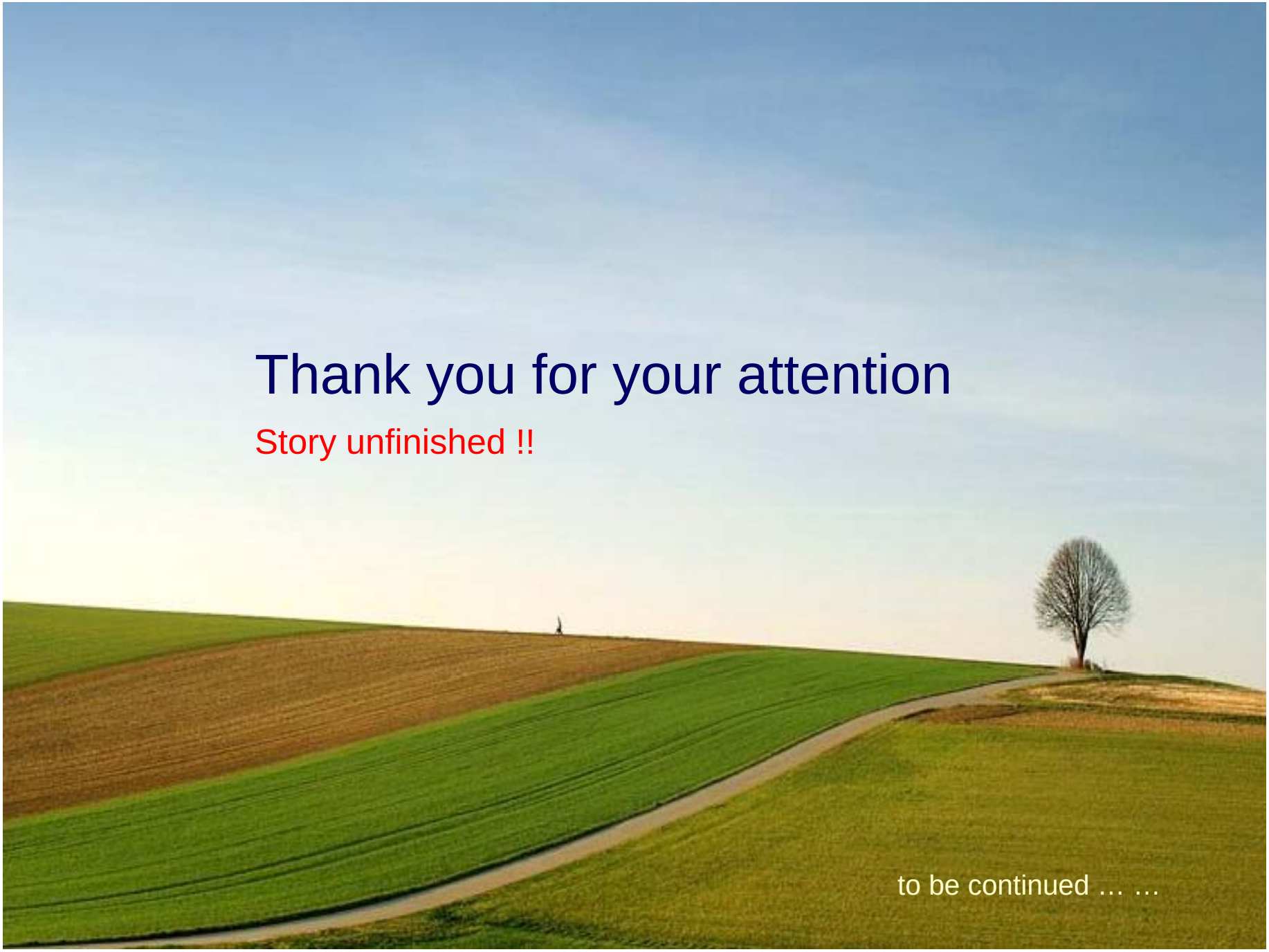
1 Expression



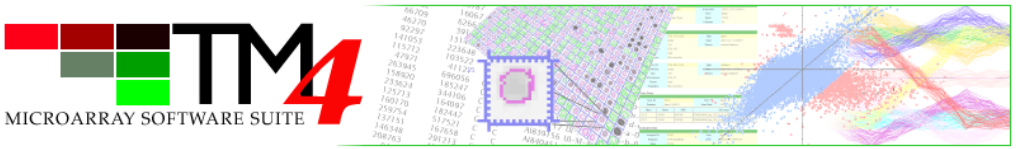
- Up regulation
- Down regulation
- Yoyo expression

http://biosys.csie.ncku.edu.tw/genefilter2/index.jsp

FIGURE SOURCE: http://www.accessexcellence.org/AB/GG/chromosome.html

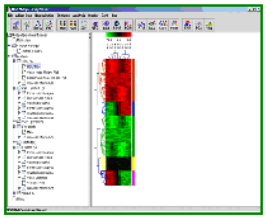

Thank you for your attention
Story unfinished !!
to be continued … …

http://www.tm4.org/mev.html



MeV: MultiExperiment Viewer

Home
MADAM
Spotfinder
MIDAS
MeV
AMP
Utilities
FAQ
Formats
Contributors
Contact Us



Normalized and filtered expression files can be analyzed using TIGR Multiexperiment Viewer (MeV). MeV is a versatile microarray data analysis tool, incorporating sophisticated algorithms for clustering, visualization, classification, statistical analysis and biological theme discovery. MeV can handle several input file formats. These include the ♦.mev♦ and ♦.tav♦ files generated by TIGR Spotfinder and TIGR MIDAS, and also Affymetrix® (♦.txt♦) and Genepix® (♦.gpr♦) files. MeV generates informative and interrelated displays of expression and annotation data from single or multiple experiments. At this final stage of the TM4 pipeline, flexibility and the variety of analysis techniques are critical, as every algorithm has strengths that can be exploited when used on certain datasets and experimental designs. The concept of modularization lends itself particularly well to this system, as novel algorithms and existing codebases from the microarray community can be integrated with the Java based MeV using a well-defined module API.

Many of the algorithms currently implemented are listed below. Bootstrapping, jackknifing and k-means support resample the dataset to generate consensus clusters. Figure of merit graphs (FOM; (Yeung et al., 2001) suggest appropriate input parameters for algorithms such as k-means. Modules to allow metabolic pathways and genomic/chromosomal maps to be viewed with expression data overlaid are in development and testing. A wizard to handle links to public database websites is also being developed. Clusters identified through any analysis method can be labeled and tracked through other analyses, providing the user with the ability to compare the results of several clustering algorithms to determine consensus and focus on genes with specified expression patterns and biological profiles.

Keep up to date by subscribing to the **MeV mailing list**.

Latest Version

Download the latest version of MeV: **v4.0**
Or download the source code:
v4.0 source
Download the previous version of MeV: **v3.1**
The v4.0 **release notes** describe changes made since version 3.1.

MeV Manual

The **Manual** is a detailed description of the workings of MeV.

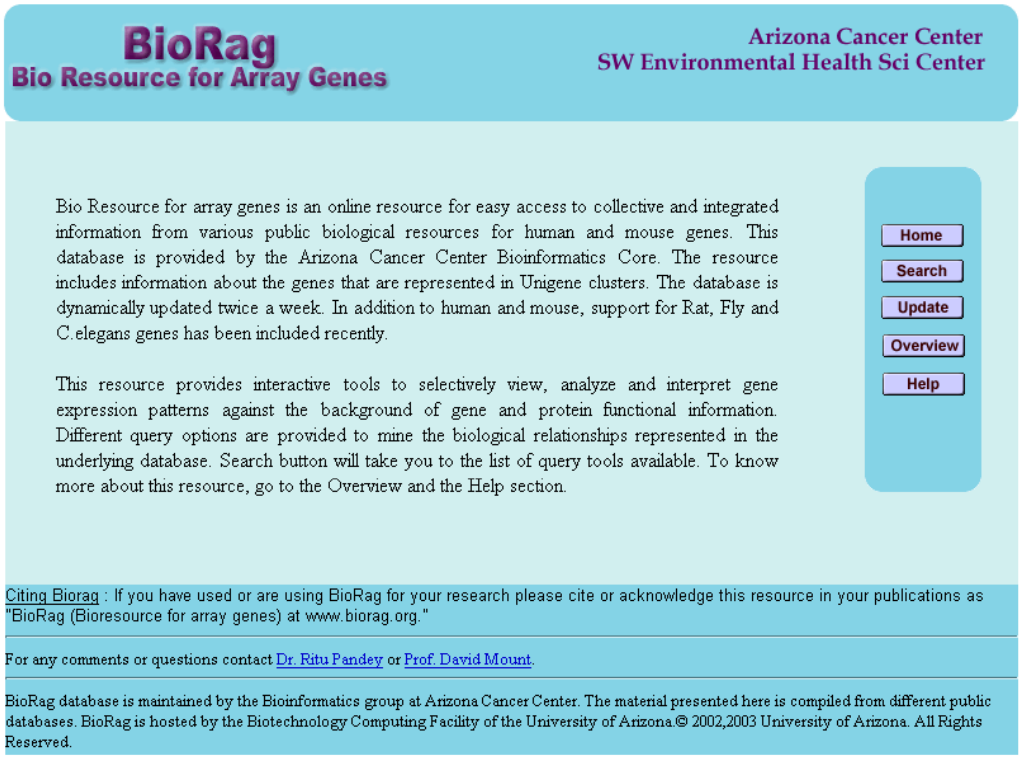
System Requirements

MeV v4.0 runs on the Windows, Mac OSX or Linux operating systems and requires Java v1.4.1 or higher.
Java 3D is required for some MeV modules, like PCA.
Download **Java v1.4.2**
Download **Java 3D v1.3.1**

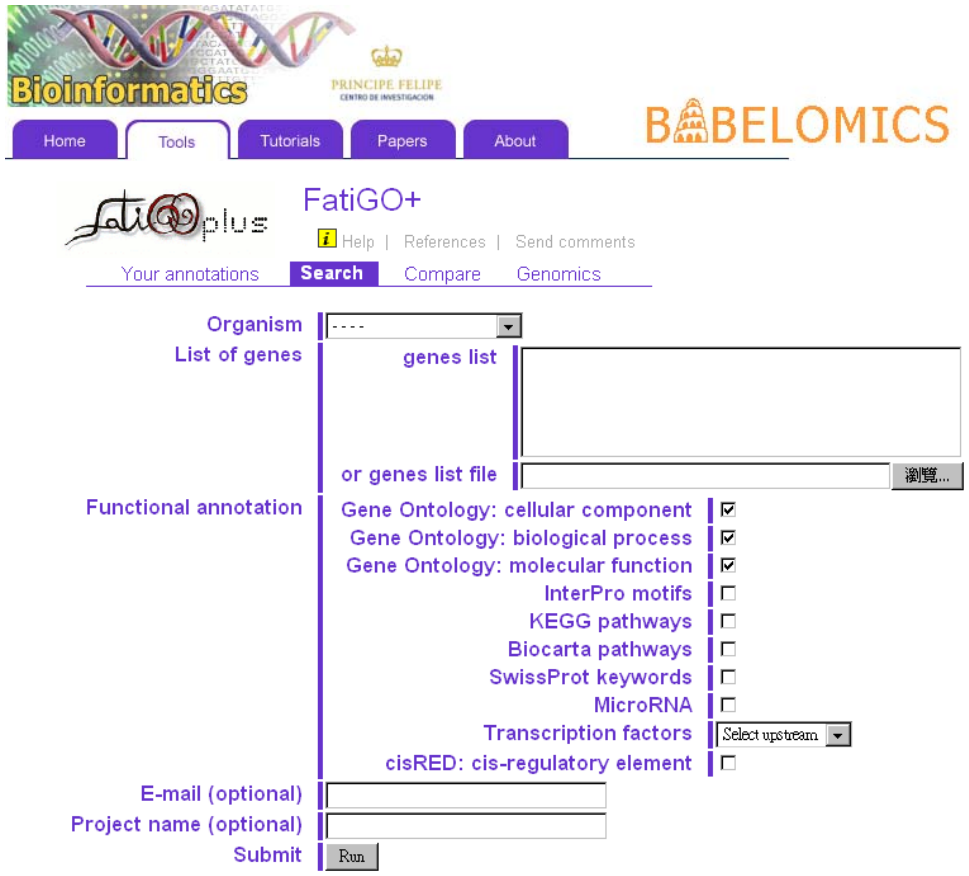
Training Documents

The powerpoint **slides** here contain the slides our group uses to teach courses on the program.

http://www.biorag.org/index.html



http://babelomics.bioinfo.cipf.es/fatigoplus/cgi-bin/fatigoplus.cgi



http://cgap.nci.nih.gov/

The CANCER GENOME ANATOMY PROJECT

The goal of the NCI's Cancer Genome Anatomy Project is to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. By collaborating with scientists worldwide, such as the Ludwig Institute for Cancer Research and Lund University, CGAP seeks to increase its scientific expertise and expand its databases for the benefit of all cancer researchers.

The CGAP Web Site

Interconnected modules provide access to all CGAP data, bioinformatic analysis tools, and biological resources allowing the user to find "in silico" answers to biological questions in a fraction of the time it once took in the laboratory.



Genes
Gene information, clone resources, SNP500Cancer, GAI, and transcriptome analysis NEW!



Tissues
cDNA library information, methods, and EST-based gene expression analysis



Pathways
Diagrams of biological pathways and protein complexes, with links to genetic resources for each known protein



RNAi NEW!
RNA-interference constructs, targeted specifically against cancer relevant genes.



Chromosomes
FISH-mapped BAC clones, SNP500Cancer, and the Mitelman database of chromosome aberrations



SAGE Genie
Analysis of gene expression using **long** NEW! and **short** SAGE tag data for both **human** and **mouse** NEW!



Tools
Direct access to all analytic and data mining tools developed for the project