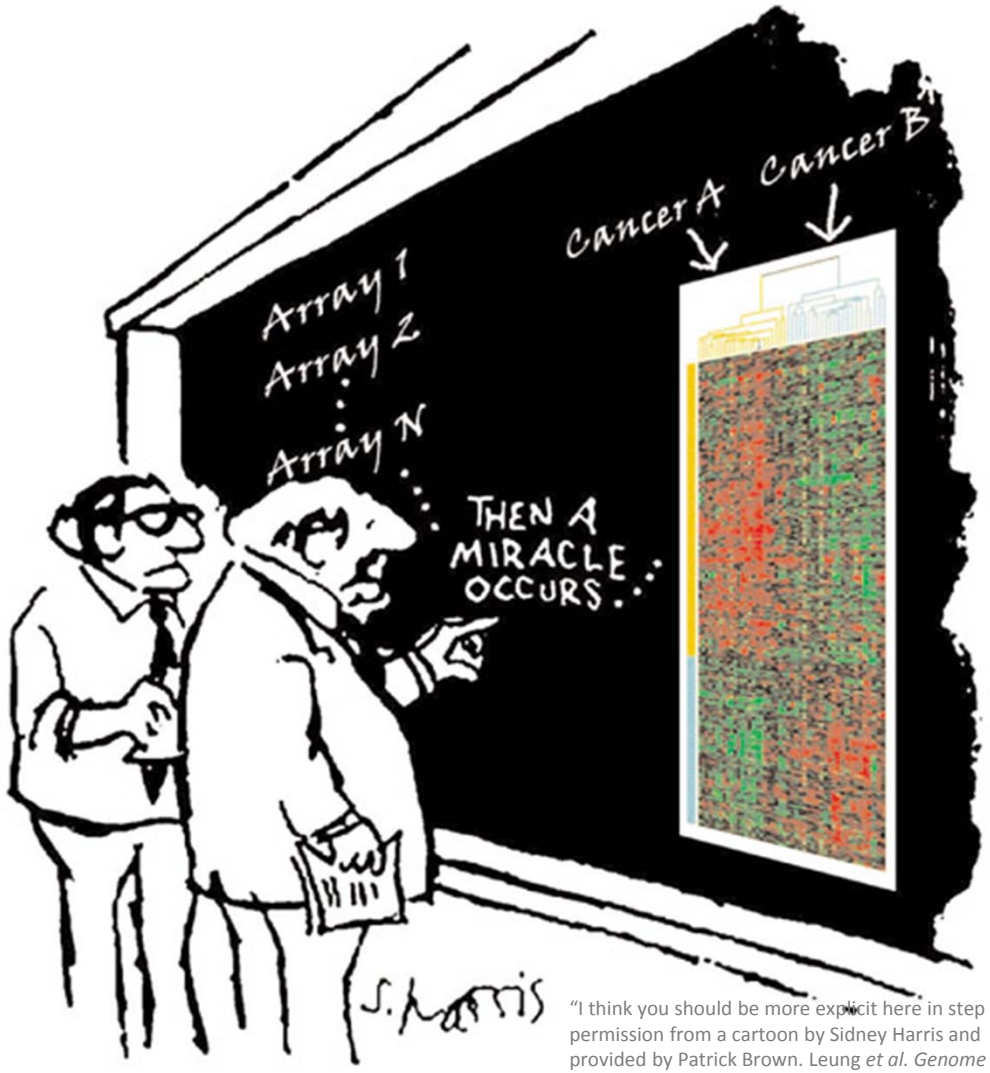
MICROARRAYS:

Chipping Away at the Mysteries of Science and Medicine



"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 discovery

van Leeuwenhoek
described single celled organisms



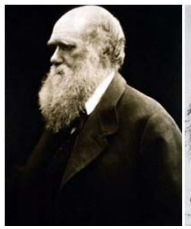
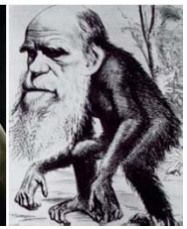

1676

Carl Linnaeus
Hierarchical classification of species




1735

Charles Darwin
"The Origin of Species"

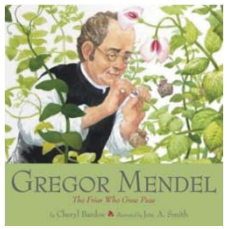
1859

Louis Pasteur
Microorganisms responsible for contamination, heating kills microorganisms



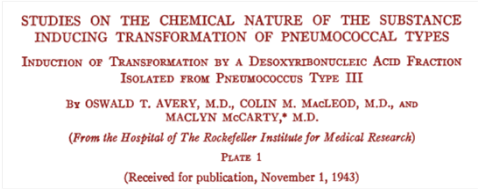

1862

Gregor Mendel
Phenotype determined by inheritable units

1866

Avery, MacLeod, McCarty
DNA is the genetic material



STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES

INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III

BY OSWALD T. AVERY, M.D., COLIN M. MACLEOD, M.D., AND MACLYN MCCARTY,* M.D.

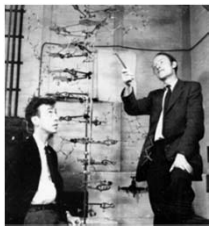
(From the Hospital of The Rockefeller Institute for Medical Research)

PLATE 1

(Received for publication, November 1, 1943)

1944

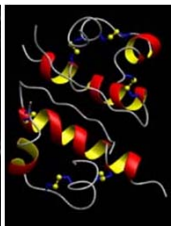
James Watson & Francis Crick
Solve structure of DNA



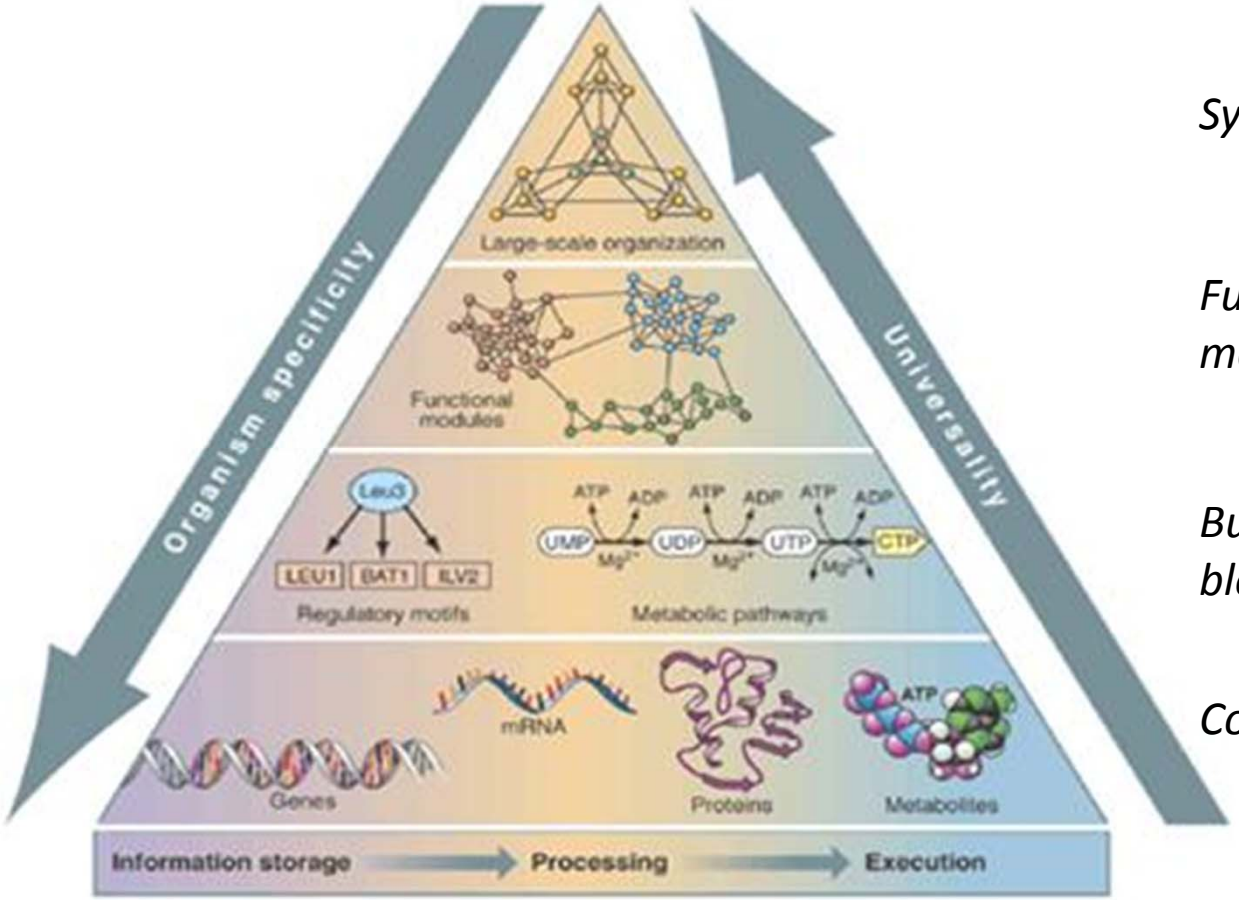

1953

1955

Frederick Sanger
Complete sequence of insulin

Life‘s Complexity Pyramid



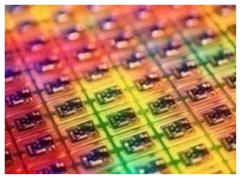
System

Functional modules

Building blocks

Components

*Z. N. Oltvai and A.-L. Barabási, Science **298**, 763 (2002)*



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.

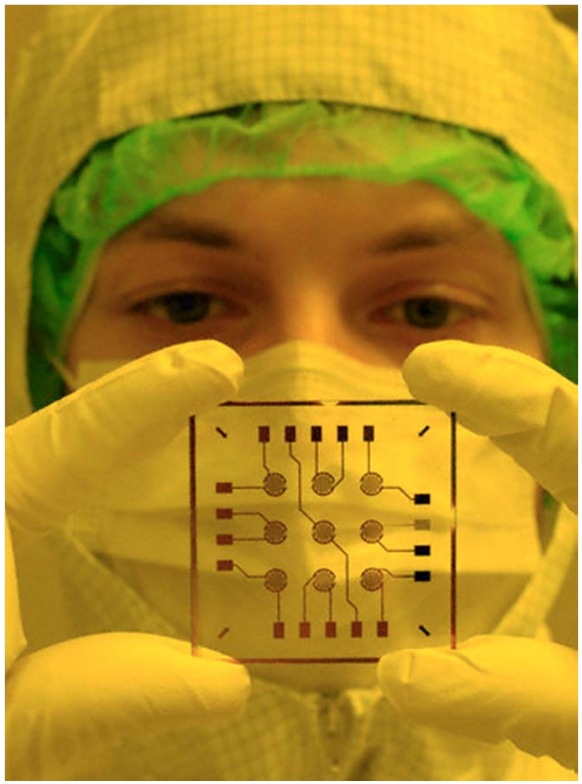
Multi-dimensional arrays are accessed using more than one index: one for each dimension.

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.

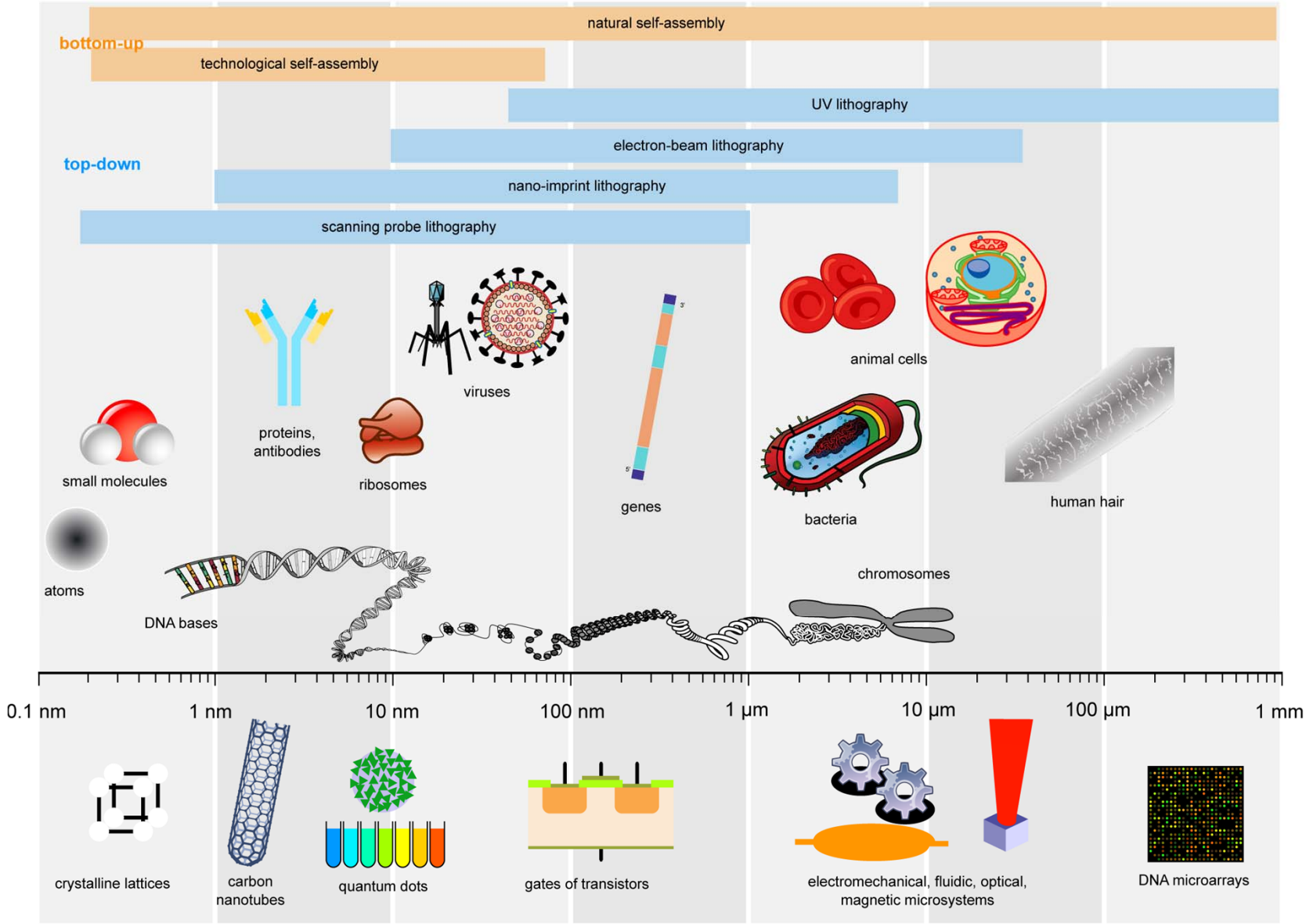


微

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.

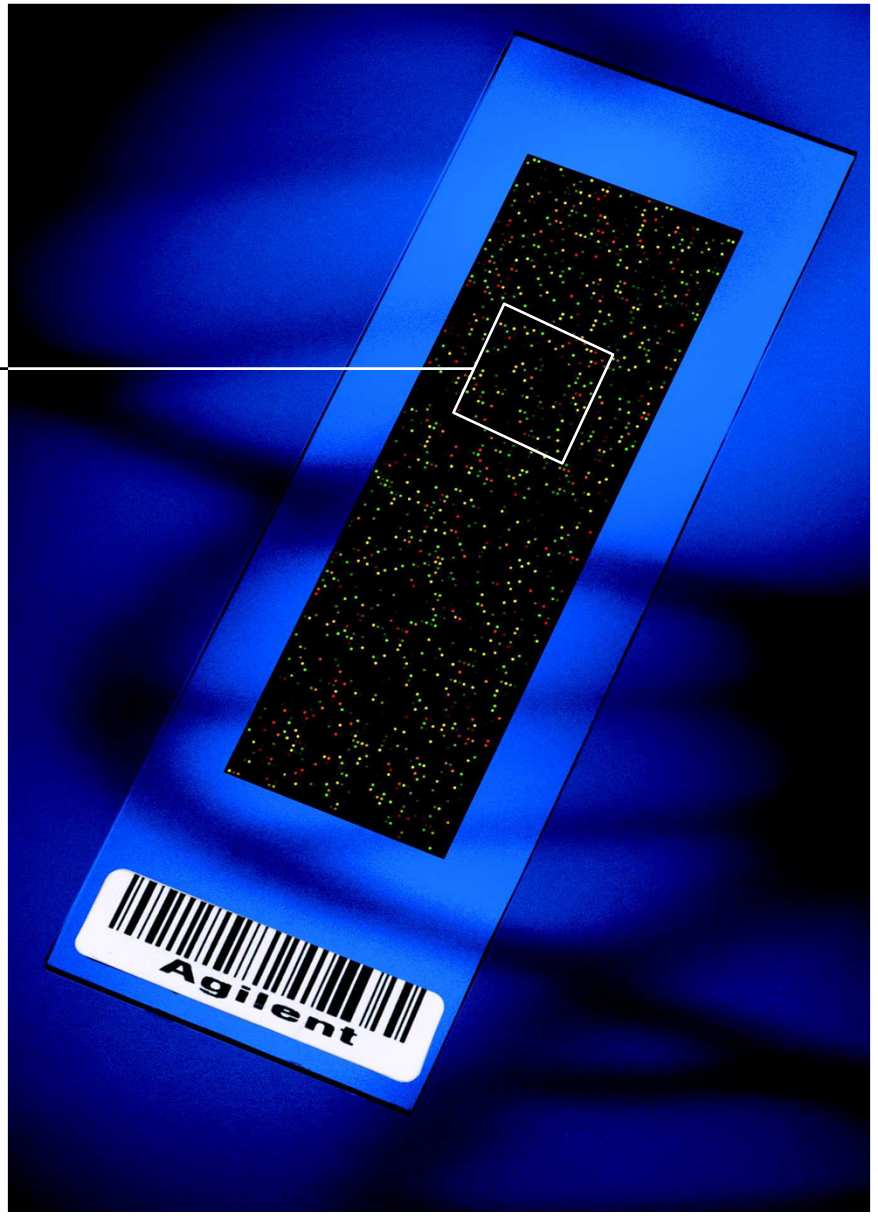


Comparison of various biological assemblies and technological devices



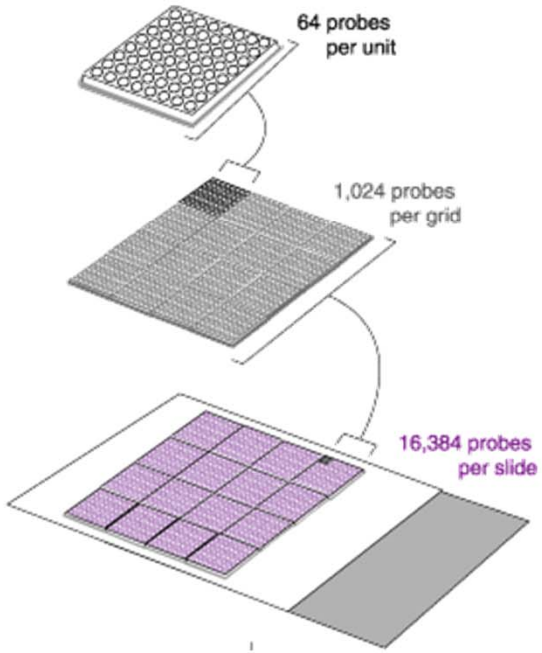
Microarrays are miniaturized biological devices consisting in molecules, for example DNA or protein, named the "probes", that are orderly arranged at a microscopic scale onto a solid support such as a membrane or a glass microscope slide.

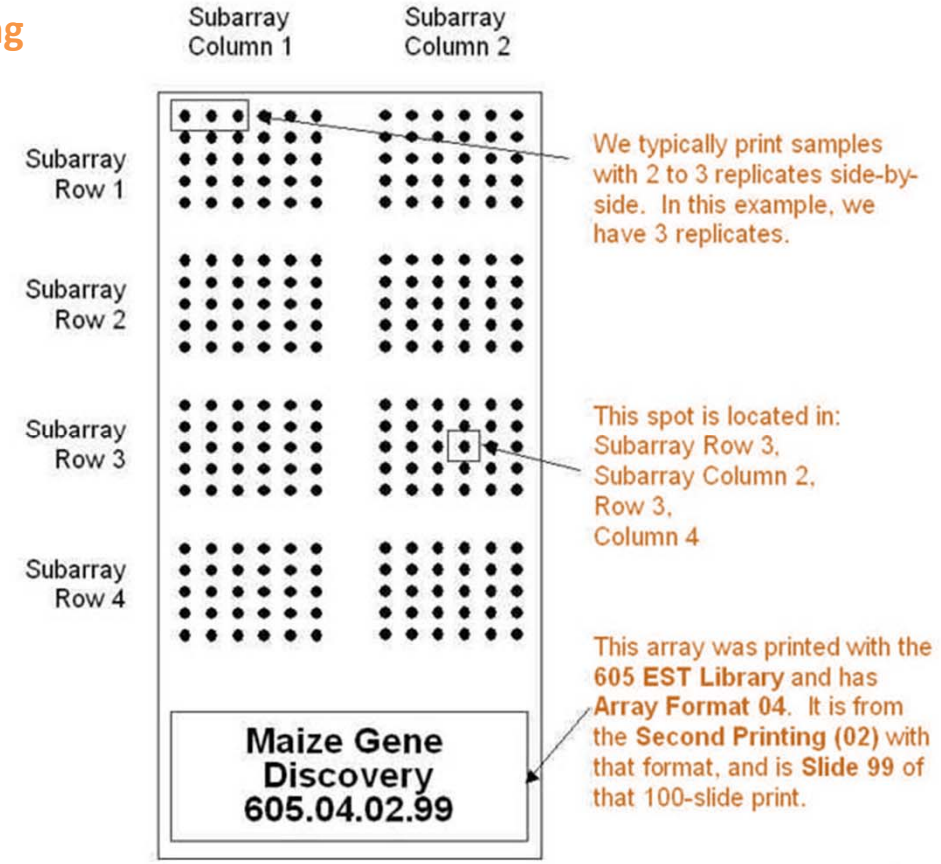




Microarray Features

1. Parallelism: Thousands of genes simultaneously
2. Miniaturization: Small chip size
3. Multiplexing: Multiple samples at the same time
4. Automation: Chip manufacturing and Reagents





What circumstances brought out this technology?

1. Fact: biological processes are complicated with many molecules working together.

   Biologists are eager to obtain the "**whole picture**"

2. Genome sequences availability

3. Computer aids



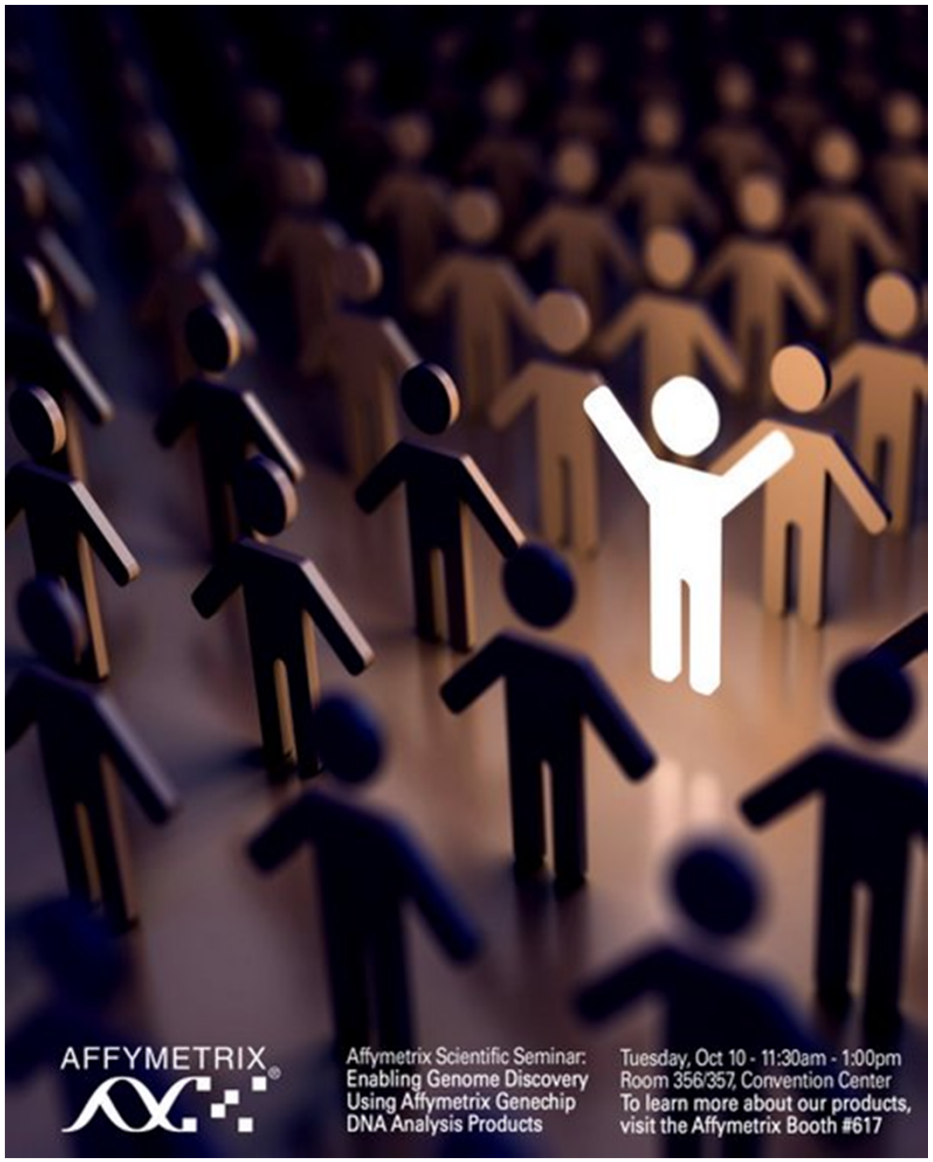
Detect 1.8 million genetic variations on a single array.



AFFYMETRIX

What problems can it solve?

1. Differing expression of genes over time, between tissues, and disease states
2. Identification of complex genetic diseases
3. Drug discovery and toxicology studies
4. Mutation/polymorphism detection (SNP's)
5. Pathogen analysis

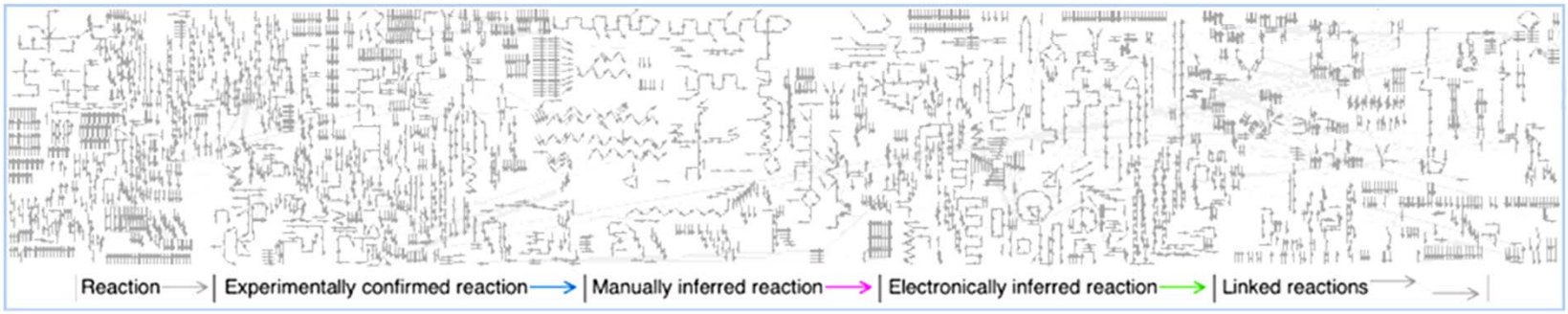


What is its PITFALL?

1. Detect transcription mRNA level, not translation protein level
2. Many factors (variations) can affect the result
    - Chip and probe design
    - Experiment design
    - Sample preparation
    - Image acquisition
    - Data normalization
    - Data analysis
    - ….
3. Success crucial:
    - You know both the **biology problem** and the **computer aids** (software, statistics).

Why Use DNA Microarrays for Expression Analysis?

1. Conventional expression analysis only allows the study of the expression of a single gene in a single experiment.
2. The highly parallel nature of microarrays allows the simultaneous study of the expression of thousands or even tens of thousands of different genes in a single experiment.
3. Microarrays allow researchers to undertake global expression analysis that is not feasible with conventional techniques.



www.reactome.org

高通量

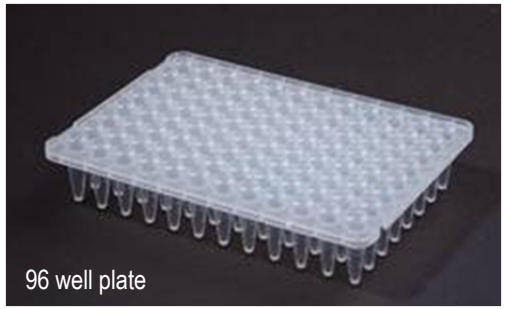
High-throughput screening (HTS), is a *method* for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry.



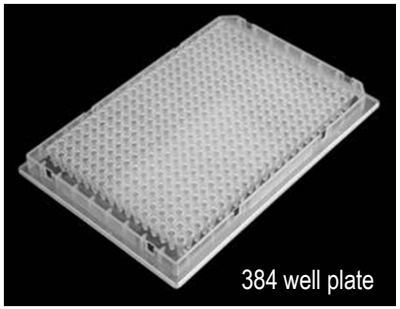






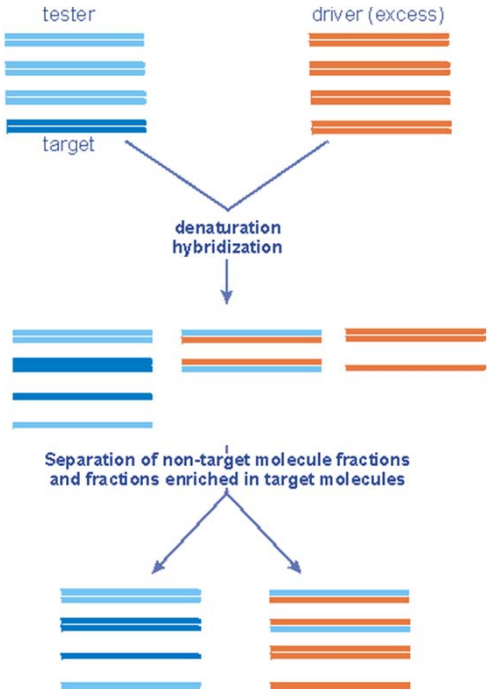


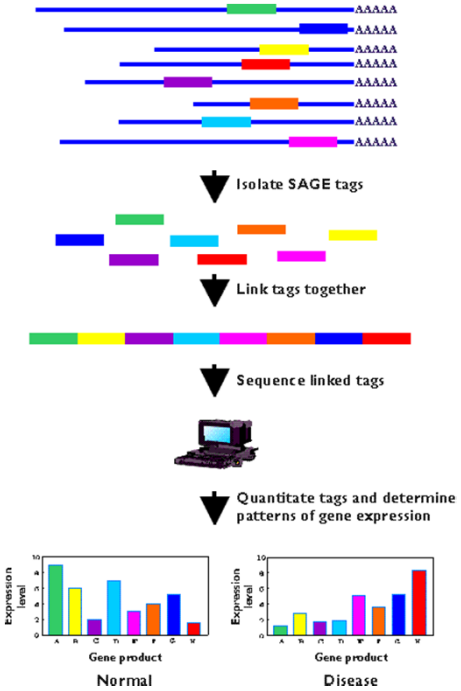


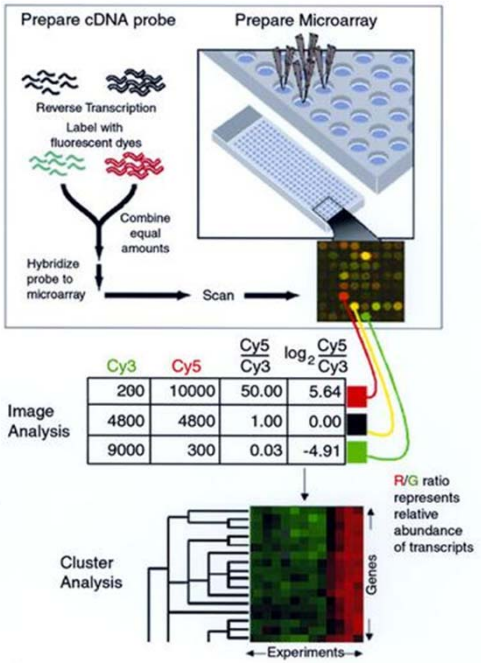


Evaluate global gene expression

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







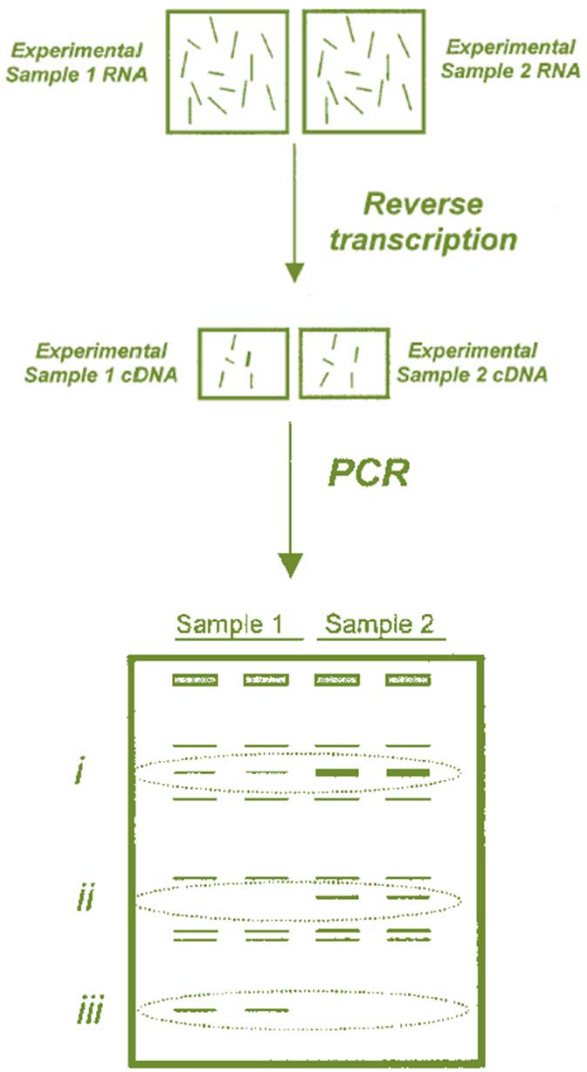
Differential displaying

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

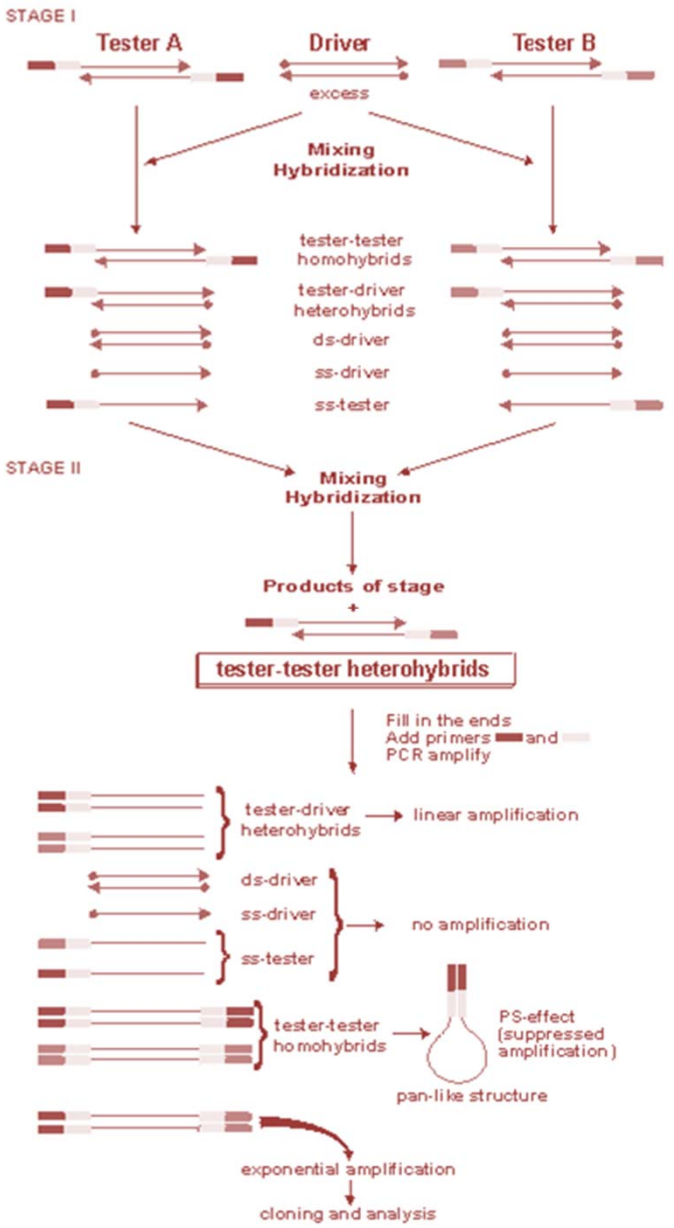


Subtractive hybridization

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

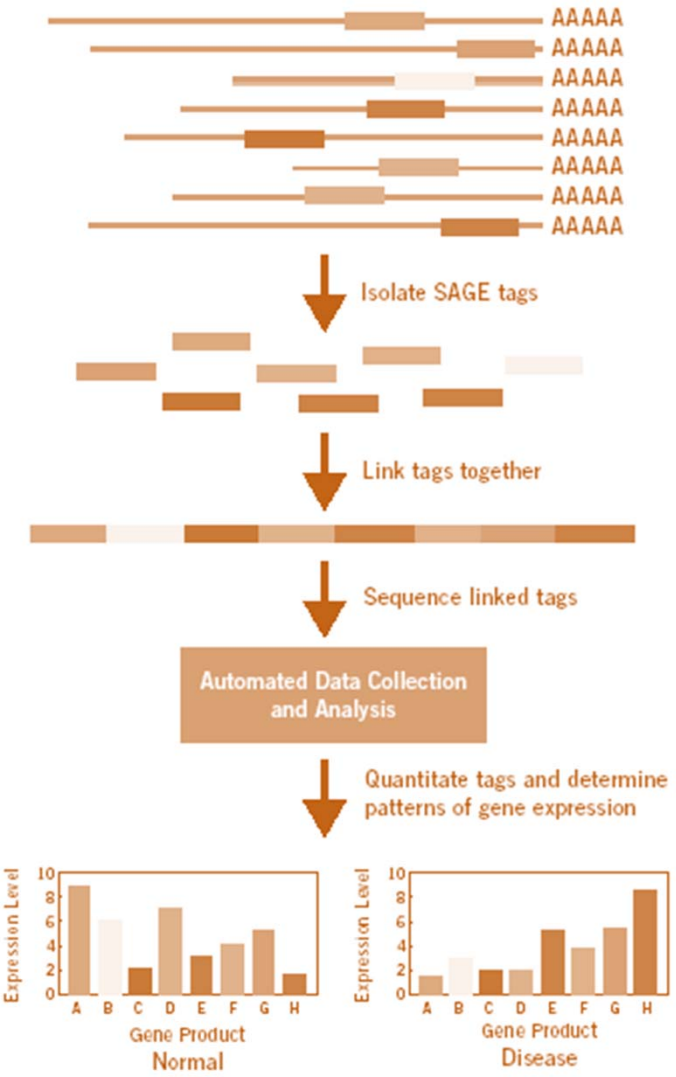


EST and SAGE

- ✓ 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.

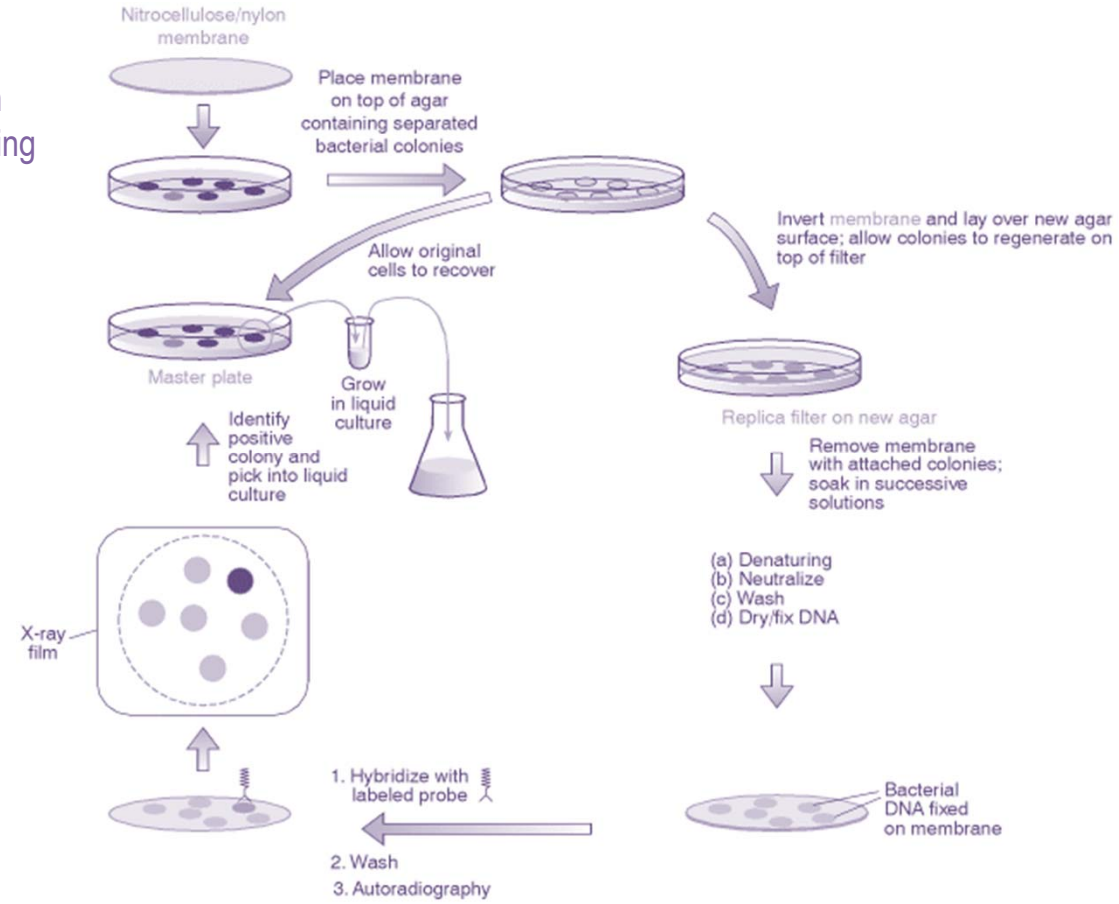


Microarray hybridization

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.



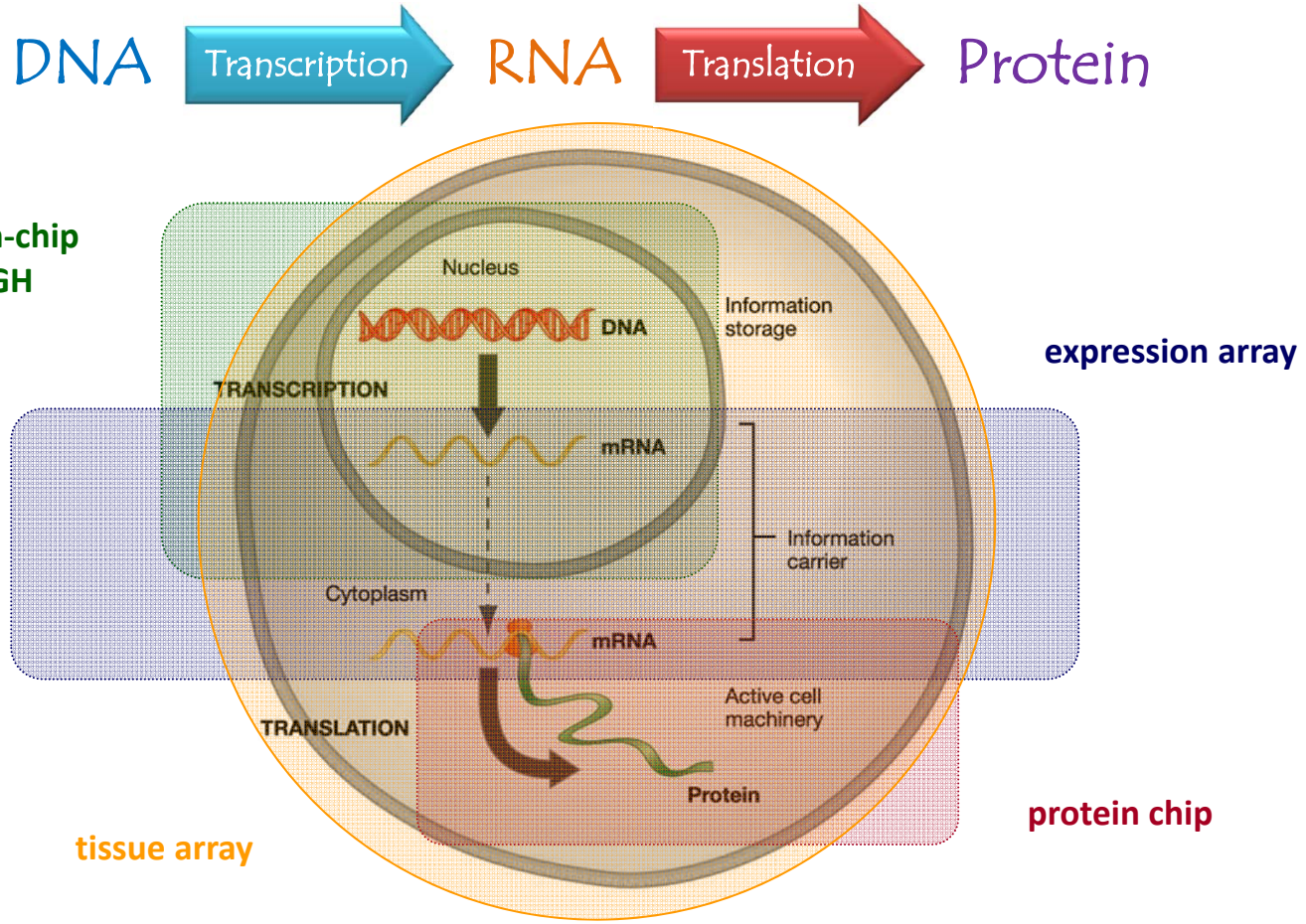
They are generally divided and differentiated by many ways

1. Type of target DNA (immobilized nucleic acid molecule) used in the array fabrication
2. Type of substrate to which the target DNA is printed or spotted
3. Methodology used to present the target DNA on the substrate
4. Density of probes on the array
5. Type of labeling for hybridization



1. **Type of target DNA (immobilized nucleic acid molecule) used in the array fabrication.**

cDNA, oligonucleotide, RNA, protein, antibody, tissue



CHIP-on-chip
array CGH

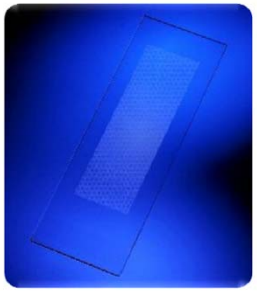
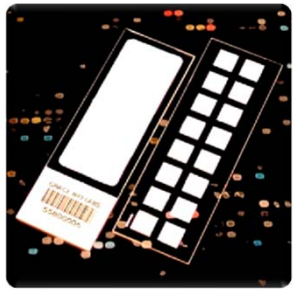
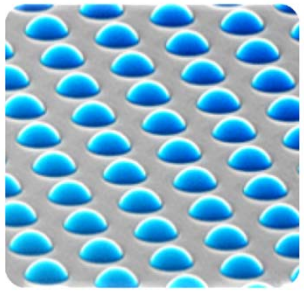
expression array

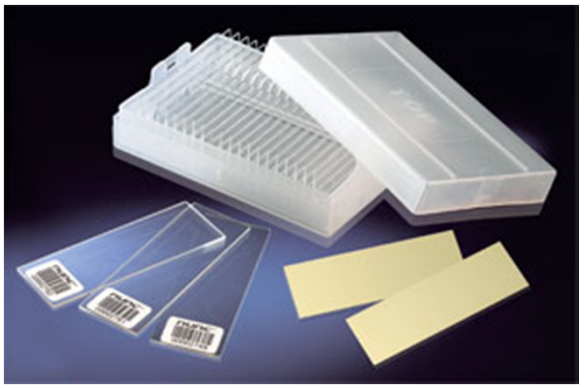
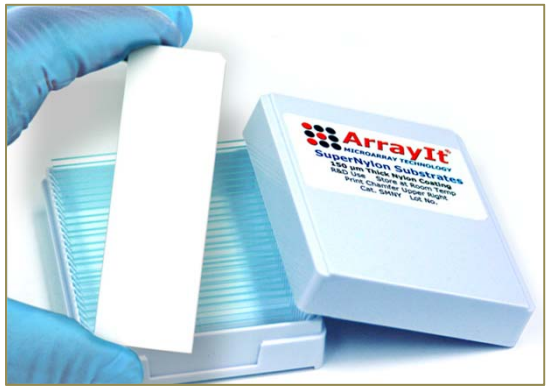
protein chip

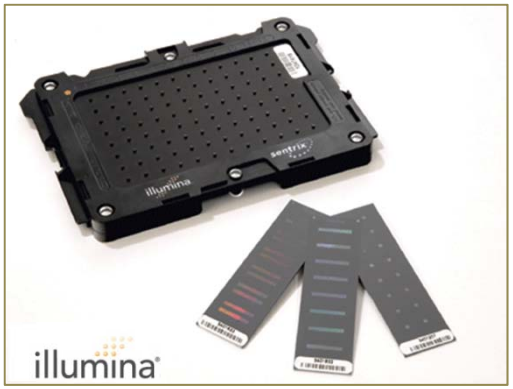
tissue array

2. Type of substrate to which the target DNA is printed or spotted.

glass, membrane, bead

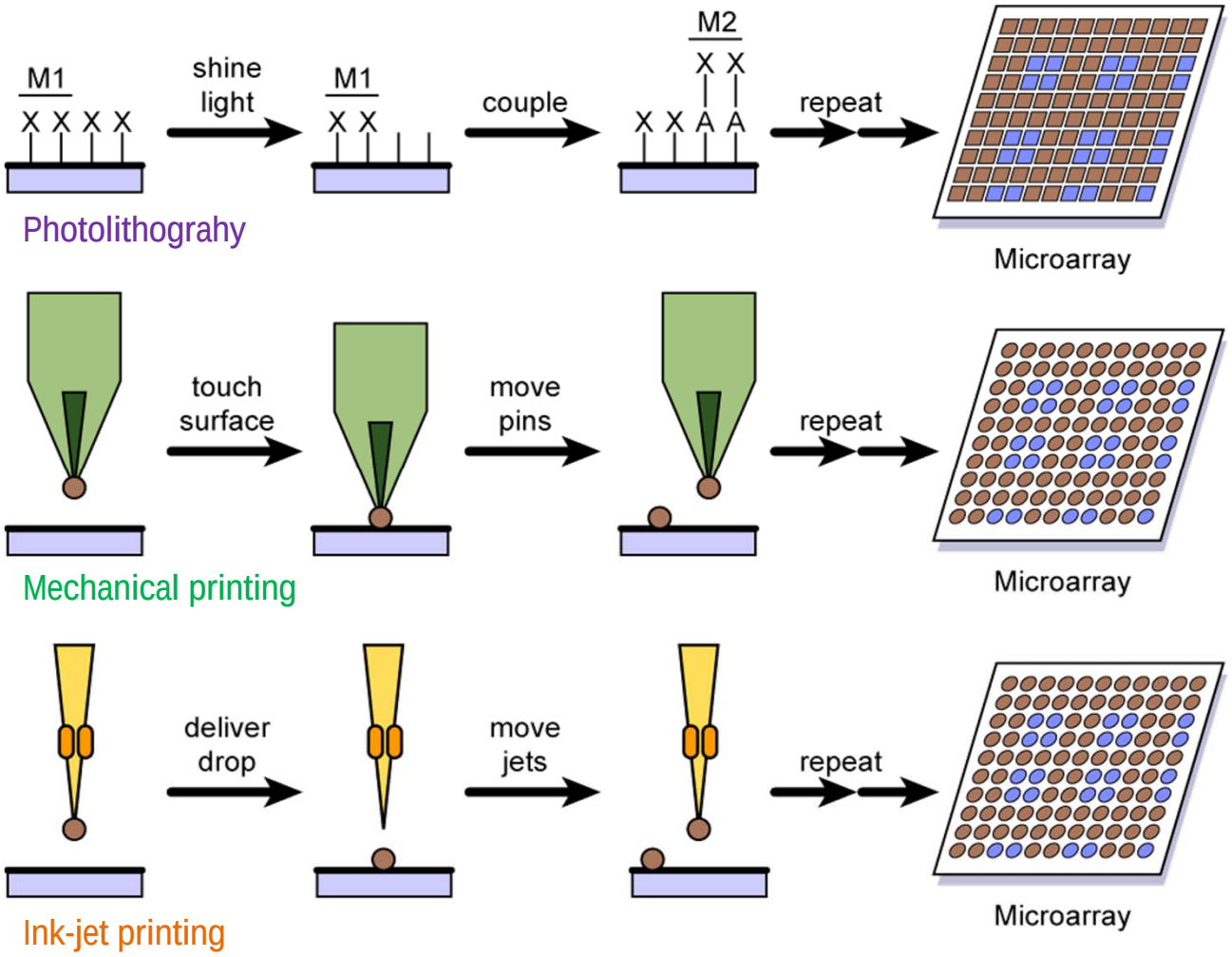





3. Methodology used to present the target DNA on the substrate.

printed directly example Affymterix

combine matrix spotted example Agilent

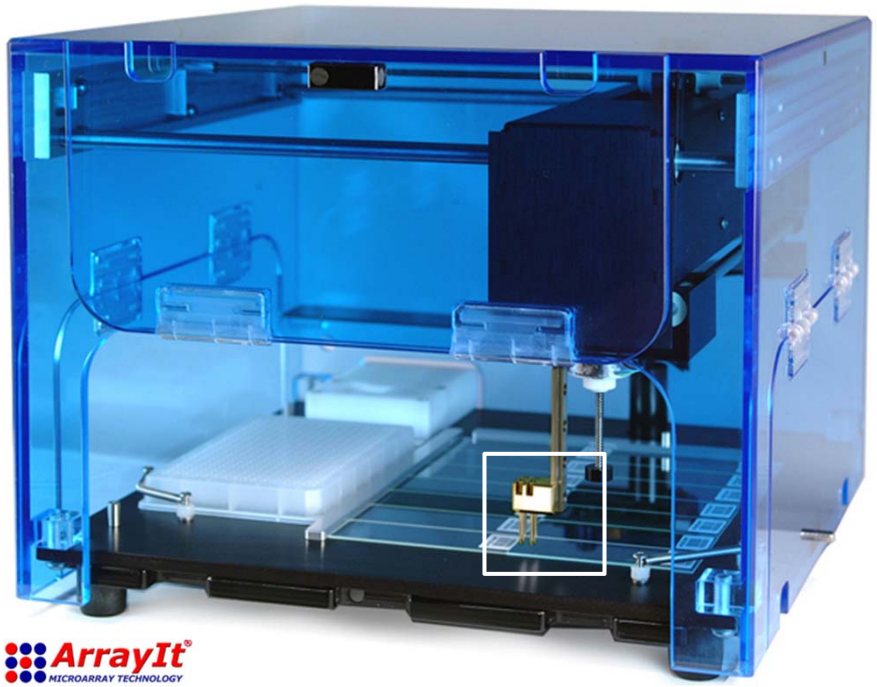


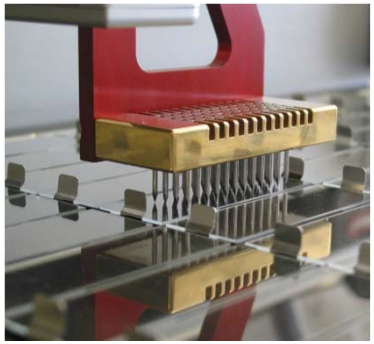
DNA spotting

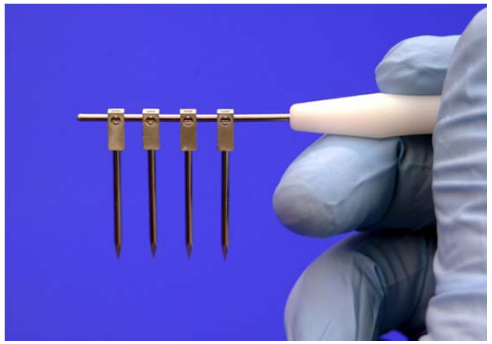
- DNA spotting usually uses multiple pins
- DNA in microtiter plate
- DNA usually PCR amplified
- Oligonucleotides can also be spotted

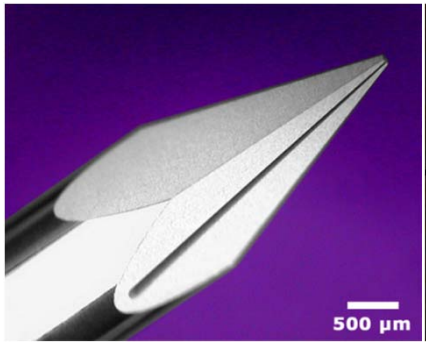


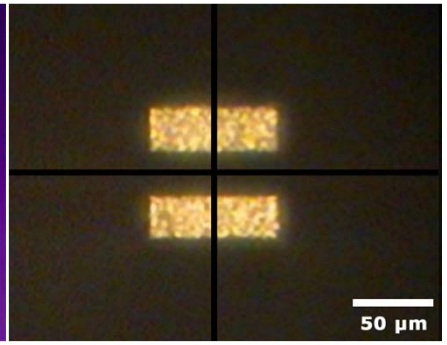






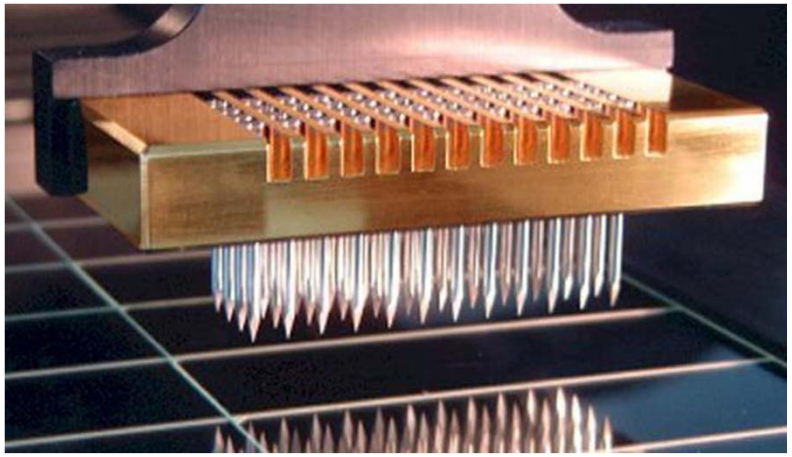






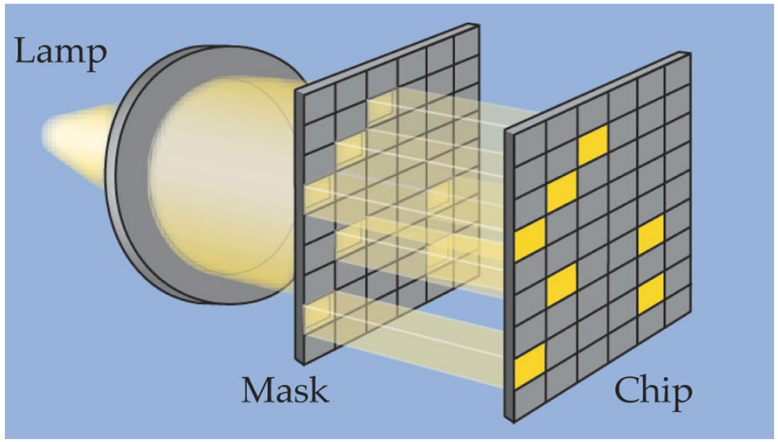
Ink-jet printer microarrays

- ✓ 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.



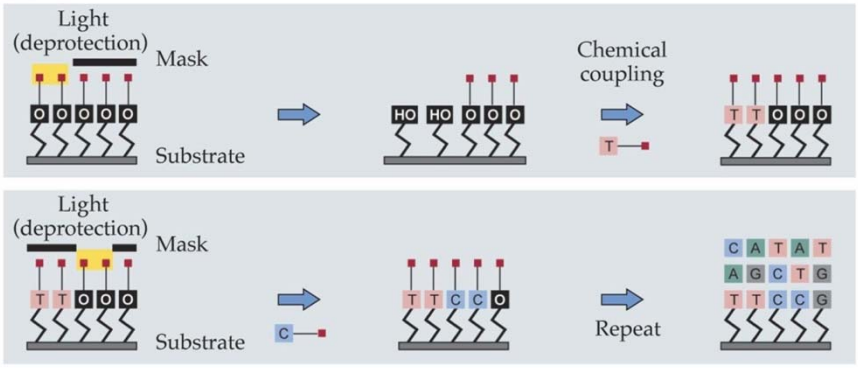
Photolithography

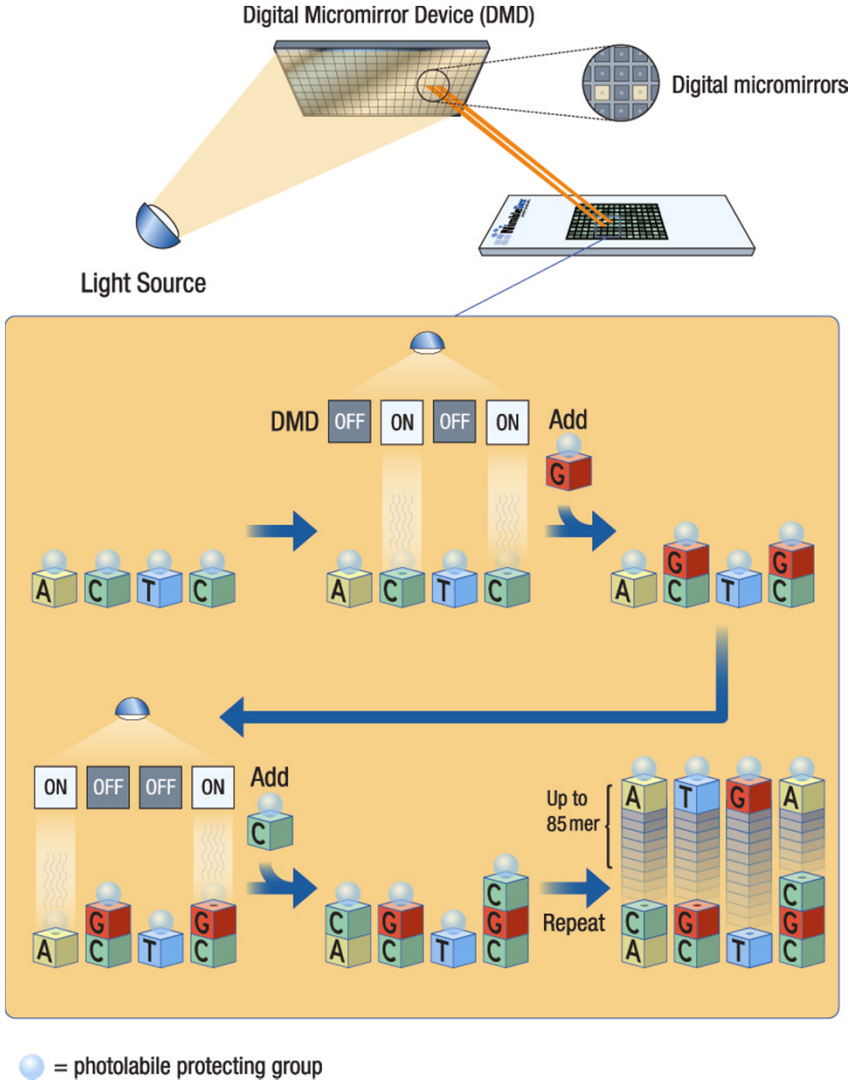
- ✓ Light-activated chemical reaction
  - For addition of bases to growing oligonucleotide
- ✓ Custom masks
  - Prevent light from reaching spots where bases not wanted
- ✓ Mirrors also used
  - NimbleGen™ uses this approach



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.

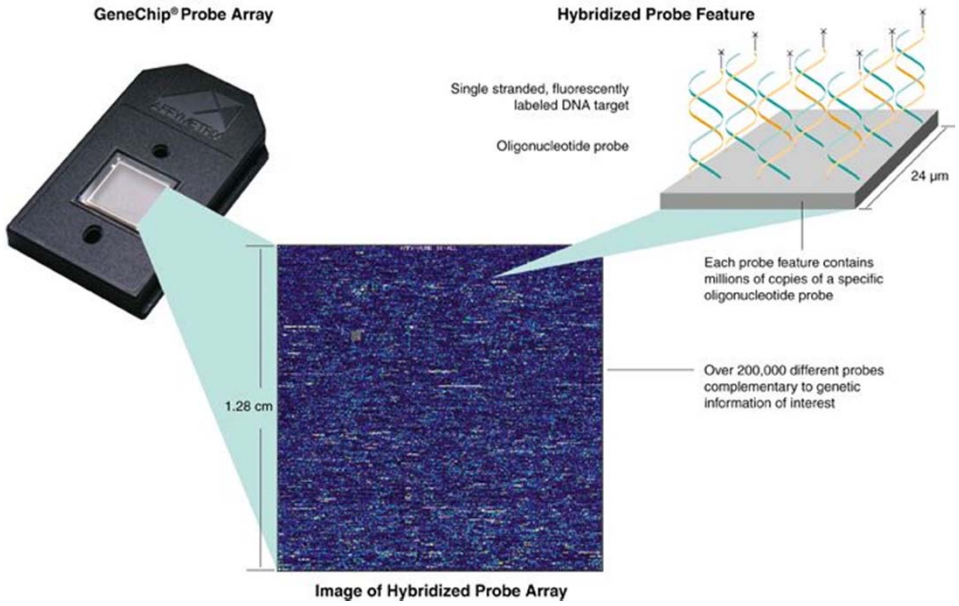




Oligonucleotide microarrays – pioneered by Affymetrix

- 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 ($4^{(20 \text{ to } 25)} = 2^{(40 \text{ to } 50)} >> 3*10^9 = 2^{30}$), 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. $T_m = 64.9+0.41*(GC\%) -675/L$)

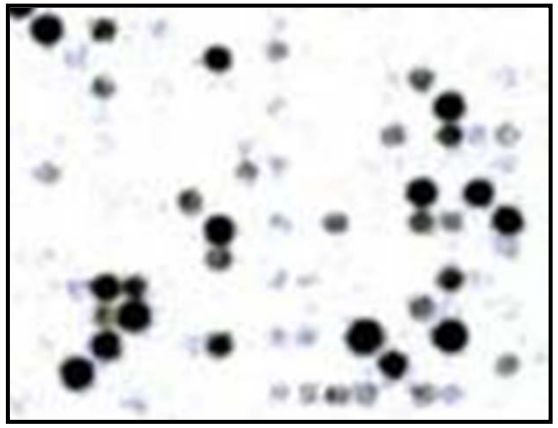


4. Density of probes on the array.

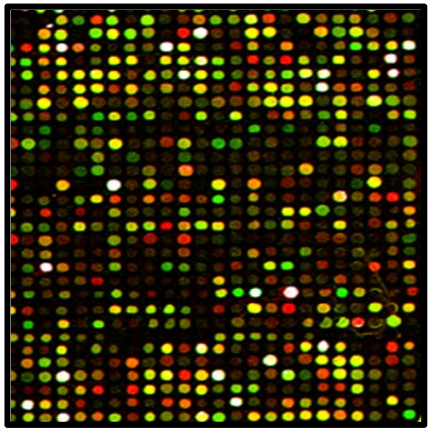
low density: 100 s (example: protein antibodies)

medium density: 1,000s to 10s of 1,000s and more (example: cDNA)

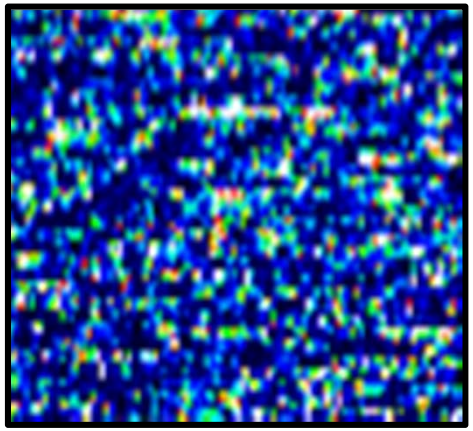
high density 100s to 1000s of 1000s (example: short oligonucloetide)



Isotope
Nylon – cDNA (300-900 nt)

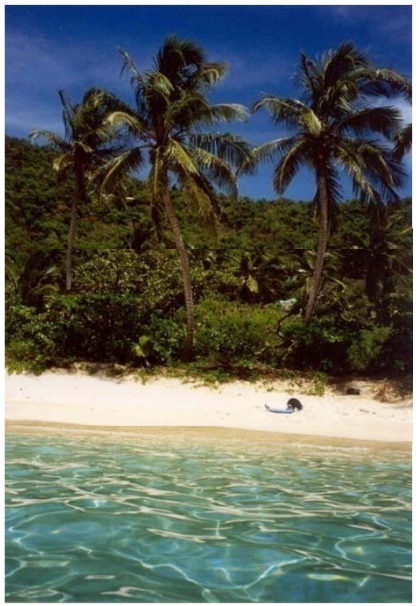


Two-colour
cDNA or Oligo (80 nt)
500 – 11,000 elements



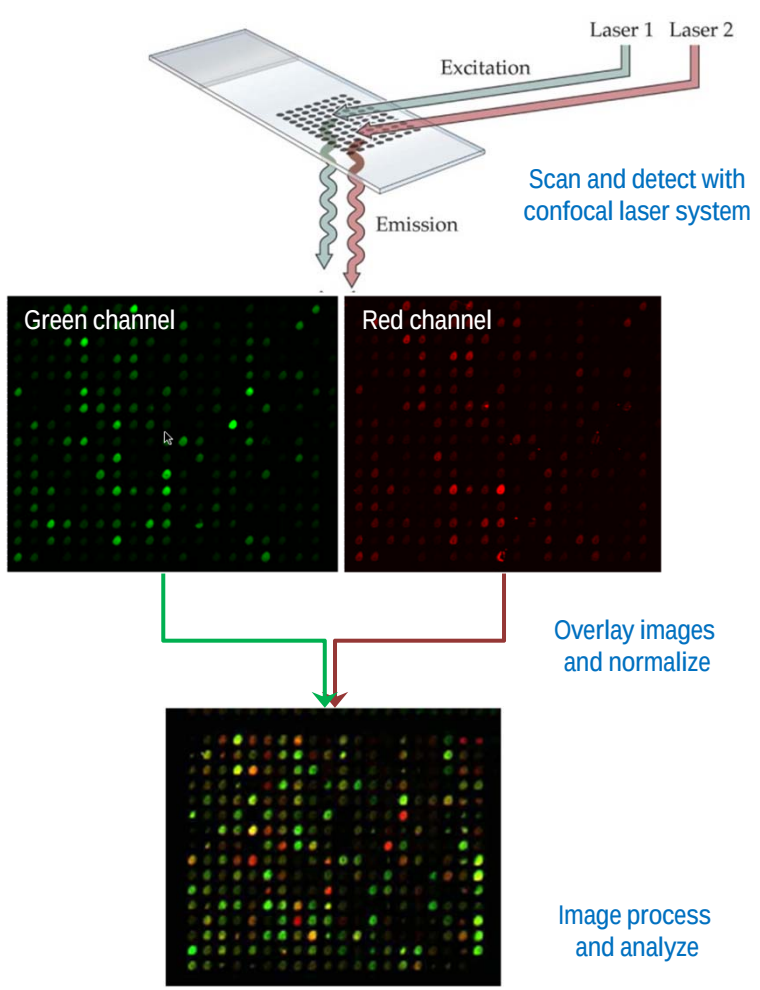
Affymetrix
Silicone – oligo (20 nt)
22 ,000 elements

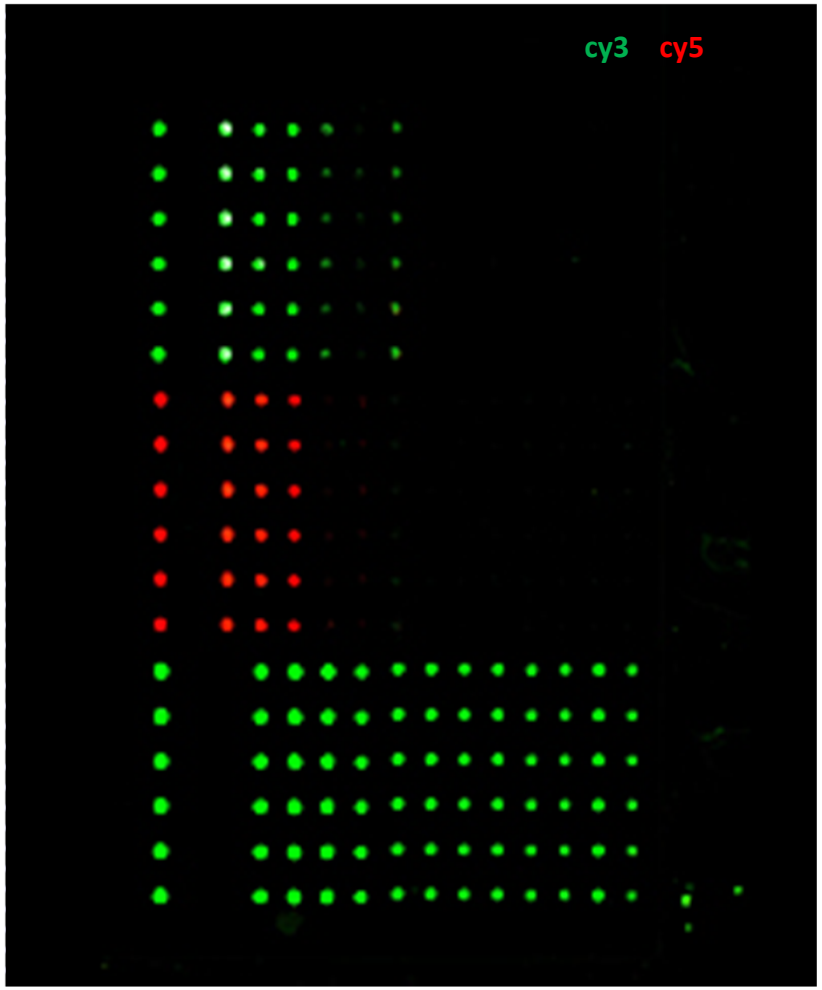
Digital Image Analogy

	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

5. Type of labeling for hybridization.

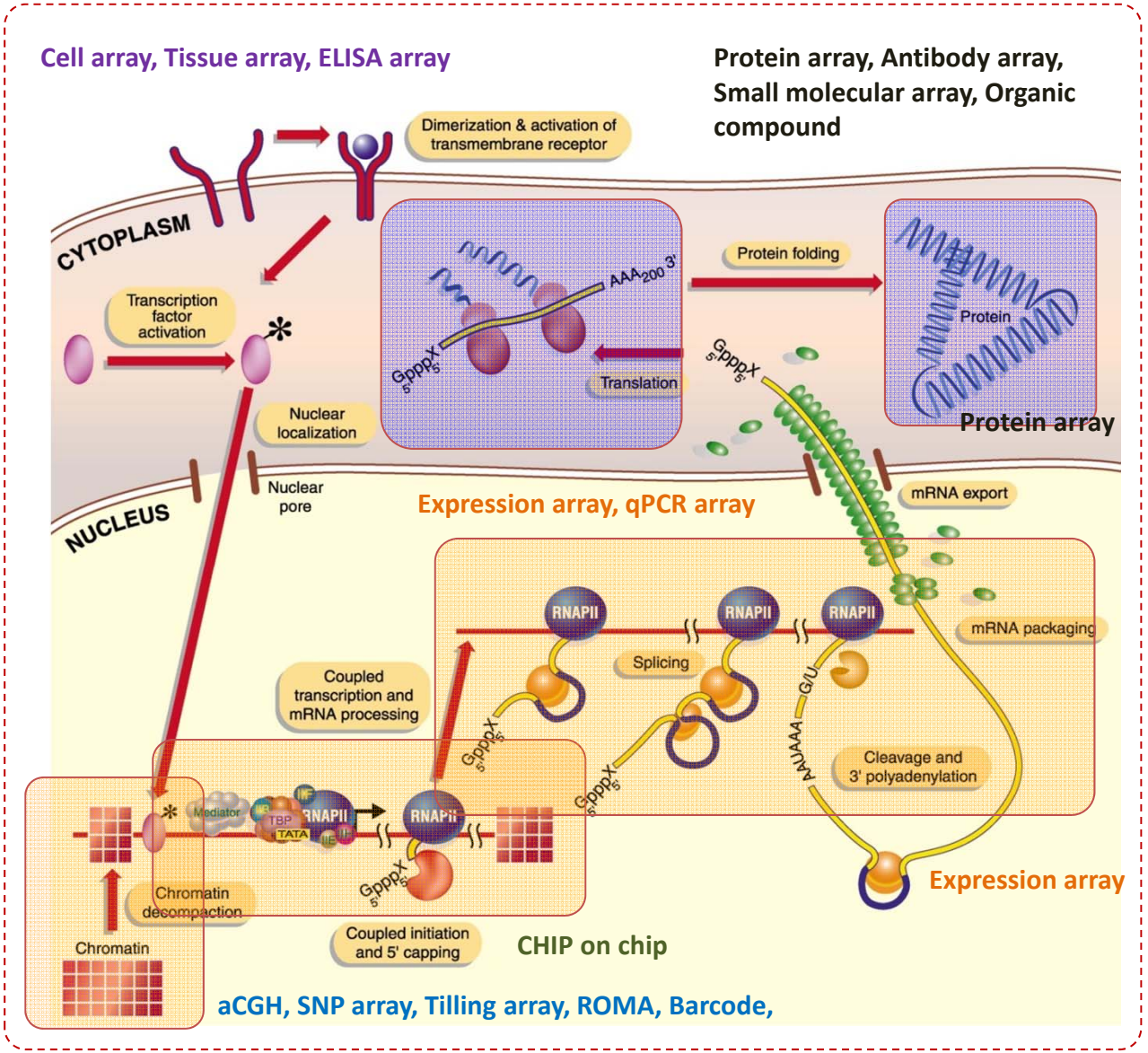
single channel using one color and double channel using two color





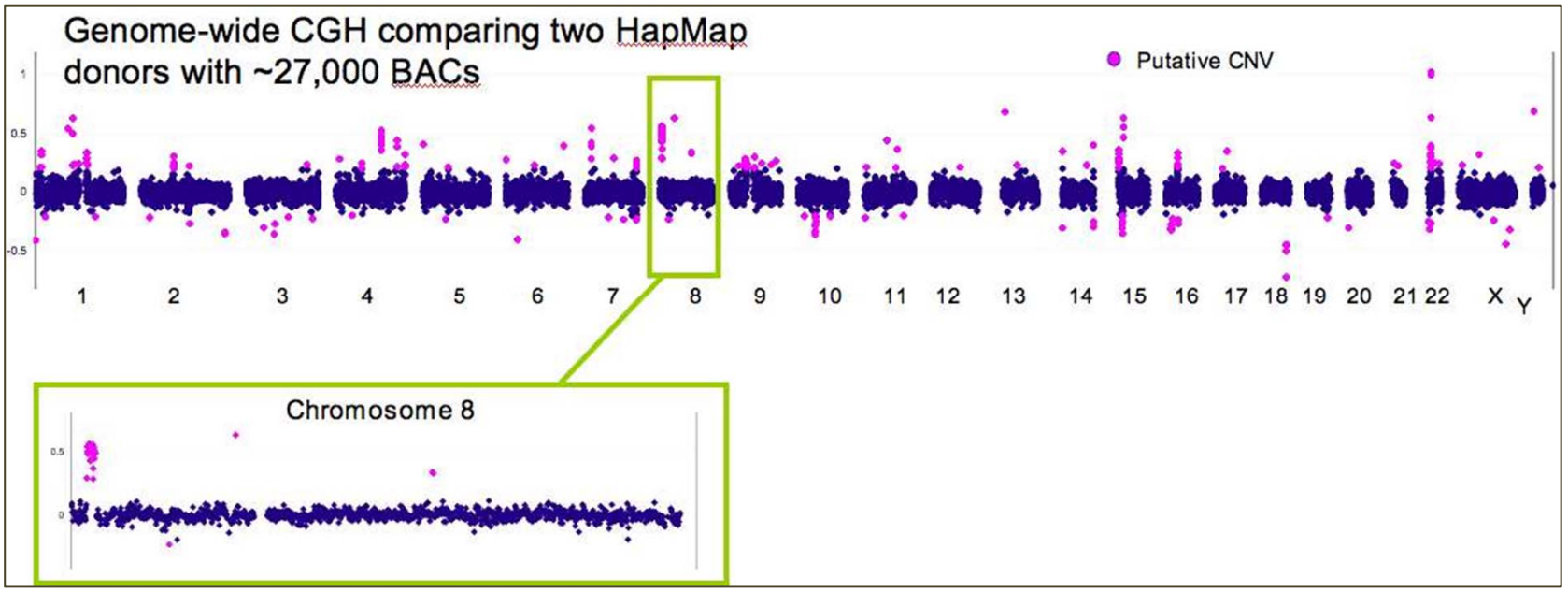
Central Dogma and Existing Microarrays Types

1. aCGH
2. SNP array
3. Tilling array
4. ROMA
5. Barcode array
6. cDNA array
7. qPCR array
8. RNA array
9. Protein array
10. CHIP on chip
11. Antibody array
12. Organic compound
13. Small molecule
14. ELISA array
15. Cell array
16. Tissue array
17. Lab-on-a-chip



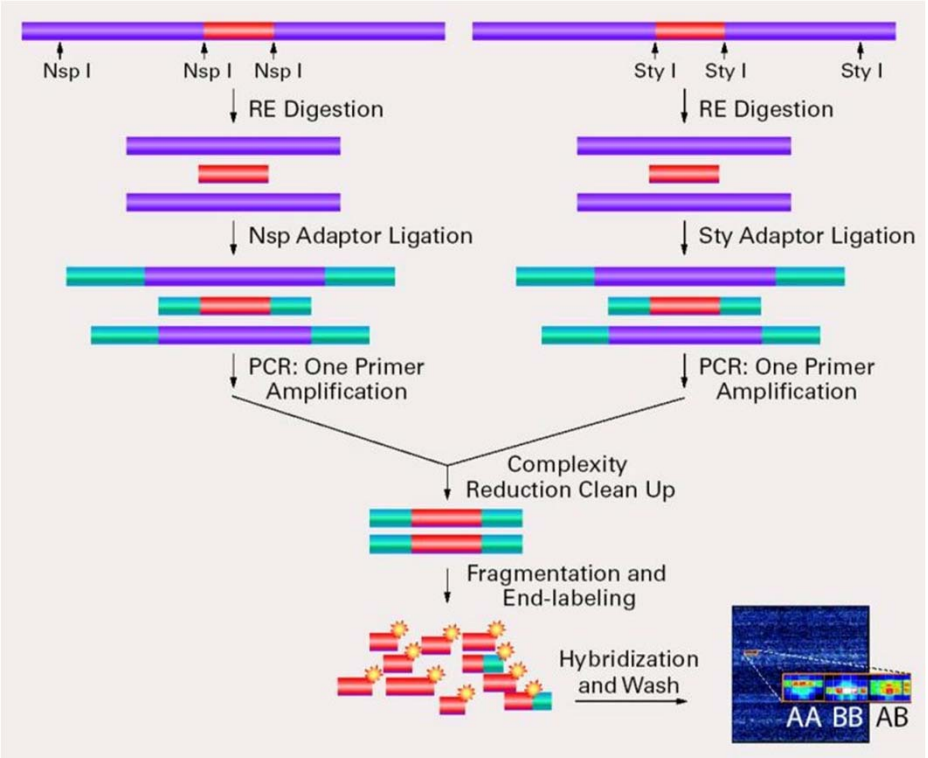
1. **aCGH**
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

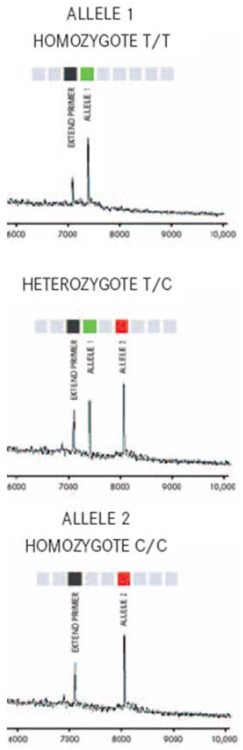
Array comparative genomic hybridization (also CMA, Chromosomal Microarray Analysis, Microarray-based comparative genomic hybridization, array CGH, a-CGH, aCGH, or virtual karyotype) genomic copy number variations at a higher resolution level than chromosome-based comparative genomic hybridization (CGH). This is a molecular-cytogenetic method for the analysis of copy number changes (gains/losses) in the DNA content of a given subject's DNA and often in tumor cells.



1. aCGH
2. **SNP array**
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

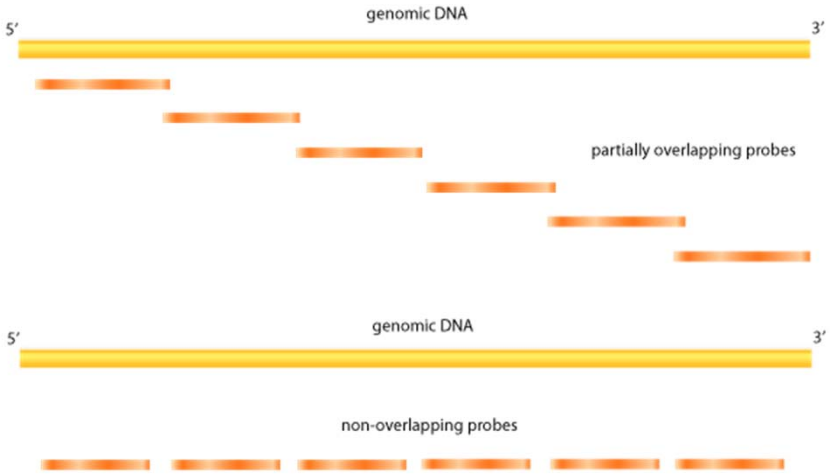
A **single nucleotide polymorphism** (**SNP**, pronounced snip) is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In molecular biology and bioinformatics, a SNP array is a type of DNA microarray which is used to **detect polymorphisms within a population**.

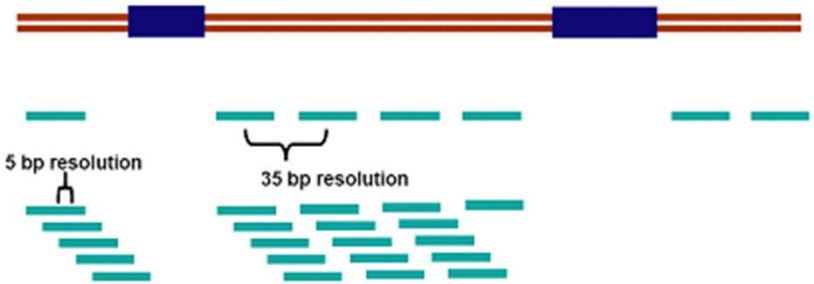




1. aCGH
2. SNP array
3. **Tilling array**
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

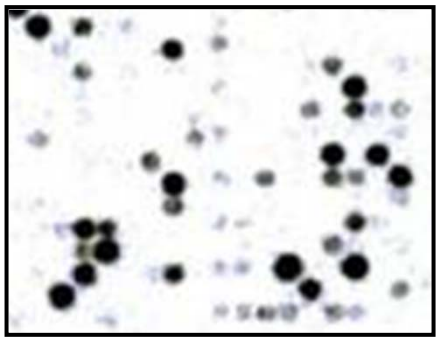
Tiling arrays differ in the nature of the probes. Short fragments are designed to **cover the entire genome or contigs of the genome**. Depending on the probe lengths and spacing different degrees of resolution can be achieved. Number of features on a single array can range from 10,000 to greater than 6,000,000, with each feature containing millions of copies of one probe. Tiling arrays can produce an unbiased look at gene expression because previously unidentified genes can still be incorporated. On top of individual gene expression analysis, other tiling arrays can be used in transcriptome mapping, ChIP-chip, MeDIP-chip and DNase Chip studies, Array CGH among others.



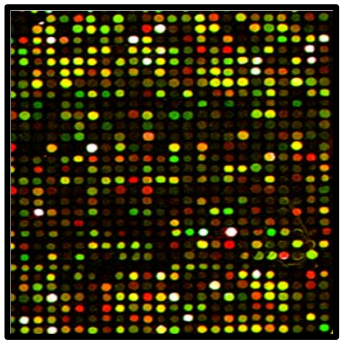


1. aCGH
2. SNP array
3. Tilling array
4. **cDNA array**
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

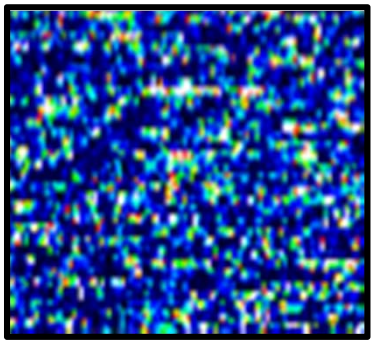
A **DNA microarray** (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface by covalent attachment to a chemical matrix. DNA arrays are different from other types of microarray only 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

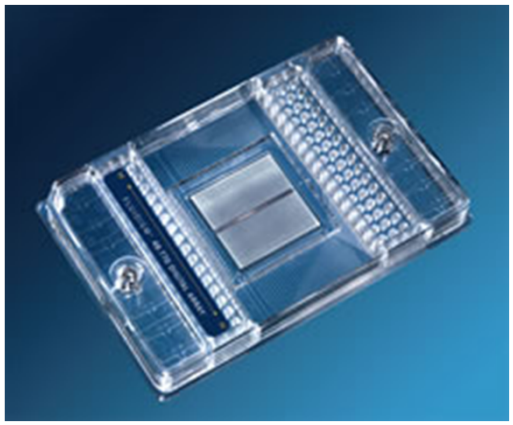


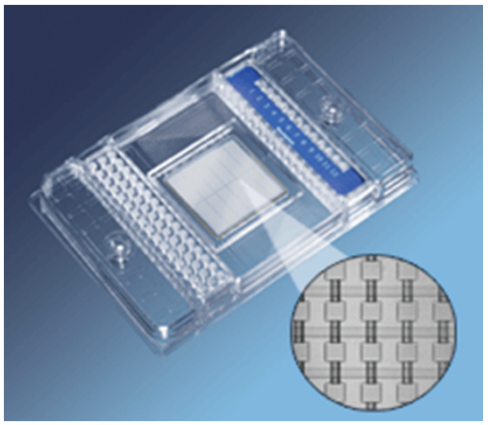
Affymetrix
Silicone – oligo (20 nt)
22 ,000 elements

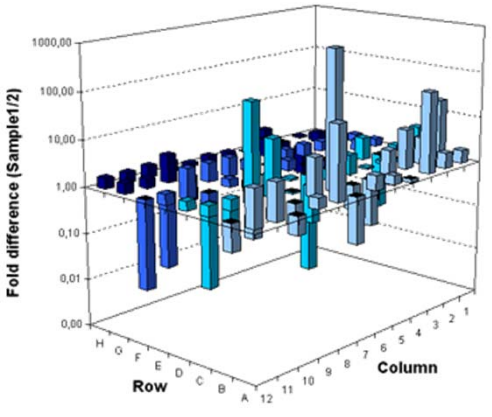
1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. **qPCR array**
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

Tiling arrays

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.

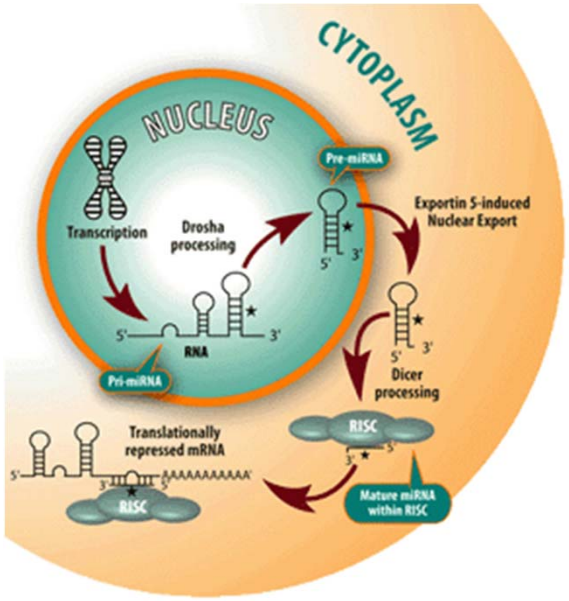


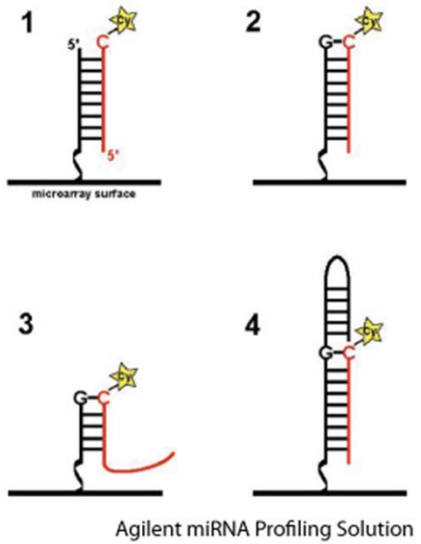




1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. **RNA array**
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

In genetics**, microRNAs** (miRNA) are single-stranded RNA molecules of 21-23 nucleotides in length, which regulate gene expression. miRNAs are encoded by genes from whose DNA they are transcribed but miRNAs are not translated into protein (non-coding RNA); instead each primary transcript (a pri-miRNA) is processed into a short stem-loop structure called a pre-miRNA and finally into a functional miRNA. Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules, and their main function is to down-regulate gene expression. MicroRNA array is a kind of arrays which is designed for the detection of **miRNA expression profile**.

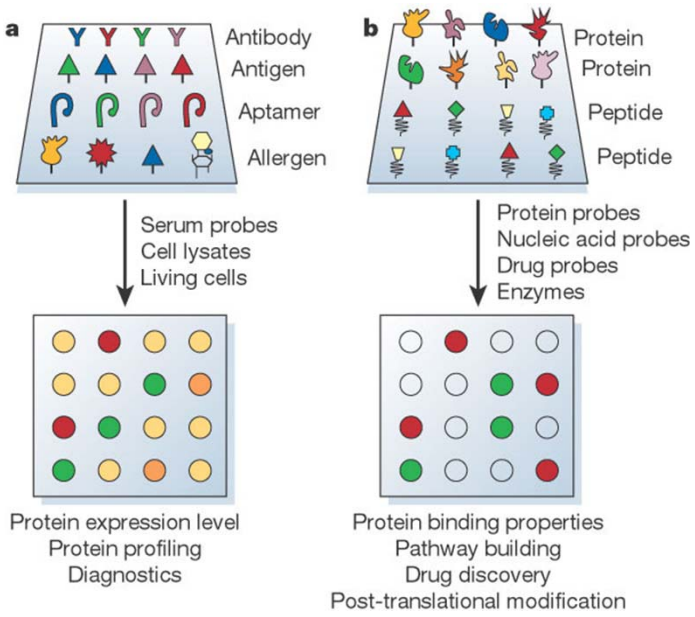




Agilent miRNA Profiling Solution

1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. **Protein array**
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

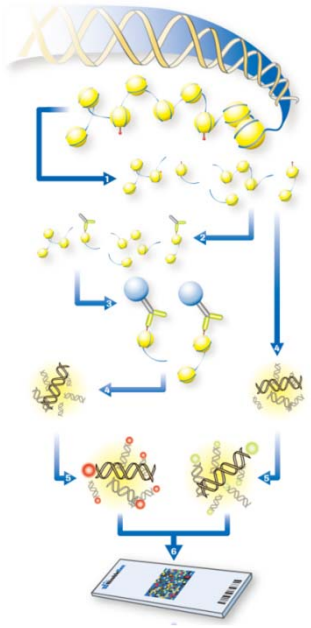
A **protein microarray** is a piece of glass on which different molecules of protein have been affixed at separate locations **in an ordered manner** thus forming a microscopic array. These are used to identify protein-protein interactions, to identify the substrates of protein kinases, or to identify 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.

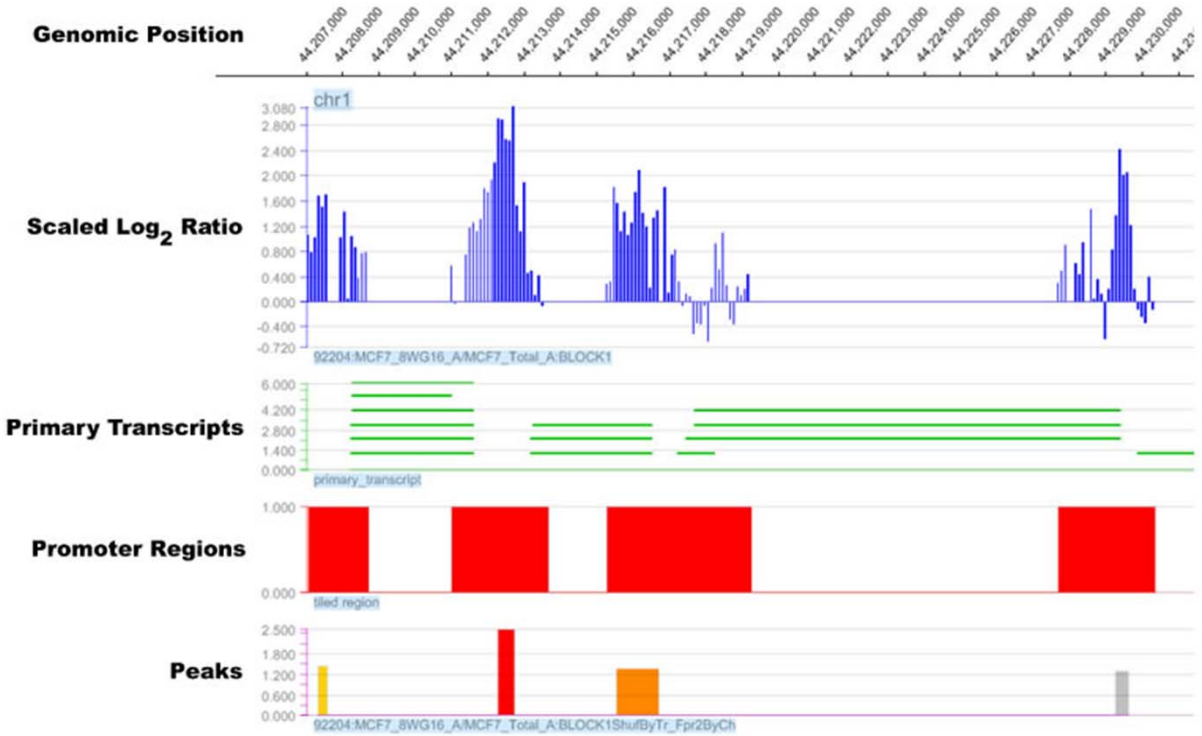




1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. **CHIP on chip**
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

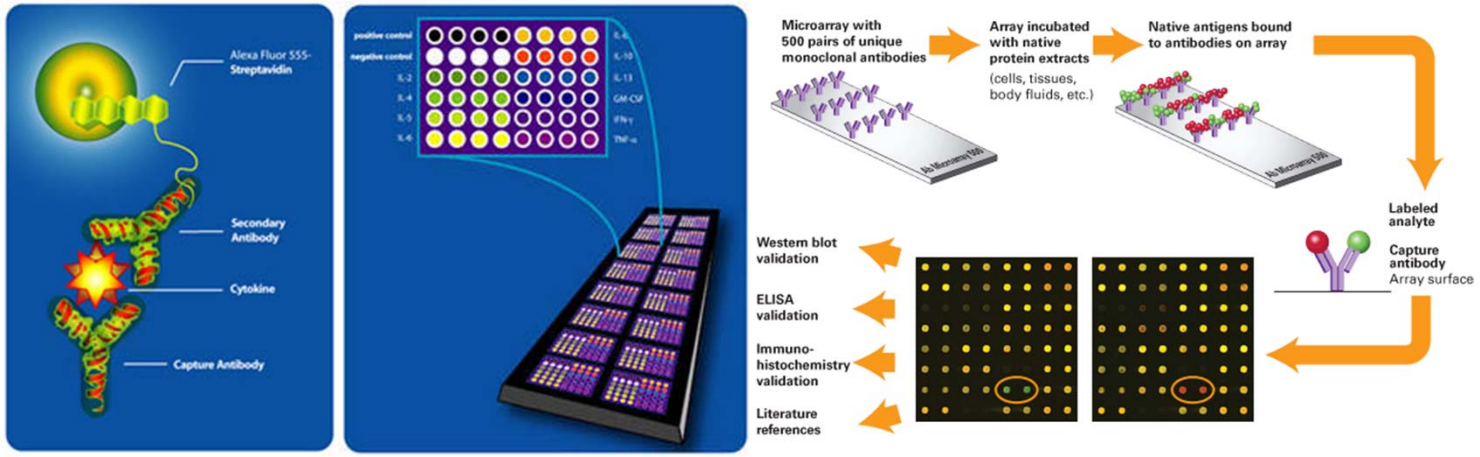
ChIP-on-chip (also known as ChIP-chip) is a technique that combines chromatin immunoprecipitation ("ChIP") with microarray technology ("chip"). Like regular ChIP, ChIP-on-chip is used to investigate **interactions between proteins and DNA** *in vivo*. Specifically, it allows the identification of binding sites of DNA-binding proteins on a genome-wide basis. One of the long-term goals ChIP-on-chip was designed for is to establish a catalogue of (selected) organisms that lists all protein-DNA interactions under various physiological conditions.





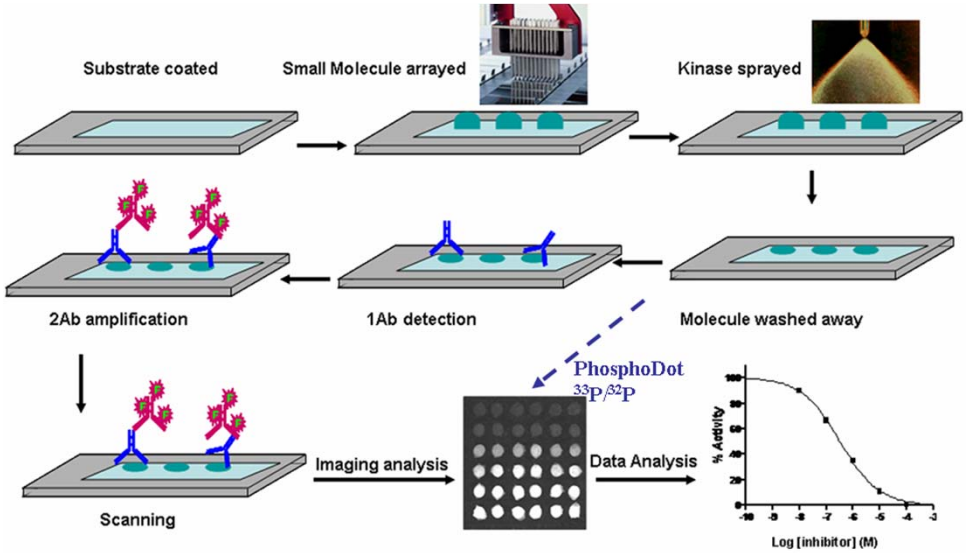
1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. **aAntibody**
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

An **antibody microarray** is a specific form of protein microarrays, a collection of capture antibodies are spotted and fixed on a solid surface, such as glass, plastic and silicon chip for the purpose of detecting antigens. Antibody microarray is often used for **detecting protein expressions** from cell lysates in general research and special biomarkers from serum or urine for diagnostic applications.



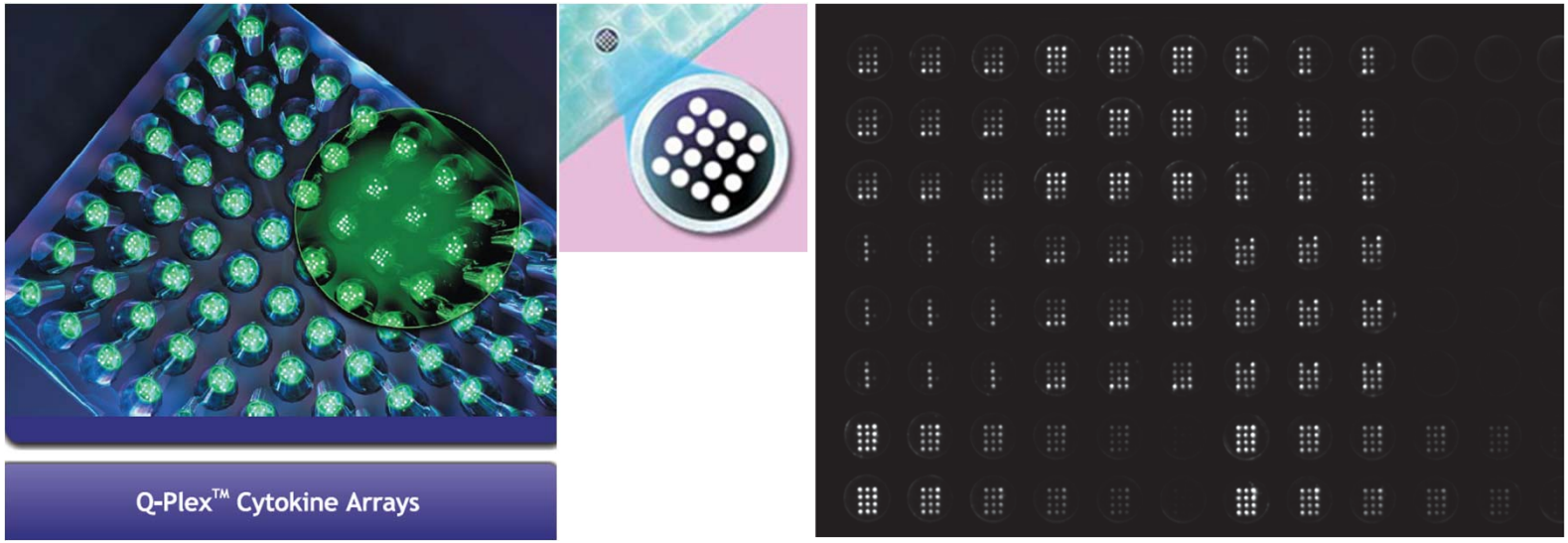
1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. **aOrganic cpd**
11. ELISA array
12. Cell array
13. Tissue array
14. Lab-on-a-chip

A **chemical compound microarray** is a collection of organic chemical compounds spotted on a solid surface, such as glass and plastic. In chemical genetics research, they are routinely used for searching proteins that bind with specific chemical compounds, and in general drug discovery research, they provide a multiplex way to search potential drugs for therapeutic targets. There are three different forms of chemical compound microarrays based on the fabrication method, **to covalently immobilize** (usually called Small Molecule Microarray), **to spot and dry** organic compounds on the solid surface without immobilization (commercial name as Micro Arrayed Compound Screening (μARCS)), and to spot organic compounds **in a homogenous solution** without immobilization and drying effect.



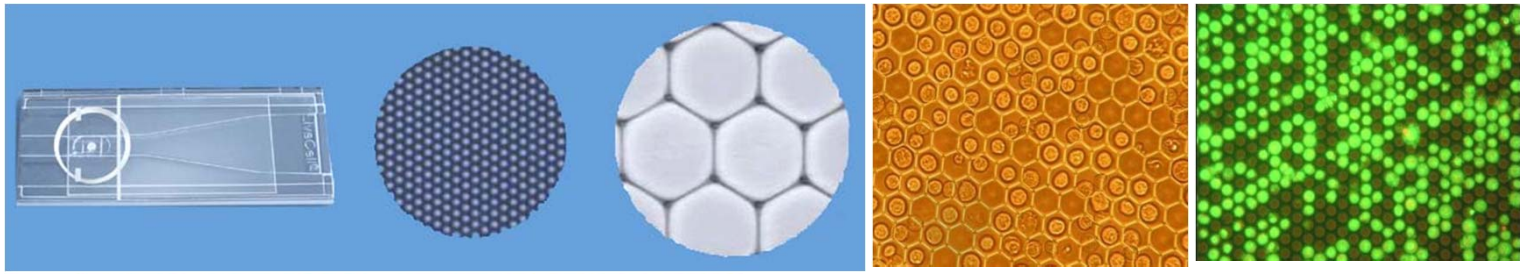
1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. **ELISA array**
12. Cell array
13. Tissue array
14. Lab-on-a-chip

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 (or array ELISA) 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.

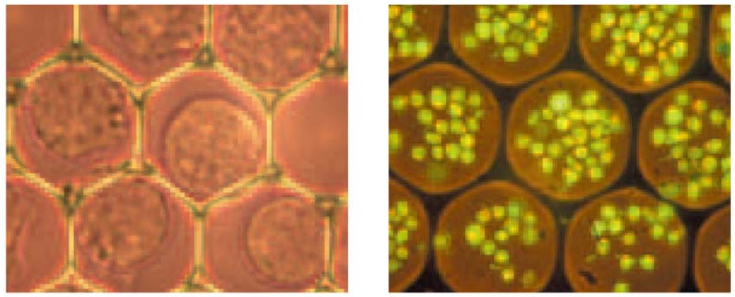


1. aCGH
2. SNP array
3. Tilling array
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. **Cell array**
13. Tissue array
14. Lab-on-a-chip

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



LIVECELL ARRAY microscope slide



The Individual Cell Array with 15 (left) and 100-micron.

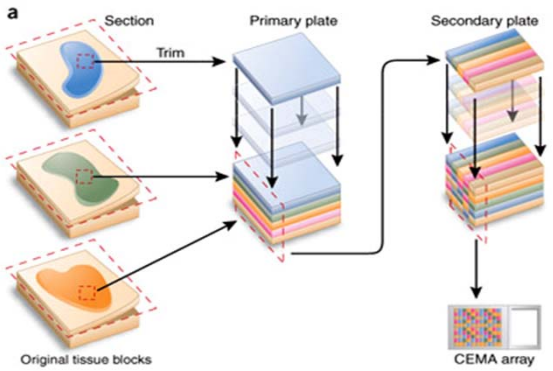
1. aCGH
2. SNP array
3. **Tilling array**
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array

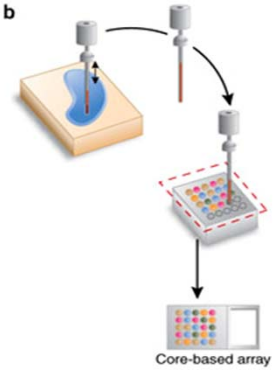
13. Tissue array

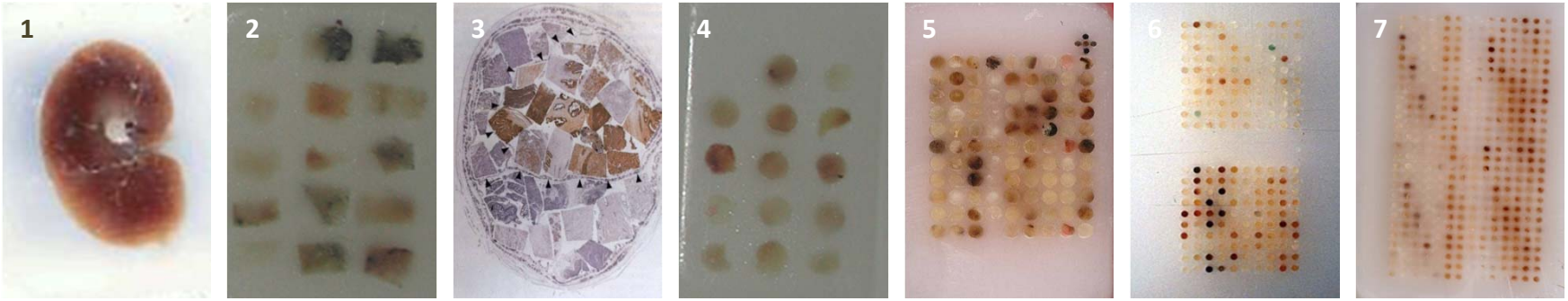
14. Lab-on-a-chip

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.





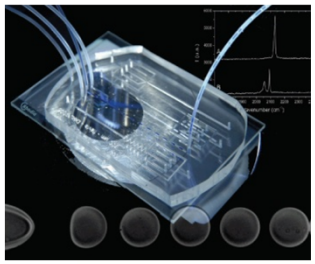


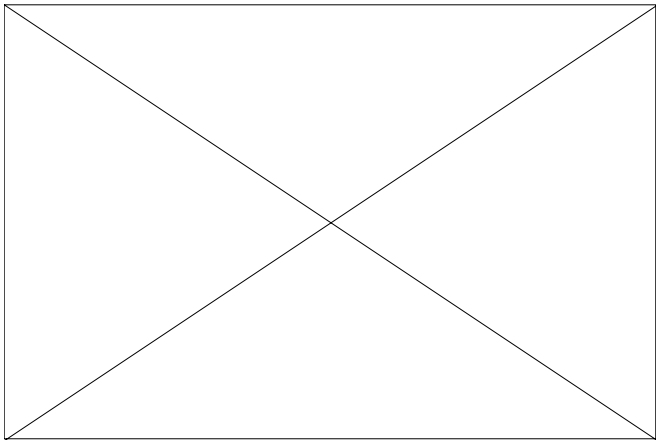


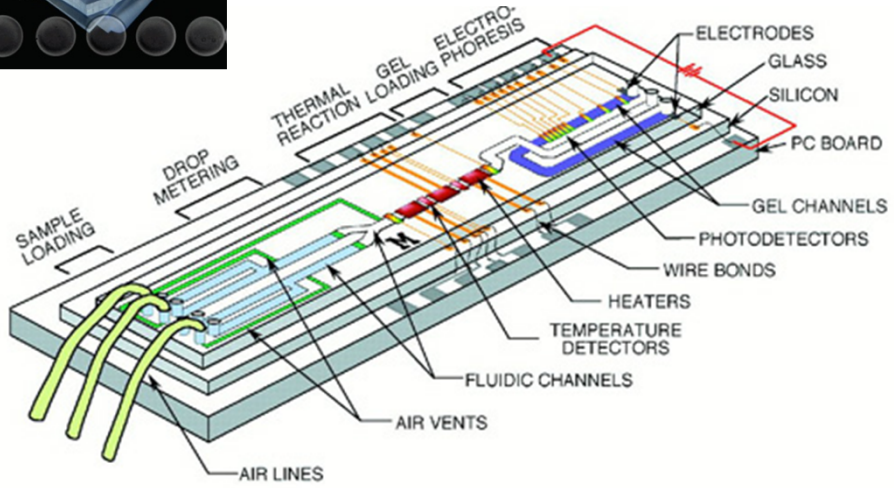
whole tissue | **tissue cocktail** | **tissue sausage** | **tissue array**

1. aCGH
2. SNP array
3. **Tilling array**
4. cDNA array
5. qPCR array
6. RNA array
7. Protein array
8. CHIP on chip
9. Antibody array
10. Organic compound
11. ELISA array
12. Cell array
13. Tissue array
14. **Lab-on-a-chip**

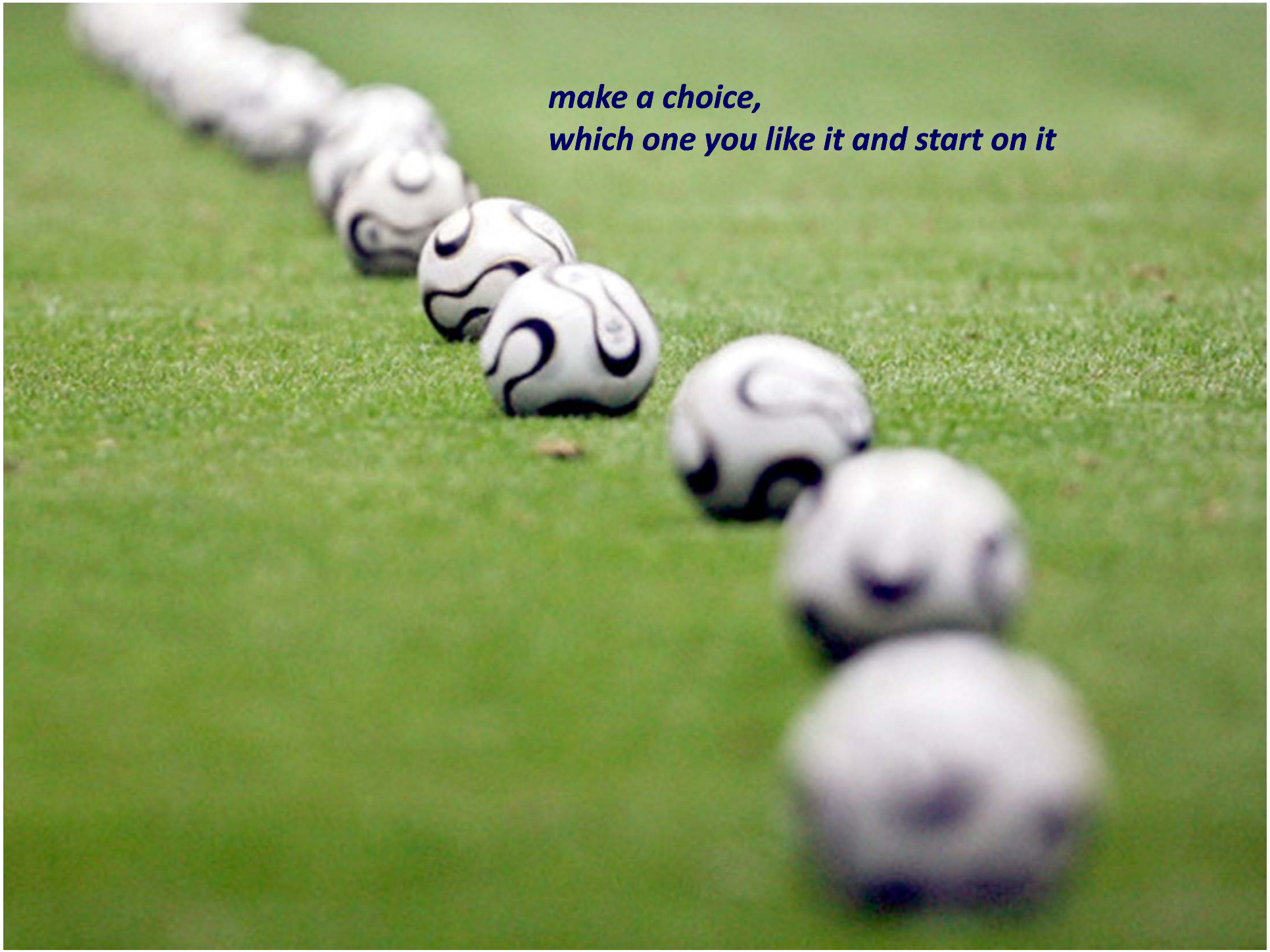
Lab-on-a-chip (LOC) is a term for devices that **integrate (multiple) laboratory functions on a single chip** of only millimeters to a few square centimeters in size and that are capable of handling extremely small fluid volumes down to less than pico liters.





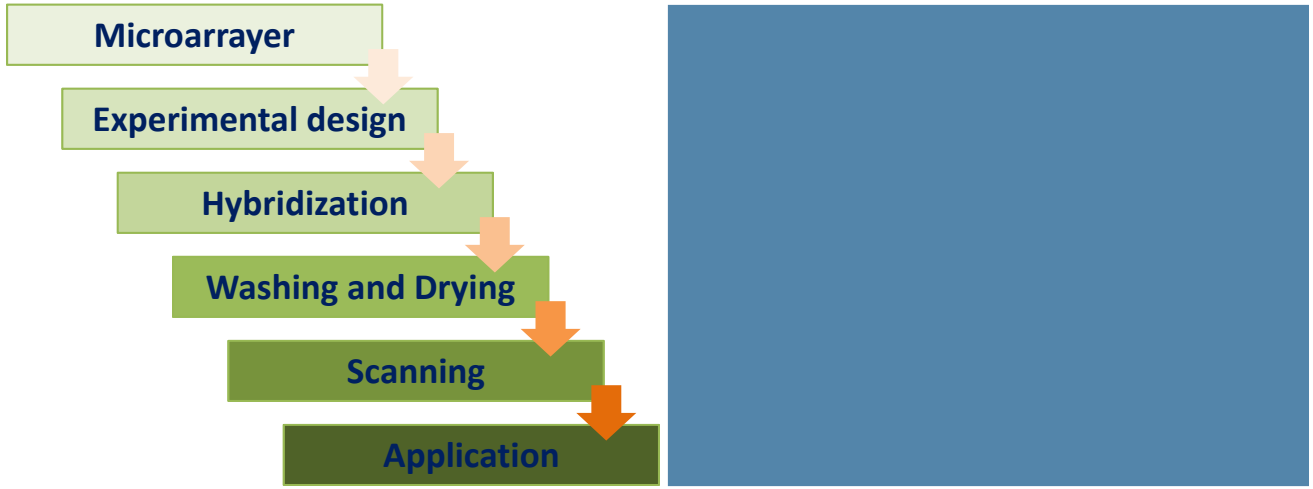





make a choice,
which one you like it and start on it

The 6 steps of a microarray experiment

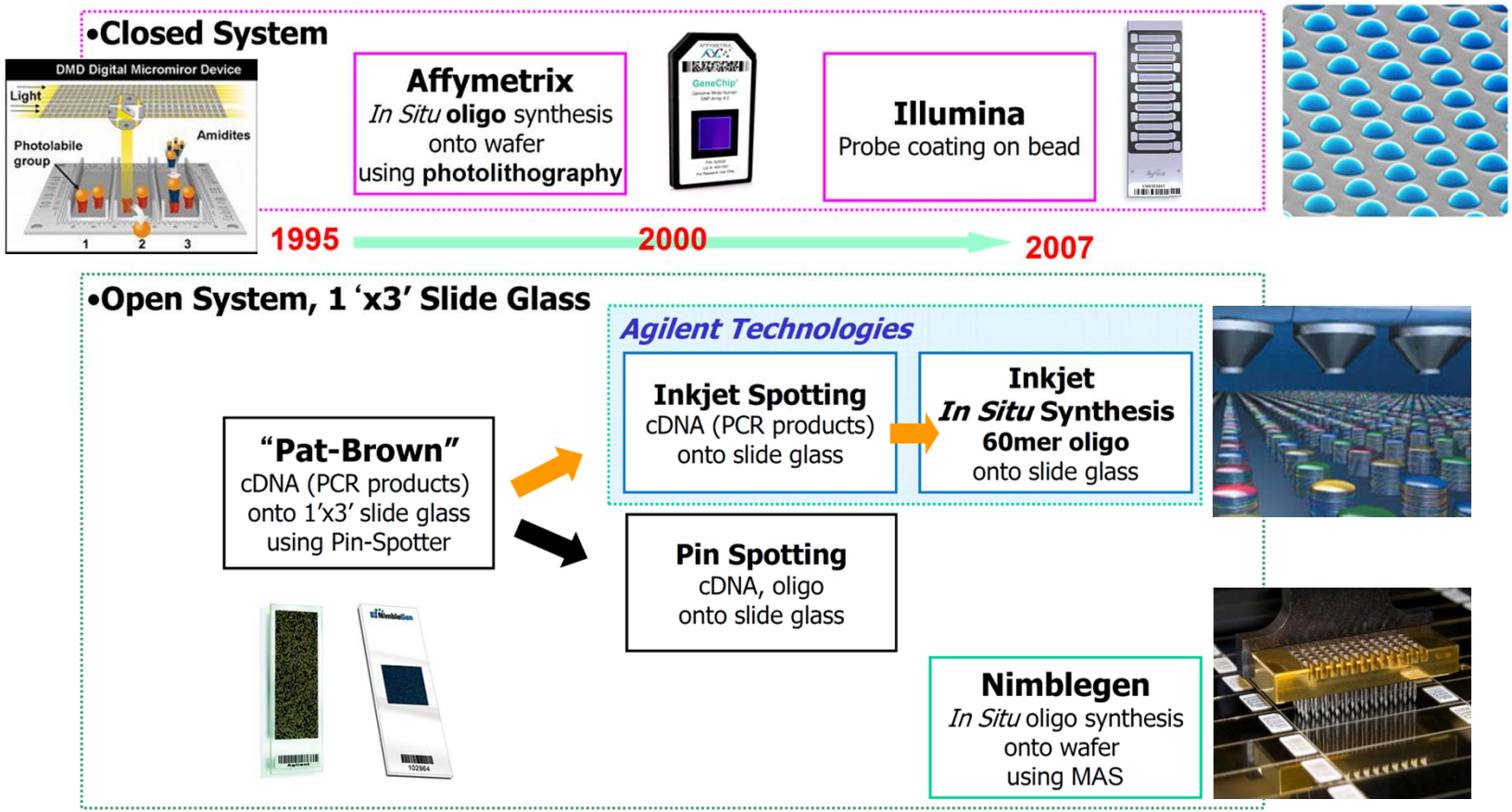
1. Manufacturing of the microarray
2. Experimental design and choice of reference: what to compare to what?
3. Target preparation (labeling) and hybridization
4. Image acquisition (scanning) and quantification (signal intensity to numbers)
5. Database building, filtering and normalization
6. Statistical analysis and data mining



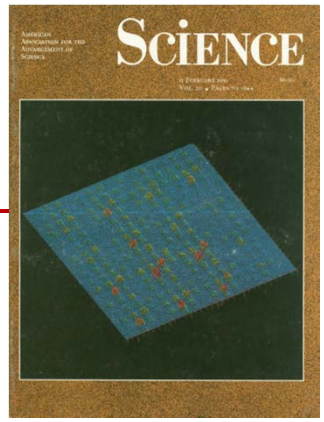
1. Manufacturing of the microarray

2. Experimental design and choice of reference: what to compare to what?
3. Target preparation (labeling) and hybridization
4. Image acquisition (scanning) and quantification (signal intensity to numbers)
5. Database building, filtering and normalization
6. Statistical analysis and data mining

The history of microarray



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



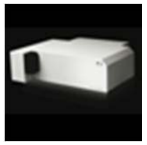
Photocamera
1989



First Microarray Scanner
1989



First Commercial Scanner
1994



Affymetrix Microarray Scanner by Hewlett-Packard
1997



Affymetrix GeneChip® Scanner 3000
2003





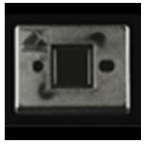
1989
First Microarray Prototype



1989
Microarray Fabrication Cell



1993
First Microarray Containing over 1 Million DNA Sequences



1994
First Commercial DNA Microarray Prototype



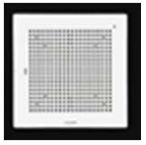
1989
500 Micron Manufacturing Mask:



1993
16-Format Manufacturing Mask



1995
49-Format Manufacturing Mask

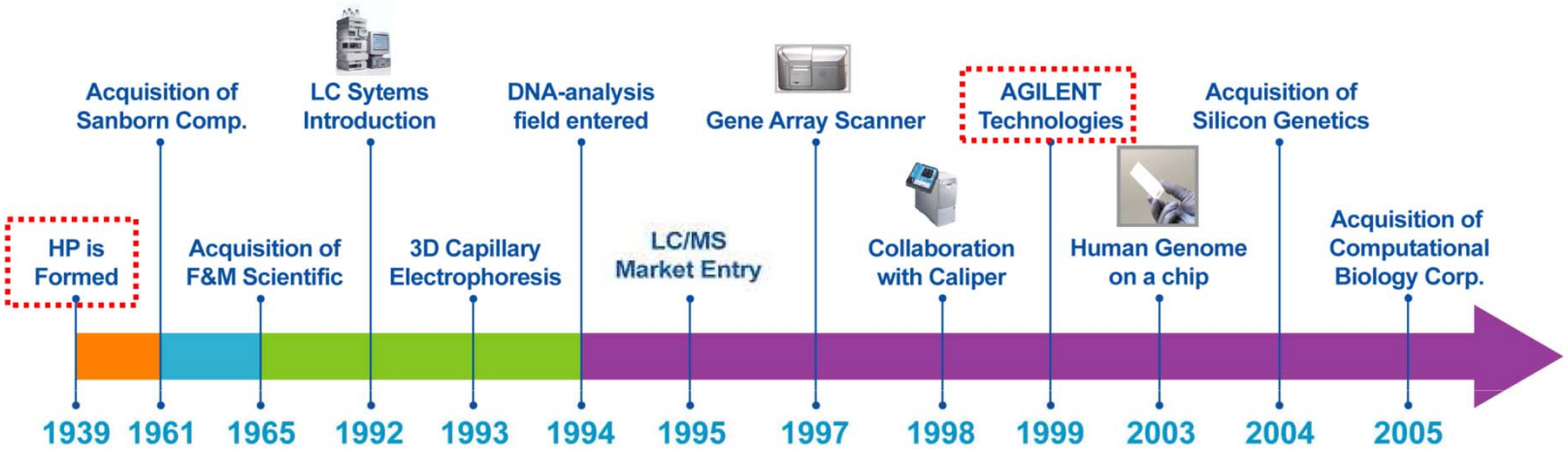


1999
400-Format Manufacturing Mask



2002
GeneChip® Human Genome U133A array

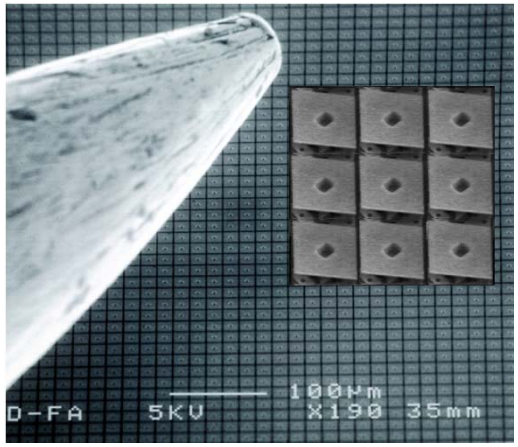
Agilent History

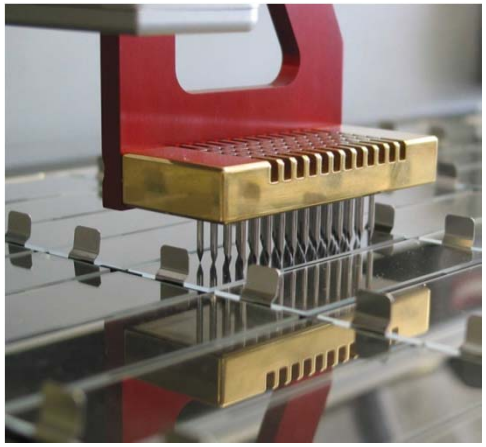


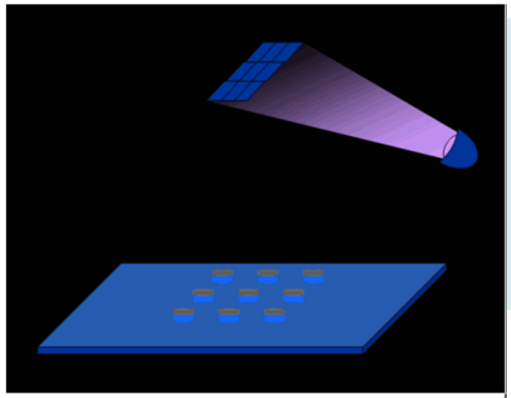


What are microarrays and how do they work ?

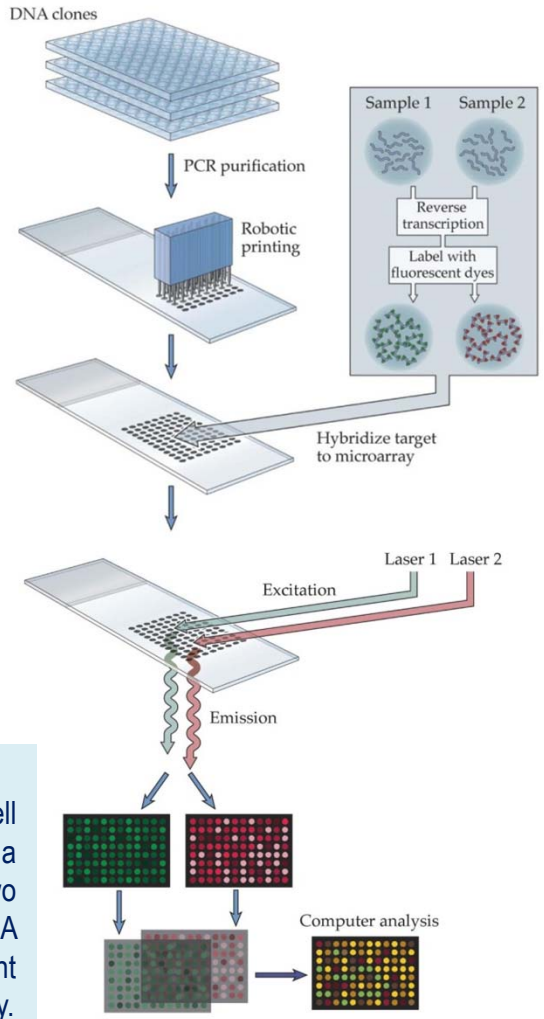
- ~10,000 spots on an array
- each spot contains ~10^7 of identical DNA of lengths from 10s to 100s of bp
- spots are either printed on the microarrays by a robot or jet, or synthesized by photolithography or by inkjet printing



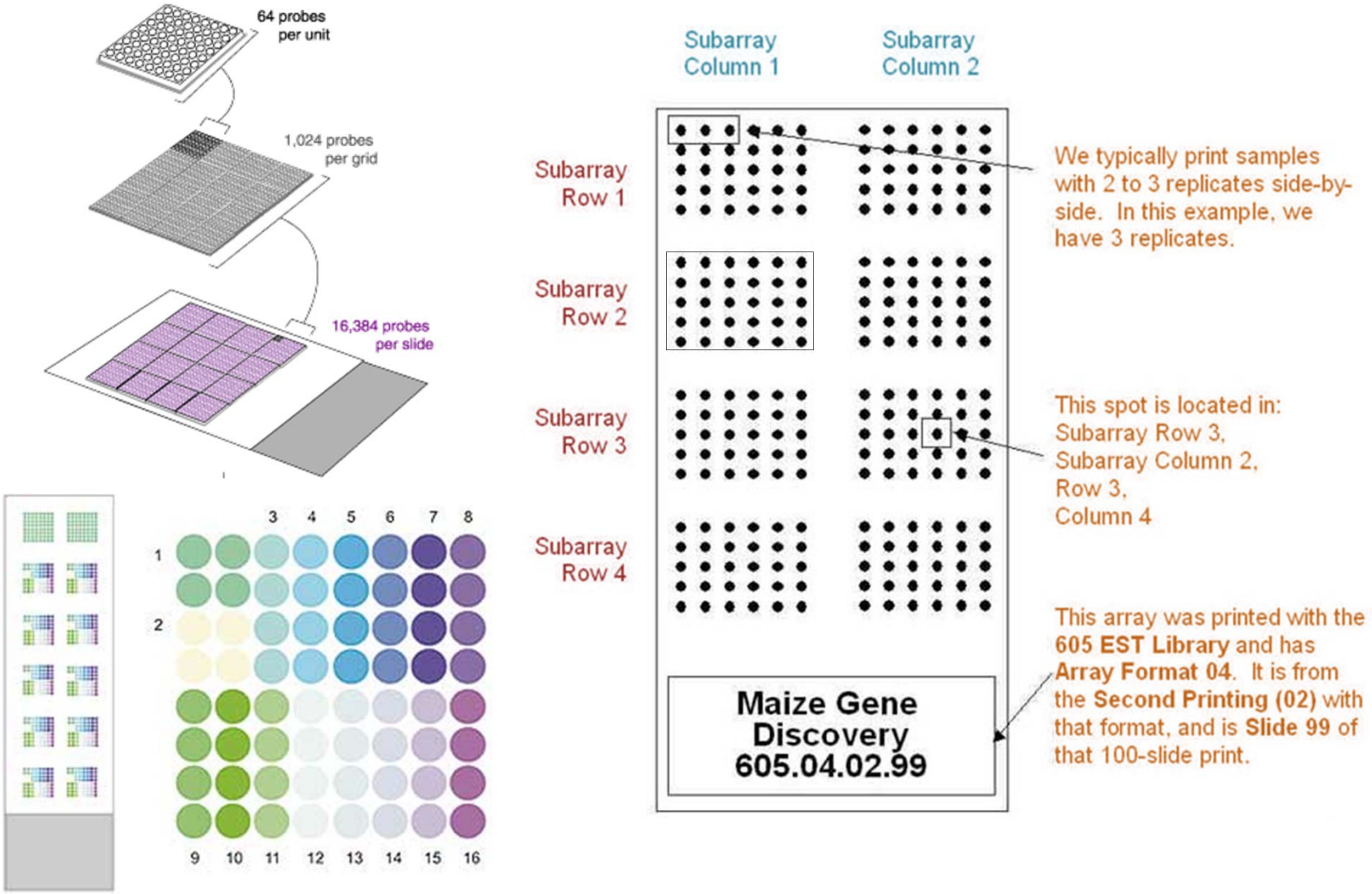



Growing oligos are protected by an acid-labile group

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.

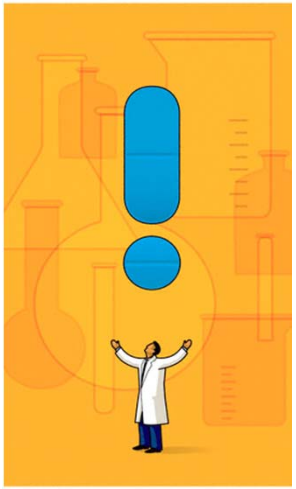


- A microarray is typically a glass or polymer slide. Smooth surface enables even deposition of surface chemistries and perfect spot morphology
- DNA molecules are attached at fixed locations called *spots* or *features*.

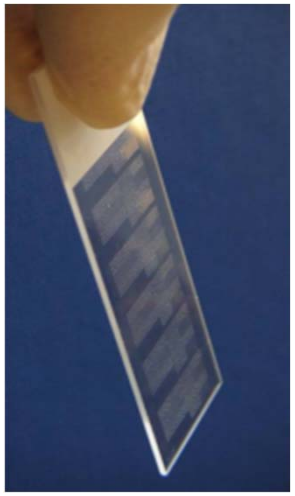


Experimental methodology

1 **Experimental Design**



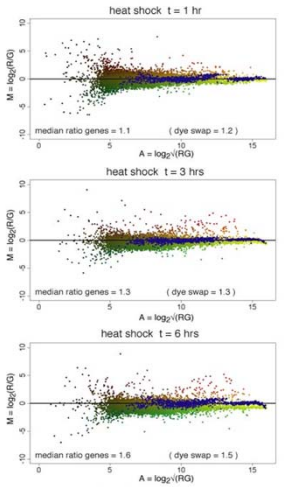
2 **Microarray Experiments**



3 **Image Processing**



4 **Data Normalization**



5 **Data Analysis**



1. Manufacturing of the microarray
2. **Experimental design and choice of reference: what to compare to what?**
3. Target preparation (labeling) and hybridization
4. Image acquisition (scanning) and quantification (signal intensity to numbers)
5. Database building, filtering and normalization
6. Statistical analysis and data mining

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

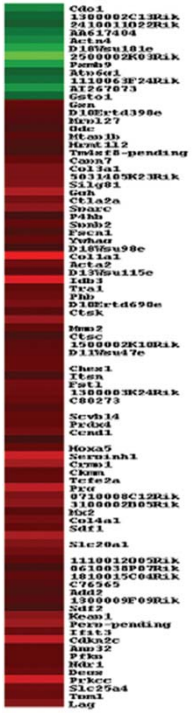
and

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

Microarray Experiment Design

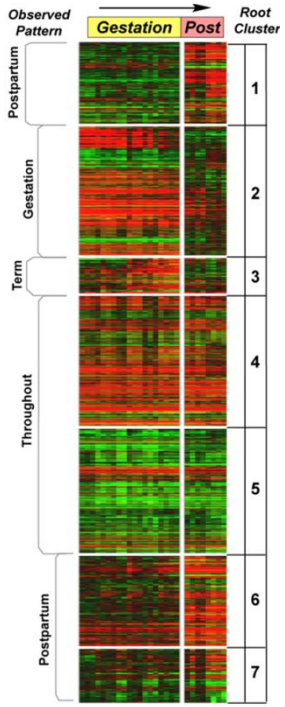
Type I: (n = 2) Differential Expression

- 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

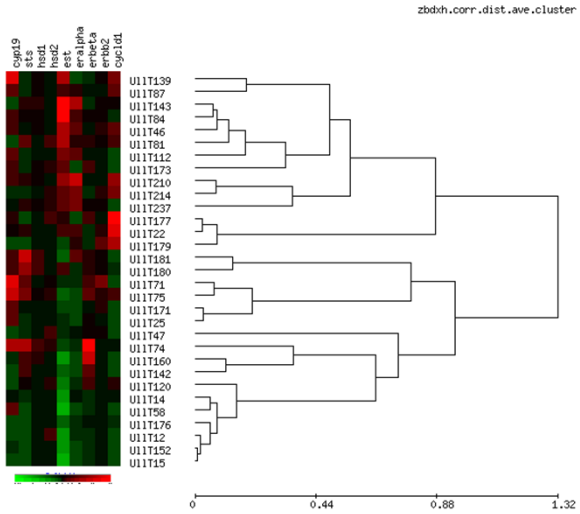
- How does the expression of gene A vary over time, tissues, or treatments?
- Do any of the expression profiles exhibit similar patterns of expression?



Reference design and Balanced block design

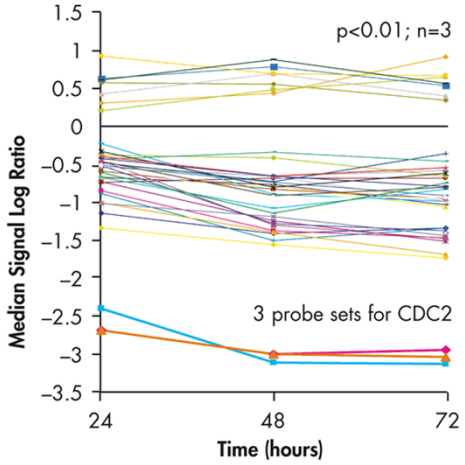
Reference design (flexible)

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.



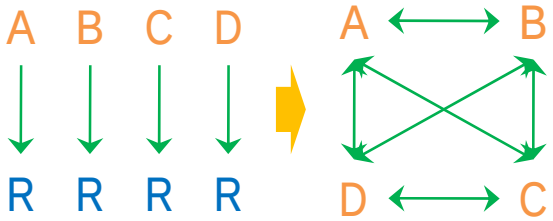
Balanced block design (efficient)

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)

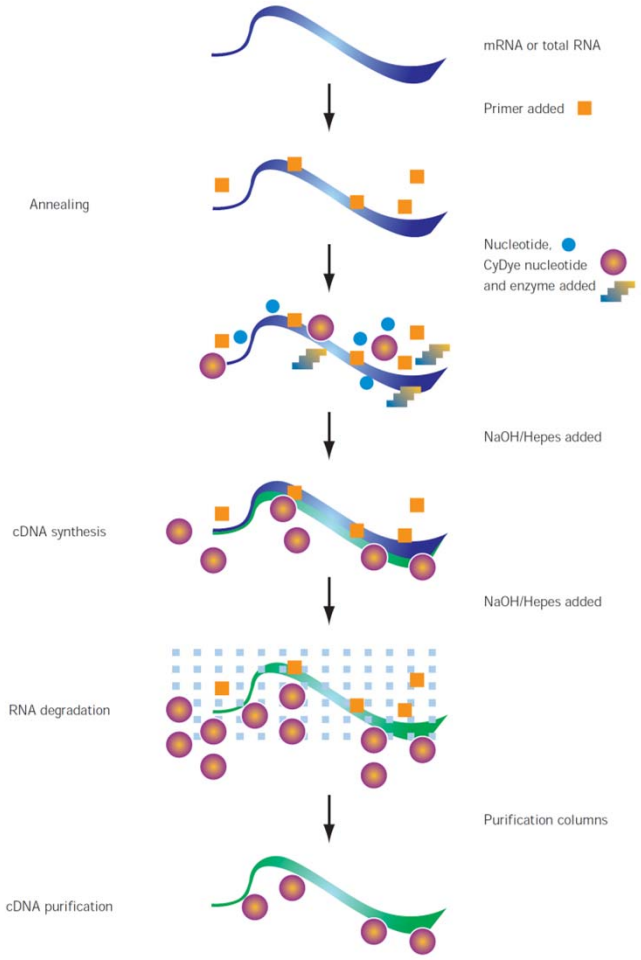
Reference RNA for reference design experiments used does not necessarily need to be physiologically relevant to your RNA. However, reference design should have at least 80% coverage for the particular slide (i.e. should contain ***mRNA representing at least 80% of genes on the slide***). RNA isolated 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. Manufacturing of the microarray
2. Experimental design and choice of reference: what to compare to what?

3. Target preparation (labeling) and hybridization

4. Image acquisition (scanning) and quantification (signal intensity to numbers)
5. Database building, filtering and normalization
6. Statistical analysis and data mining



❶

Attachment

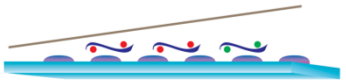
Genes of interest are spotted onto a solid surface by the array spotter. These are known as the targets. Attachment chemistry will often be required to ensure that the DNA remains attached to the slide surface throughout the hybridization process.



❷

Hybridization

Hybridization buffer containing a known amount of labelled sample DNA—often referred to as probe— is then placed on the slide surface. A coverslip can then be carefully placed on top of the slide.



The slide is then incubated in a humid environment for up to 16 h. During this time the labelled probe is in contact with the targets on the slide. If the sequence homology is good then the probe will adhere to the target.



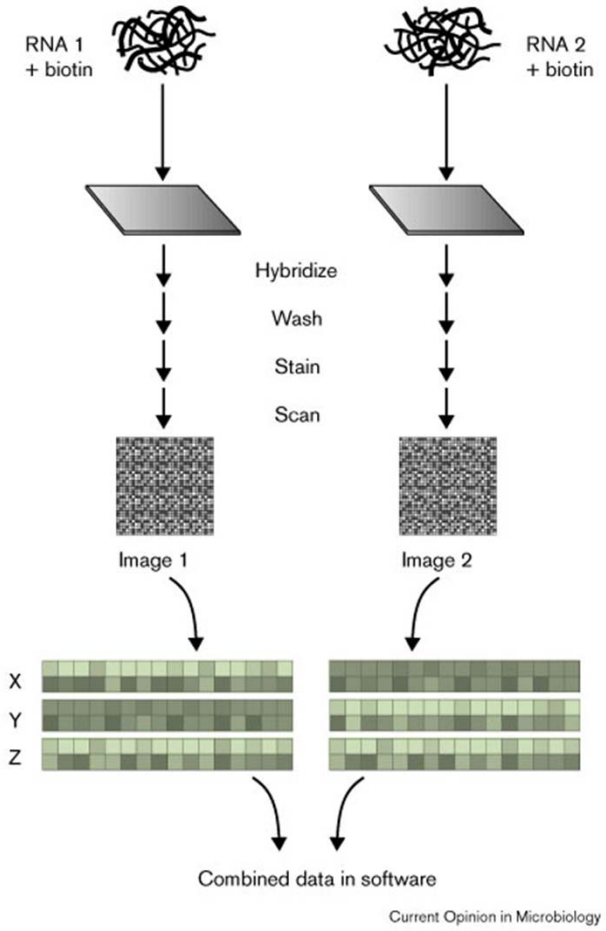
❸

Washing

Once the hybridization is complete, the slides are washed, and buffer and probe of little or no homology to the target will be washed away, leaving the labelled probe of high homology attached to the target and available for detection.



1. Manufacturing of the microarray
2. Experimental design and choice of reference: what to compare to what?
3. Target preparation (labeling) and hybridization
4. **Image acquisition (scanning) and quantification (signal intensity to numbers)**
5. Database building, filtering and normalization
6. Statistical analysis and data mining



Samples

Platform

Raw Data *(image, 16bit .tif file)*

(1) image acquisition
(2) spot location

Preprocessed Data *(.gal file)*

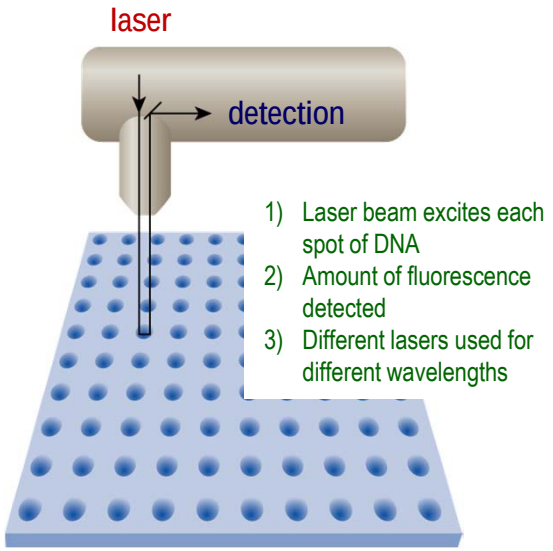
(3) computation of spot intensities

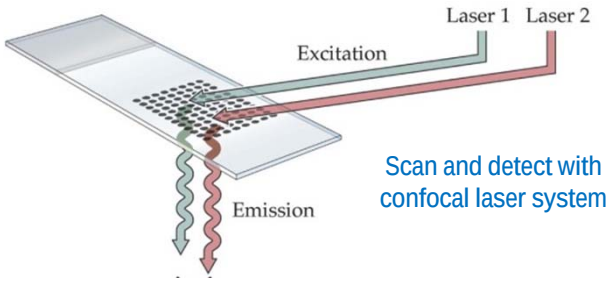
Data Analysis (different levels)

(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 (uncompress
- Separate image for each fluorescent sample
  - channel 1, channel 2





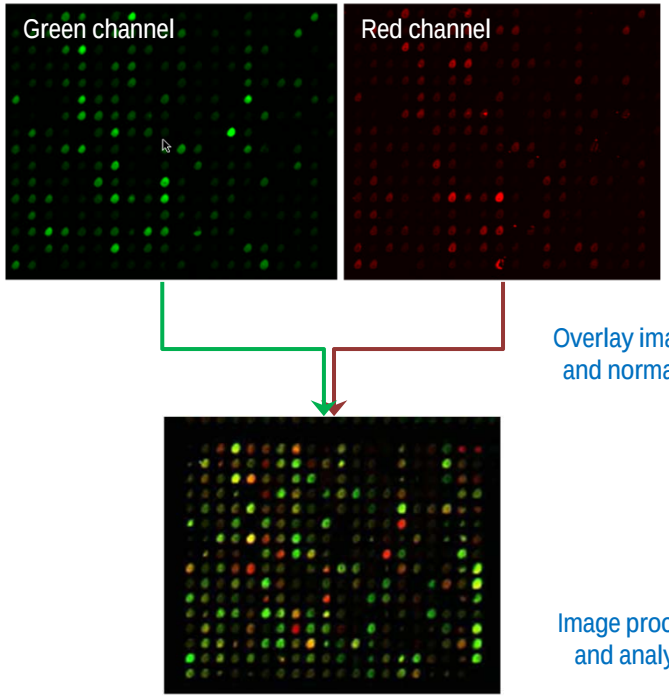
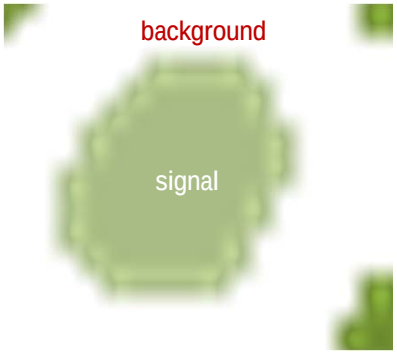
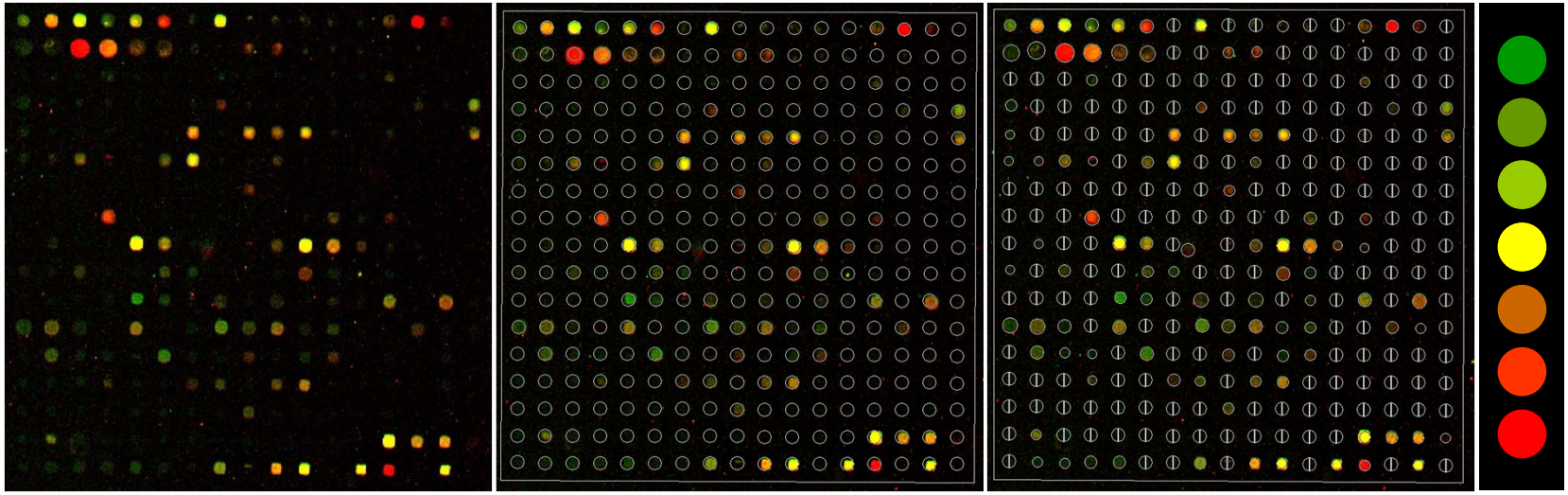


Image Processing

- **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

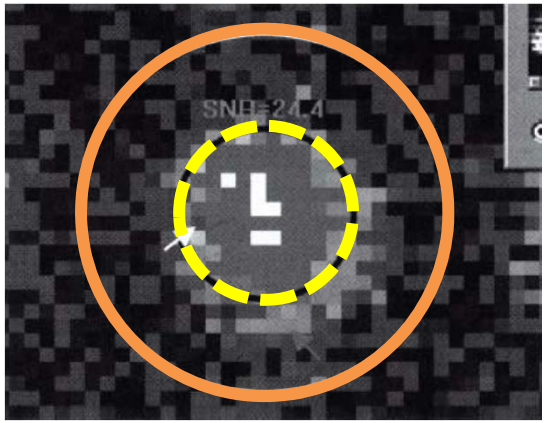
Quality Measurements

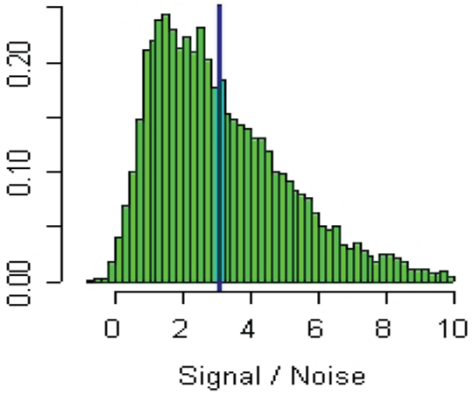
- **Array**
  - Correlation between duplicate spot intensities
  - Percentage of spots with negative signals
  - Distribution of actual spot signal area vs. idealized
  - Inter-array consistency
- **Spot**
  - Signal / Noise ratio
  - Variation in pixel intensities within spots

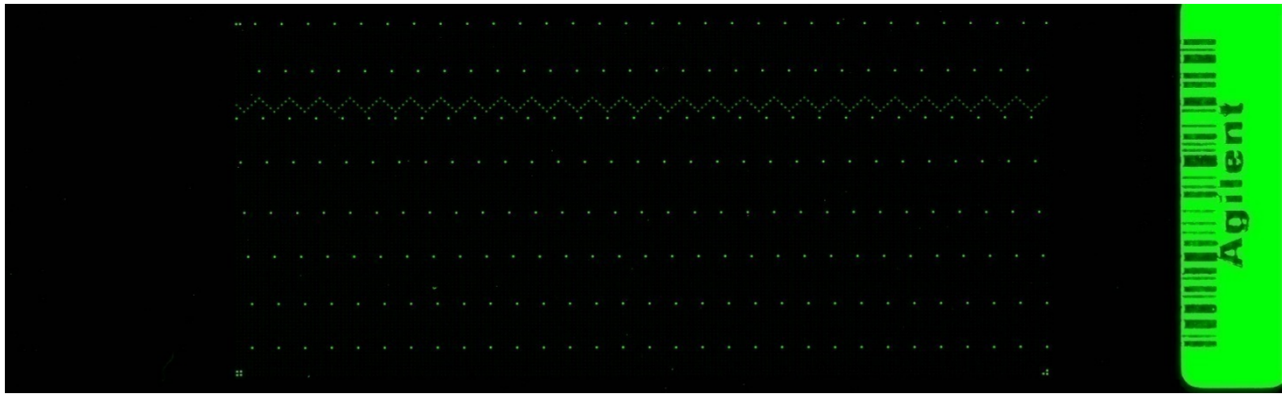
$$\text{SNR (signal-to-noise ratio)} = \frac{\text{Median signal intensity*}}{\text{Median noise value}}$$

, where signal = foreground, noise = background

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



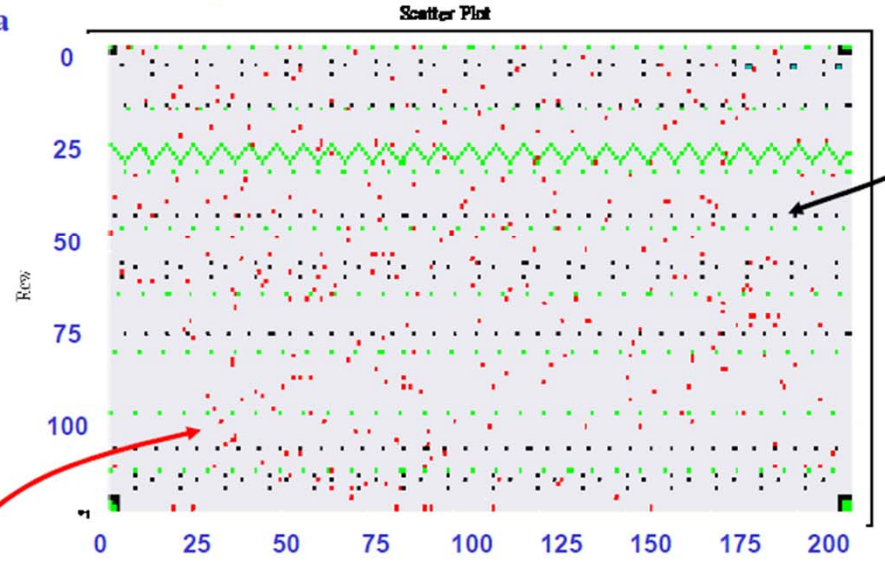




Agilent Human Array includes:

2.Negative control
- Thundered features
- non-homologous to human sequences

3. Experimental probes
- Thundered genes with 10 replicates to monitor data

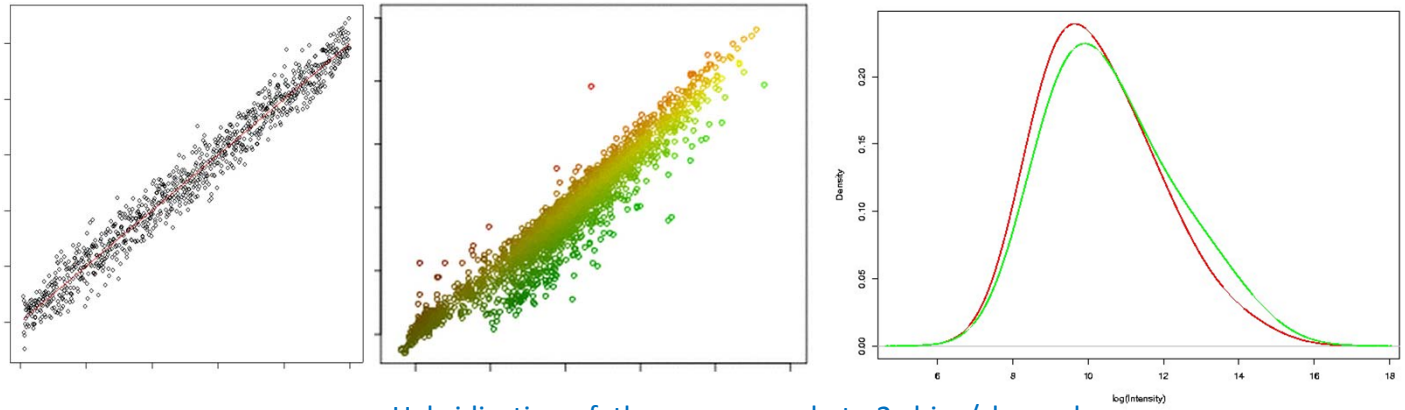


1. Spiken-In RNA controls
- Thundered features

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

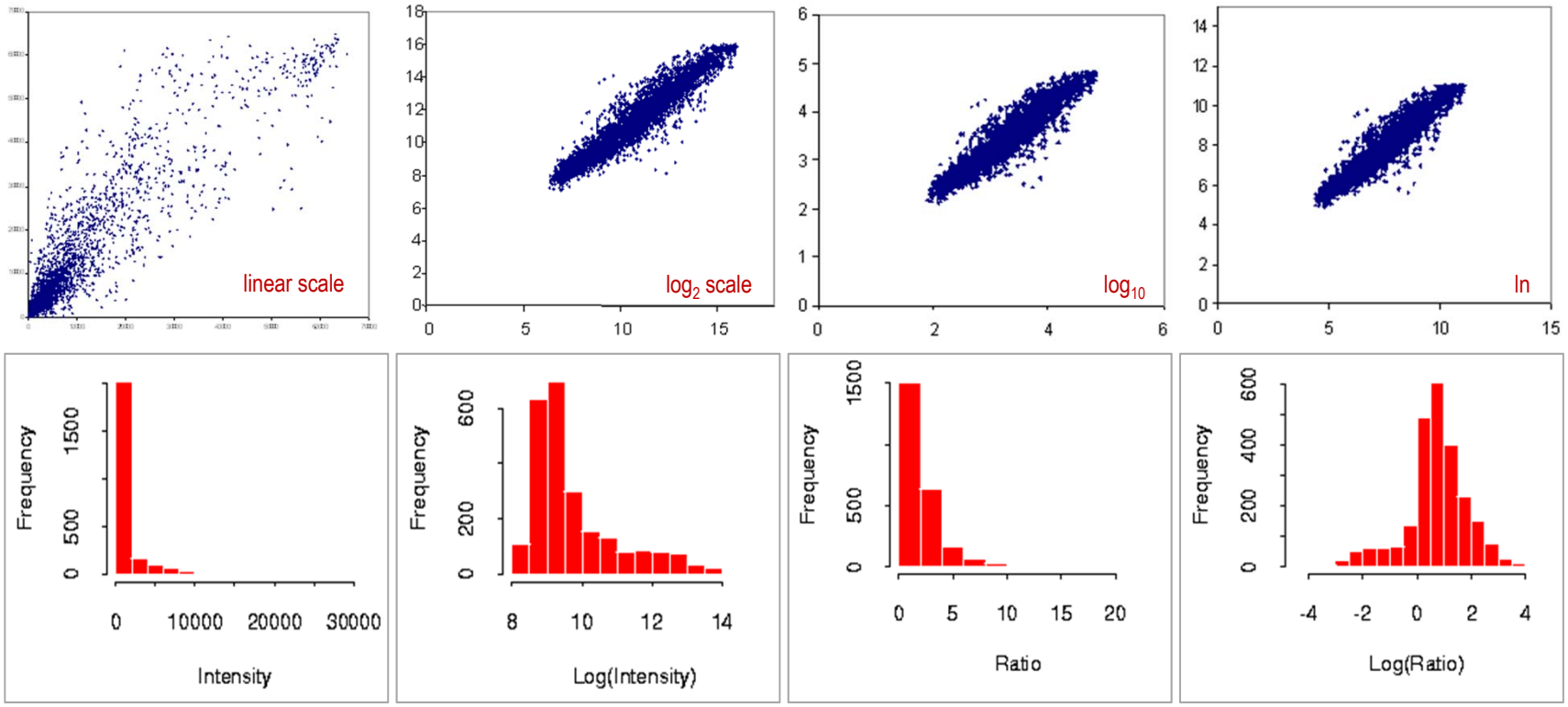
- Ideally: scatter plot coincides with the x=y diagonal
  - Due to Random errors: we expect to see a 'cloud' around the x=y diagonal.
- In practice: Both Random and Systematic measurement errors (Bias)
  - Due to Biases scatter plots are not centered around the x-y diagonal

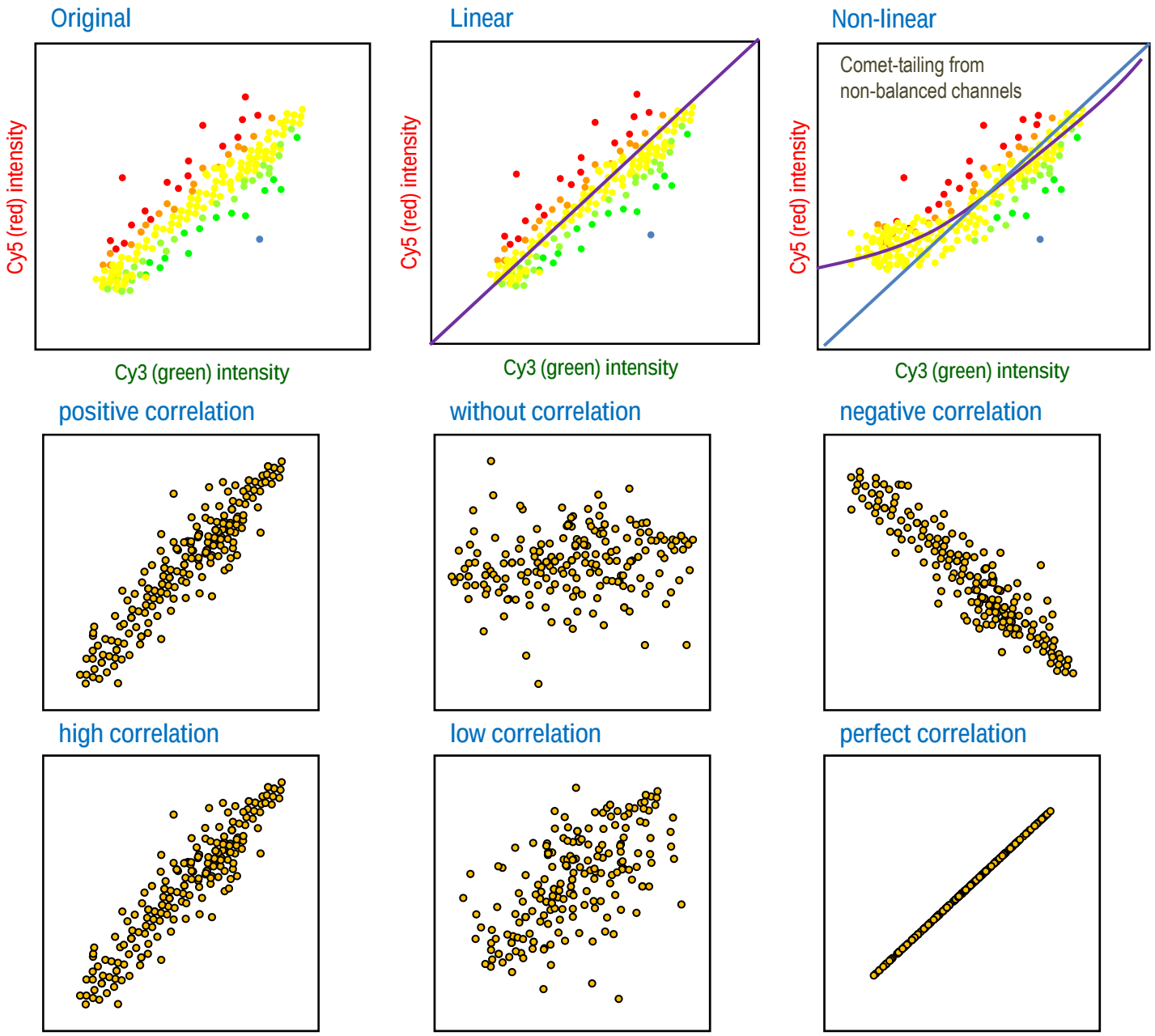


Hybridization of the same sample to 2 chips/channels

Visualizing the expression data – *log* transformation

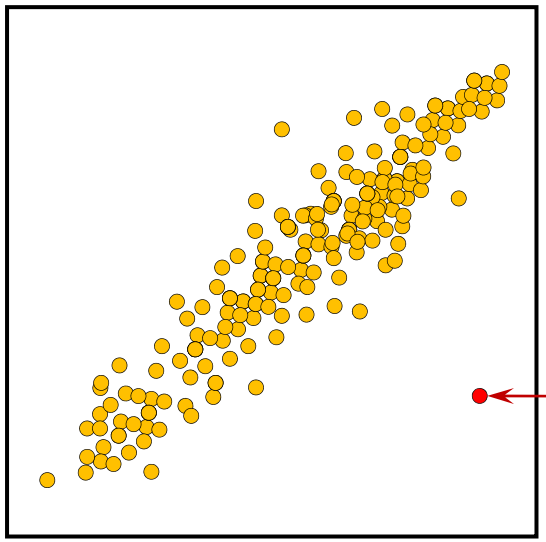
- Makes variation of intensities and ratios of intensities more independent of absolute magnitude
- Evens out highly skewed distributions
- Gives more realistic sense of variation
- Approximates normal distribution
- Treats up- and down- regulated genes symmetrically



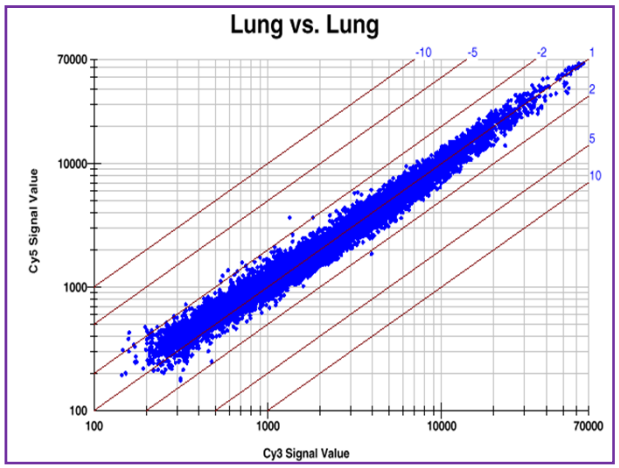

Original
Cy5 (red) intensity
Cy3 (green) intensity
Linear
Cy5 (red) intensity
Cy3 (green) intensity
Non-linear
Comet-tailing from
non-balanced channels
Cy5 (red) intensity
Cy3 (green) intensity
positive correlation
without correlation
negative correlation
high correlation
low correlation
perfect correlation

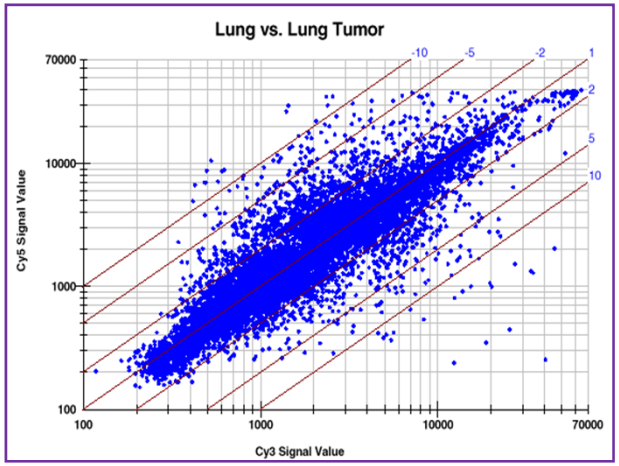
Correlation and Outliers

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



Experimental error or something important?

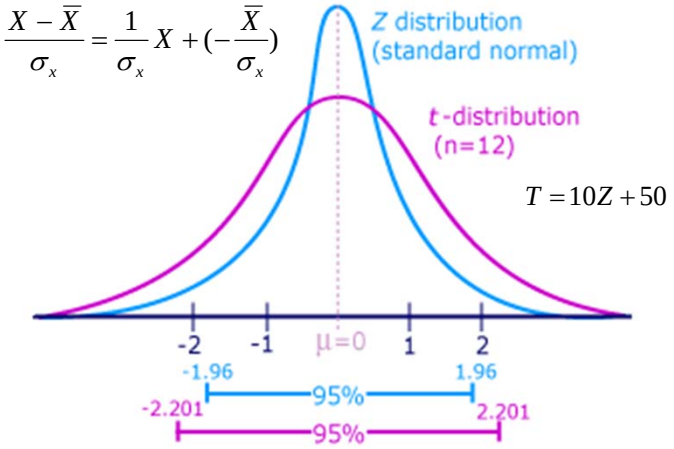




1. Manufacturing of the microarray
2. Experimental design and choice of reference: what to compare to what?
3. Target preparation (labeling) and hybridization
4. Image acquisition (scanning) and quantification (signal intensity to numbers)
5. **Database building, filtering and normalization**
6. Statistical analysis and data mining

- **Reduces systematic (multiplicative) differences between two channels of a single hybridization or differences between hybridizations**
- **Sources of Systematic Bias**
  - Different dye labeling efficiencies
  - Scanning (laser and detector, chemistry of the fluorescent label)
  - Differences in concentration of DNA on arrays (plate effects)
  - Differences in total mRNA in one sample versus another or mRNA degradation
  - Printing or tip problems
  - Uneven hybridization
- **Several Methods:**
  - Global mean method
  - Linear regression method
  - Curvilinear methods
    (e.g. LOWESS; **LO**cally **WE**ighted **S**catterplot **S**moothing)
  - Variance model methods

$$Z = \frac{X - \overline{X}}{\sigma_x} = \frac{1}{\sigma_x} X + (-\frac{\overline{X}}{\sigma_x})$$

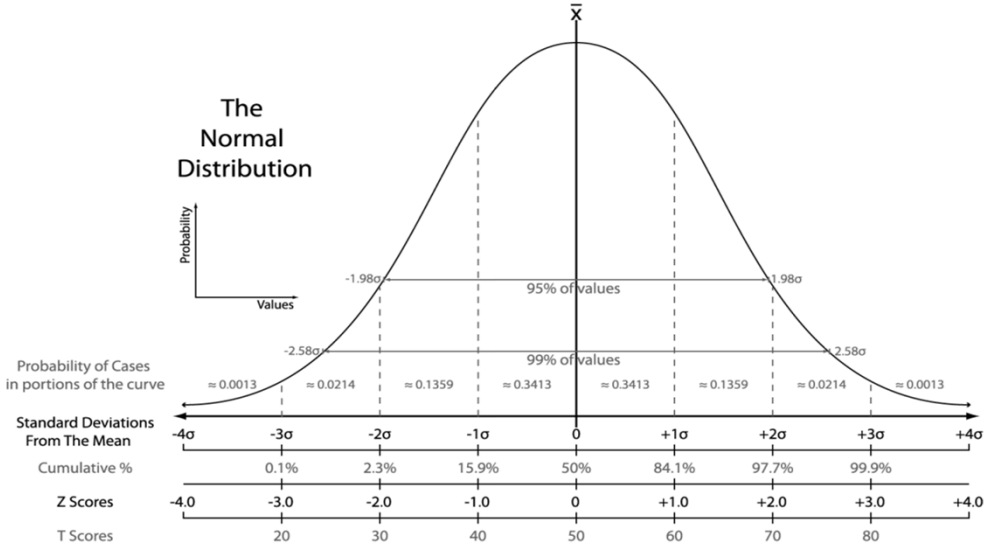


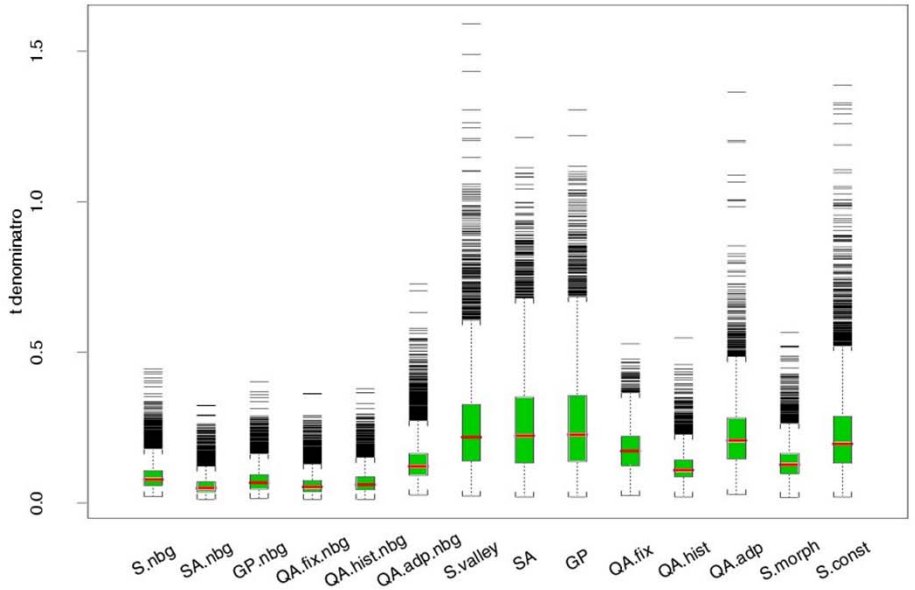
- Sigma (z) Score - the deviation from the mean divided by the standard deviation.

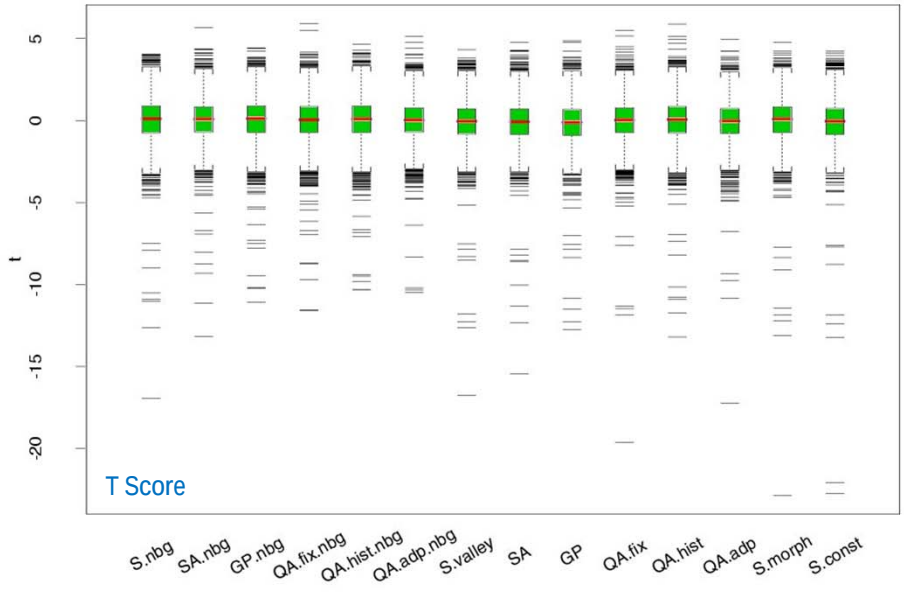
$$z_i = (x_i - \mu) / \sigma$$

- 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.

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





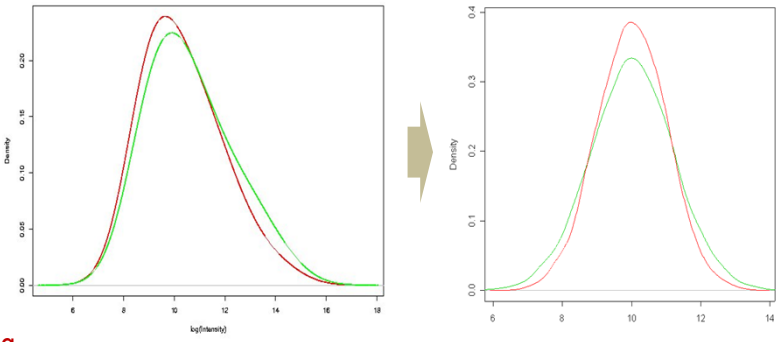


Global Mean method -cross slide normalization

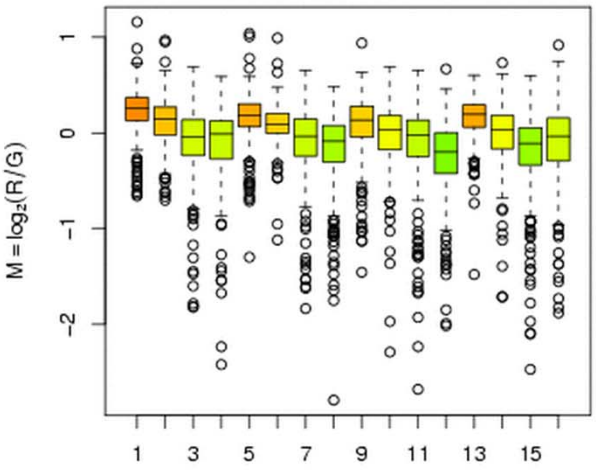
- A single normalization factor (k) is computed for balancing chips/channels:

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

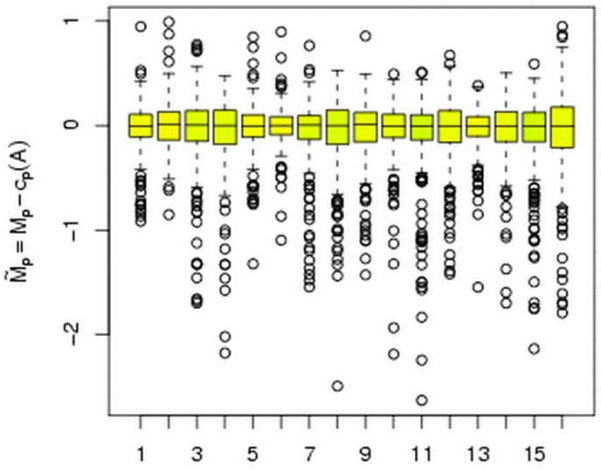
- Multiplying intensities by this factor equalizes the mean (median) intensity among compared chips
  - Calculate mean intensity of all spots in channels 1 & 2
    e.g. μch1 = 20,000
    μch2 = 25,000
    μch2/μch1 = 1.25
  - On average, spots in ch2 are 1.25X brighter than spots in ch1
  - To normalize,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

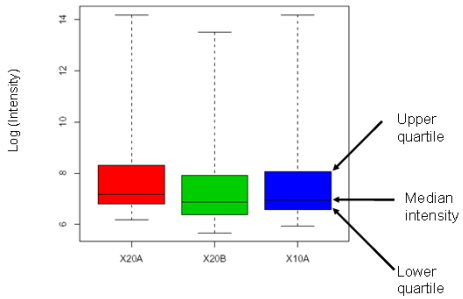


Before Normalization



After Scaling



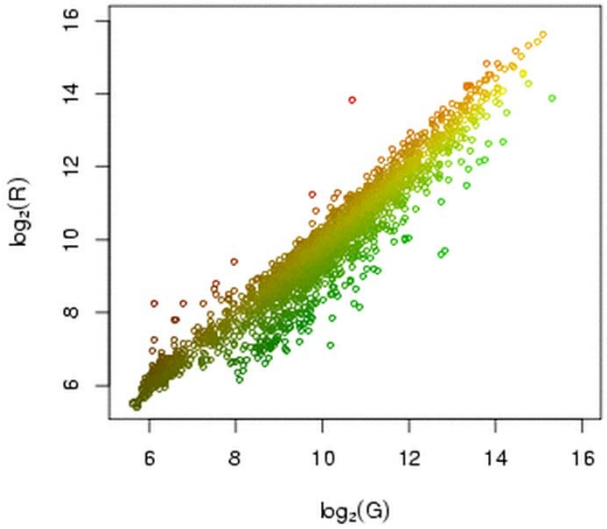


Global (*LOWESS*) Normalization -within slide normalization

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

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

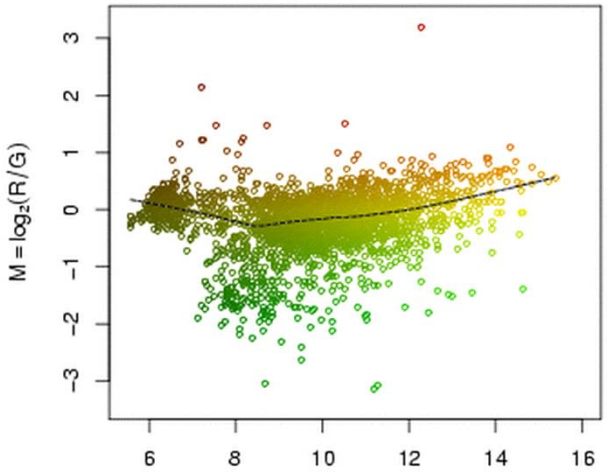
R = red channel signal
G = green channel signal
(background corrected or not)



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)

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

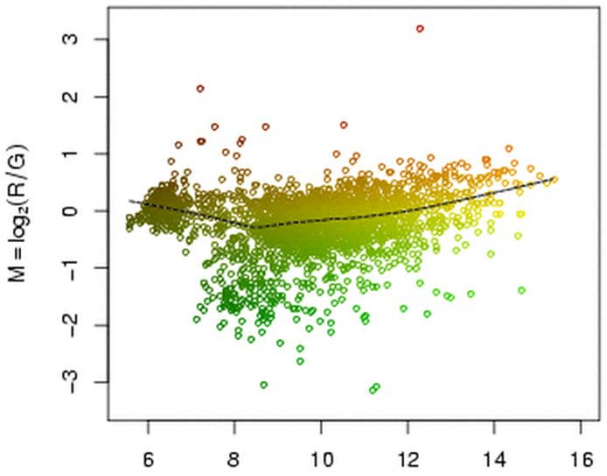


LOWESS Normalization

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

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

$M = \log_2(R/G)$

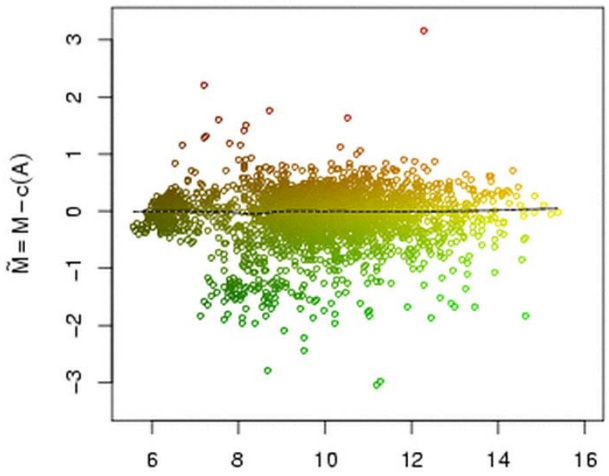


Global normalized data $\{(M, A)\}_{n=1..n}$

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

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

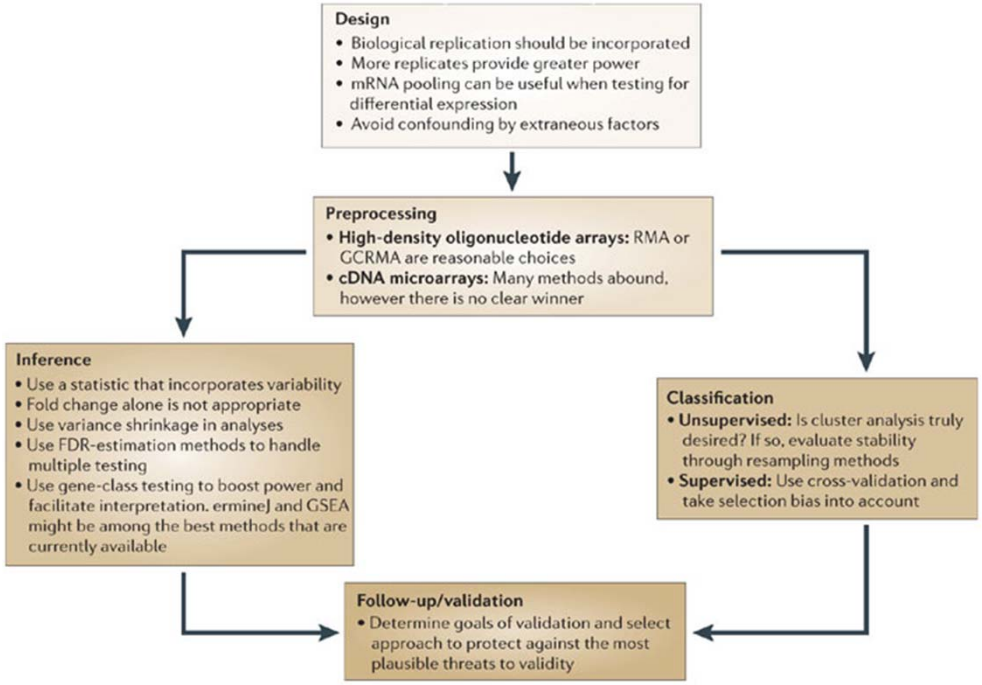
$c(A)$ is an *intensity dependent* function.
or c(A) is the (lowess fit) to the MA-plot.



1. Manufacturing of the microarray
2. Experimental design and choice of reference: what to compare to what?
3. Target preparation (labeling) and hybridization
4. Image acquisition (scanning) and quantification (signal intensity to numbers)
5. Database building, filtering and normalization

6. Statistical analysis and data mining

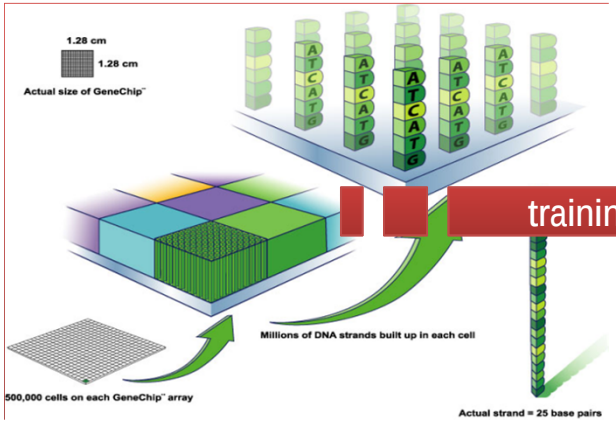
Guidelines for the statistical analysis of microarray experiments.



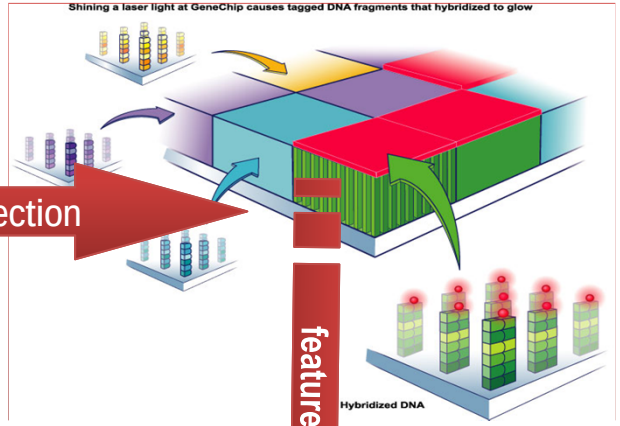
Copyright © 2006 Nature Publishing Group
Nature Reviews | Genetics

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

Single-Test Platform of Microarray & Knowledge Discovery



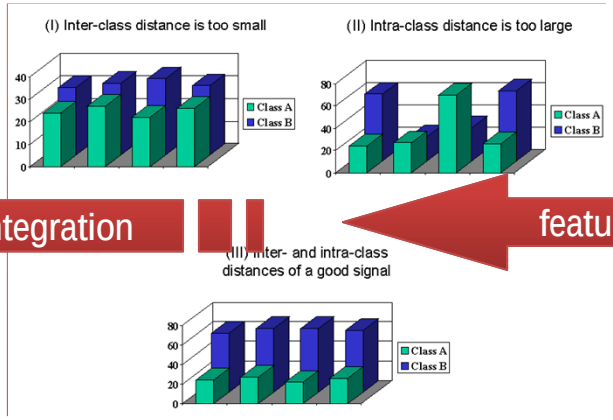
training data collection



feature generation

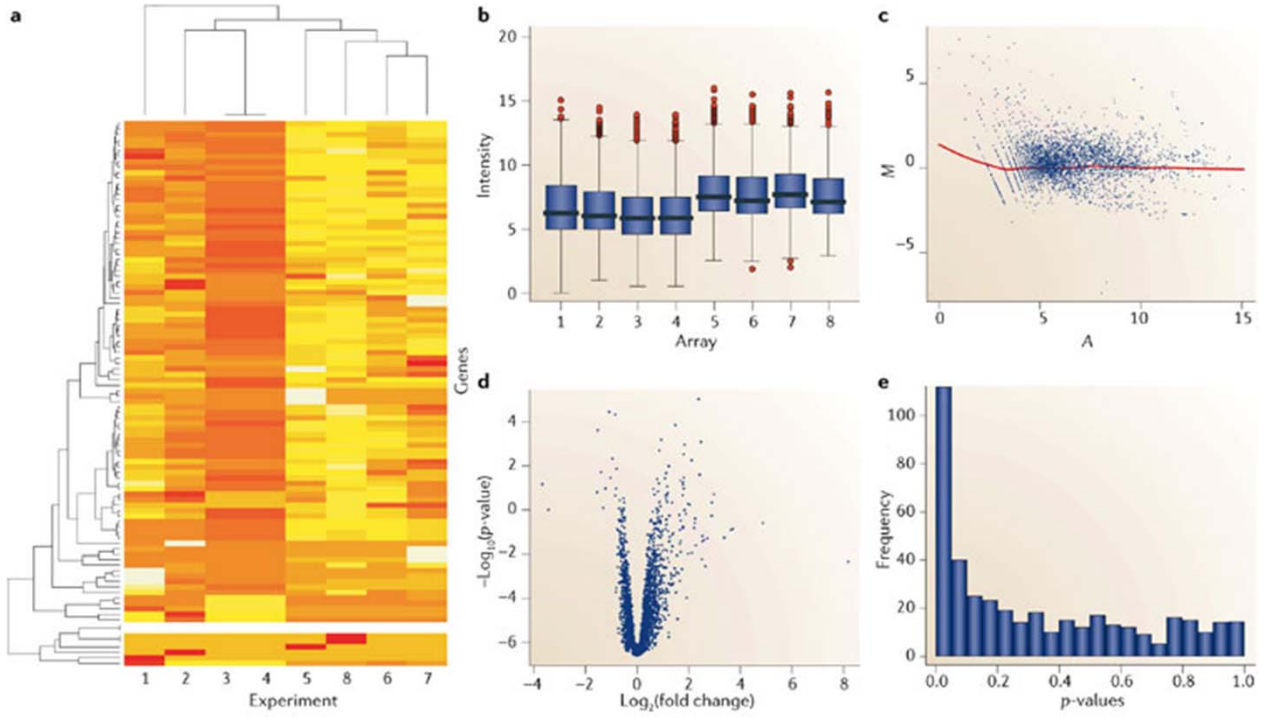
	00-0586-U	00-0586-U	00	U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative			Avg Diff	Abs Call	
AFFX-Murl	5	2			297.5	A	M16762 Mouse int
AFFX-Murl	3	2			554.2	A	M37897 Mouse int
AFFX-Murl	4	2		9	308.6	A	M25892 Mus musc
AFFX-Murl	1	3		19	141	A	M83649 Mus musc
AFFX-BioE	13	1		19	9340.6	P	J04423 E coli bioB
AFFX-BioE	15	0		19	12862.4	P	J04423 E coli bioB
				19	8716.5	P	J04423 E coli bioB
				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

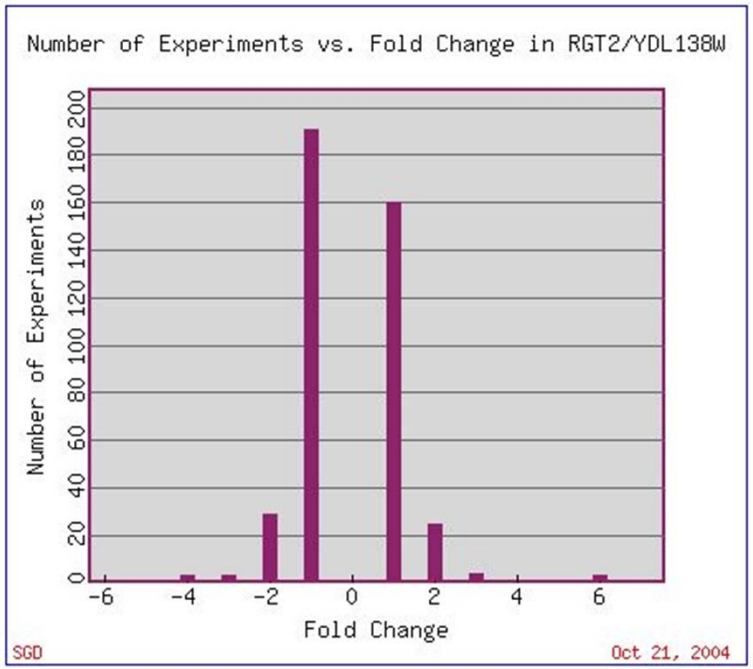
1. Find the genes that change expression between experimental and control samples
2. Classify samples based on a gene expression profile
3. Find patterns: Groups of biologically related genes that change expression together across samples and treatments

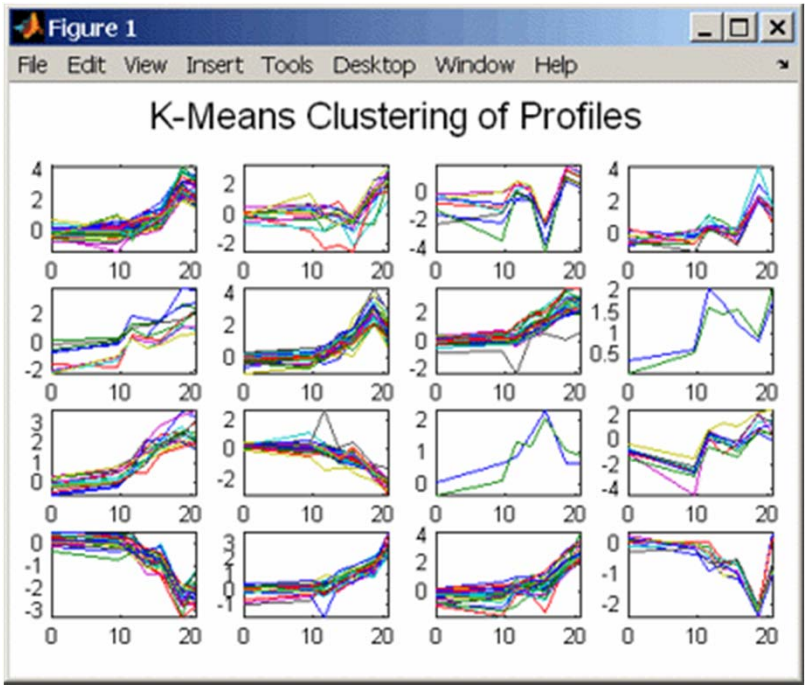


Allison DB *et al.* (2005) Microarray data analysis: from disarray to consolidation and consensus *Nat Rev gene.* 7: 55–65 doi:10.1038/nri1749

Basic Data Analysis

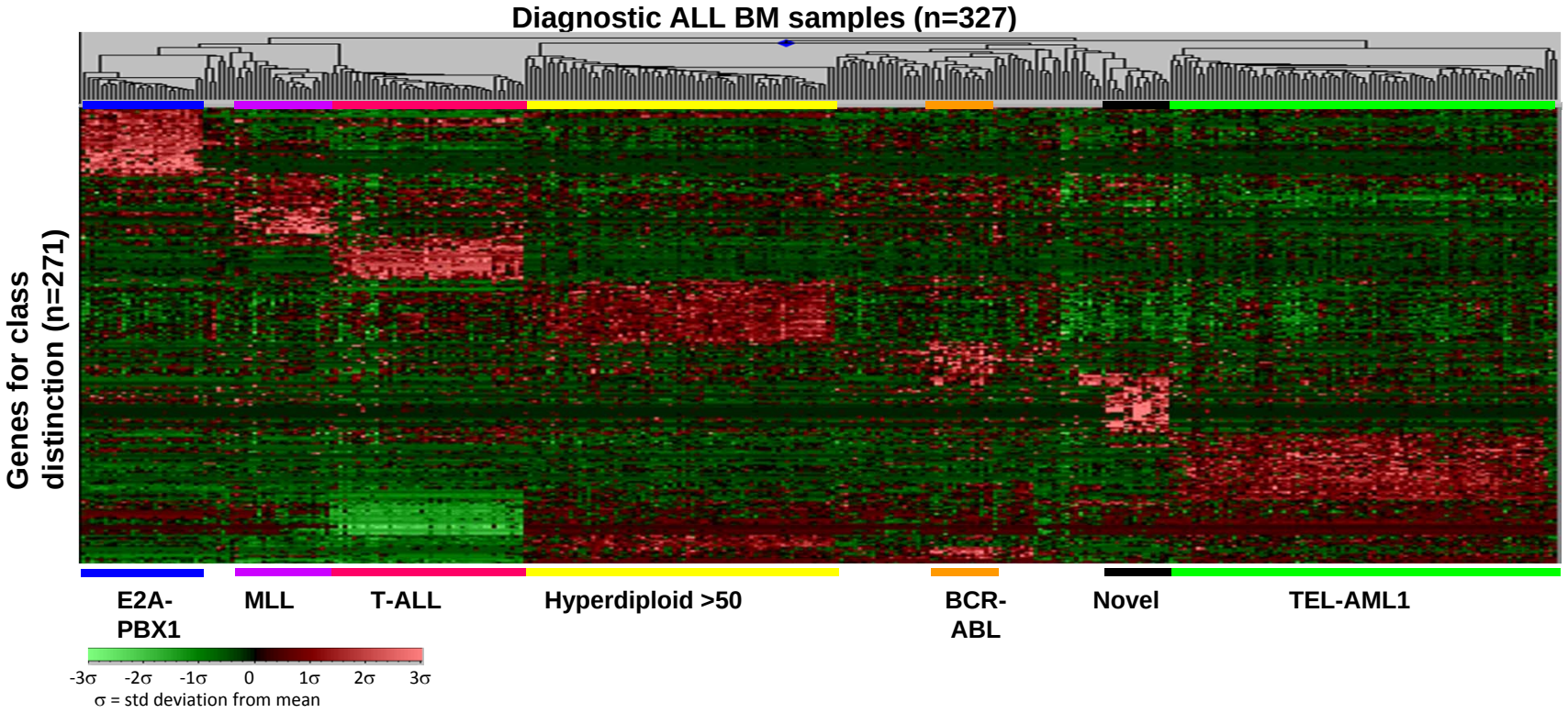
- Fold change (relative increase or decrease in intensity for each gene)
- Set cutoff filter for low values (background +noise)
- Cluster genes by similar changes - only really meaningful across multiple treatments or time points
- Cluster samples by similar gene expression profiles





Higher Level Microarray data analysis

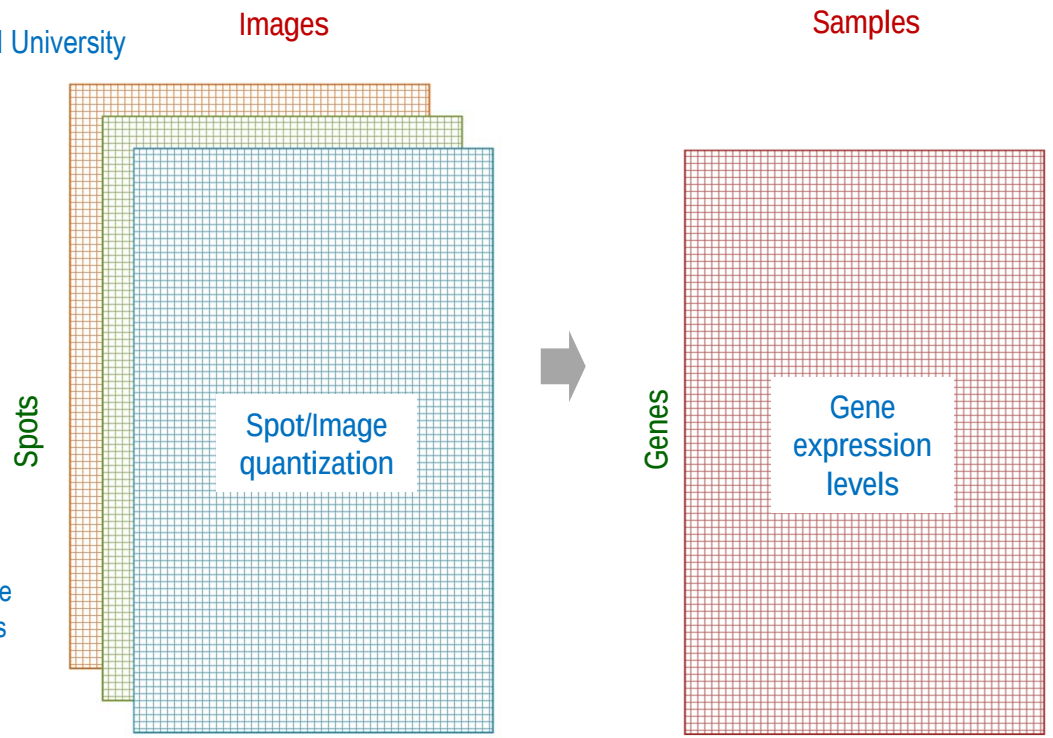
- Clustering and pattern detection
- Data mining and visualization
- Controls and normalization of results
- Statistical validatation
- Linkage between gene expression data and gene sequence/function/metabolic pathways databases
- Discovery of common sequences in co-regulated genes
- Meta-studies using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

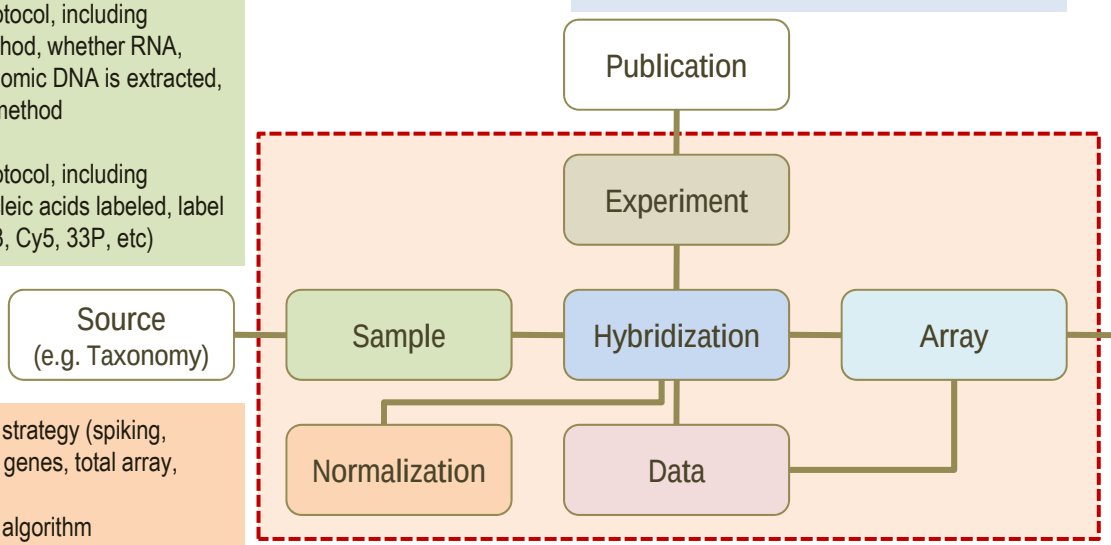
- A need to establish a public repository or repositories for microarray gene expression data became apparent in 1998, which requires data standards
- MGED-1 (Microarray Gene Expression Database) Group: November 14-15, 1999, Cambridge, UK
  - Established *five* working groups, including the microarray 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. The solution (e.g. concentration of solutes)
2. Blocking agent
3. Wash procedure
4. Quantity of labelled target used
5. Time, concentration, volume, temperature
6. Description of the hybridization instruments

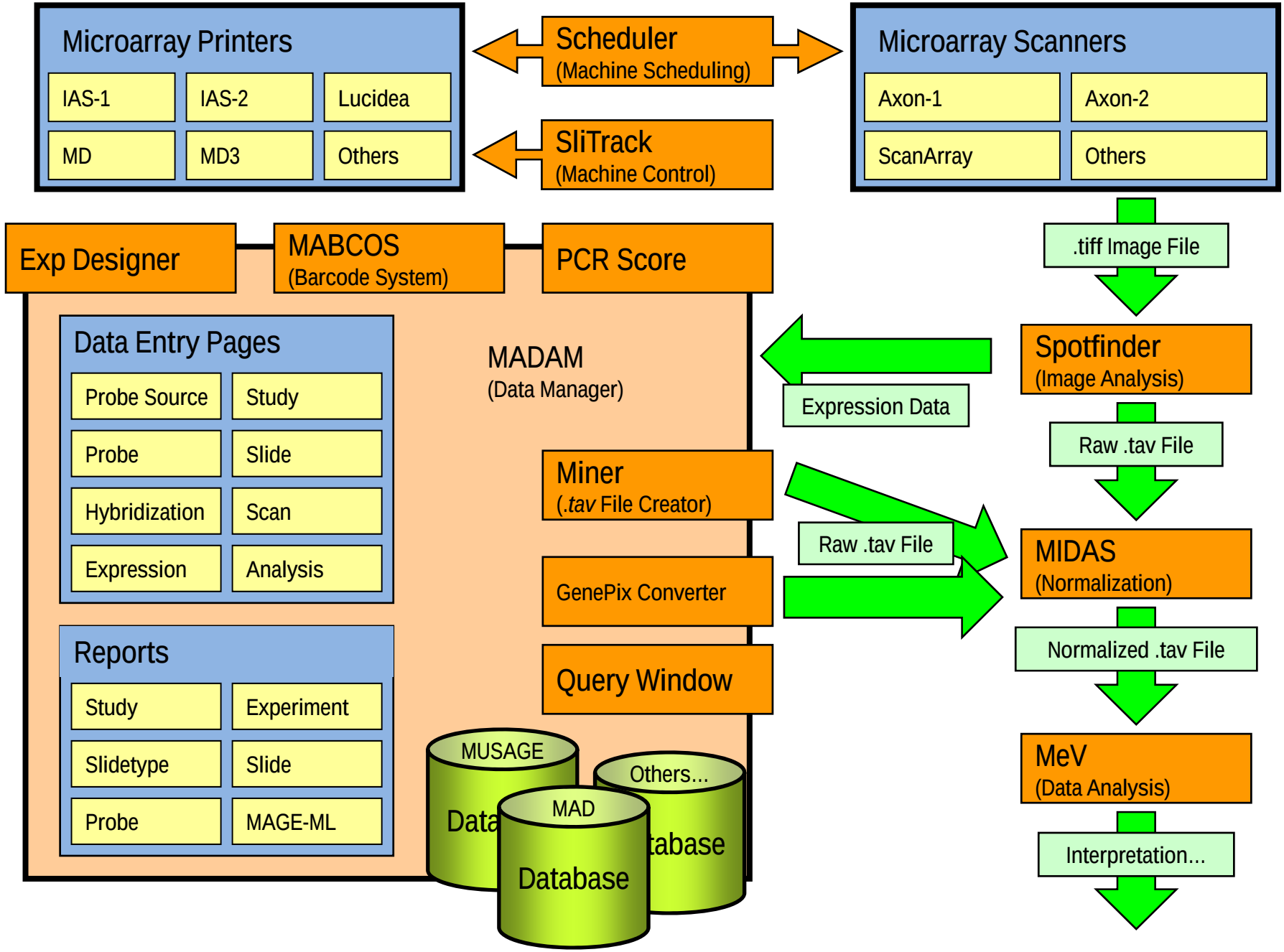
1. Sample source and treatment
2. Hybridization extract preparation
3. Laboratory protocol, including extraction method, whether RNA, mRNA, or genomic DNA is extracted, amplification method
4. Labeling
5. Laboratory protocol, including amount of nucleic acids labeled, label used (e.g. Cy3, Cy5, 33P, etc)

1. Author, contact information, citations
2. Type of experiment (e.g., time course, normal vs diseased comparison)
3. Experimental factors – i.e. tested parameters in the experiment (e.g. time, dose, genetic variation, response to a compound)
4. List of organisms used in the experiment
5. List of platforms used

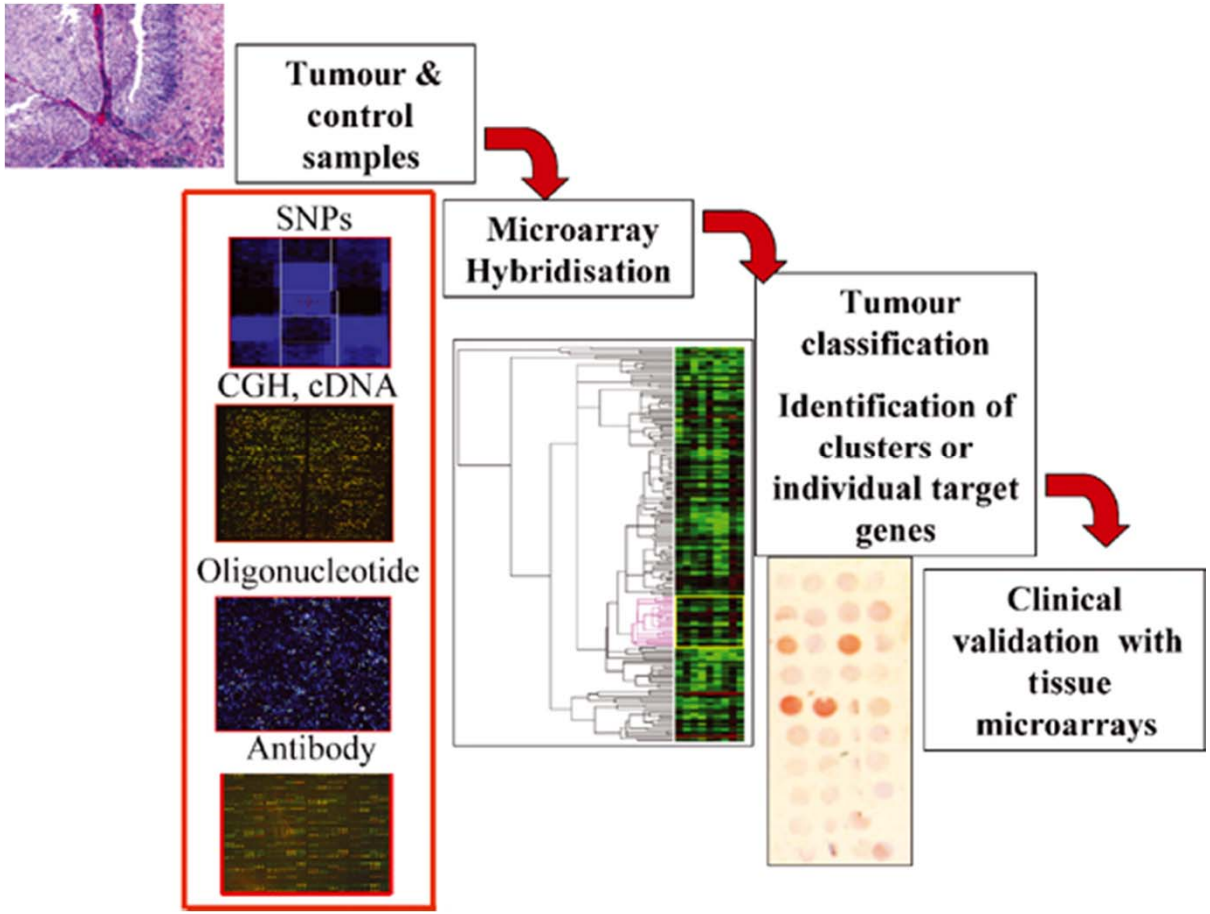
1. Normalization strategy (spiking, housekeeping genes, total array, other)
2. Normalization algorithm
3. Control array elements

1. Array design related information (e.g. platform type – *in situ* synthesized or spotted, array provider, surface type – glass, membrane, other, etc)
2. Properties of each type of elements on the array, that are generated by similar protocols (e.g. synthesized oligos, PCR products, plasmids, colonies, others) – may be simple or composite (Affymetrix)
3. Each element (spot) on the array

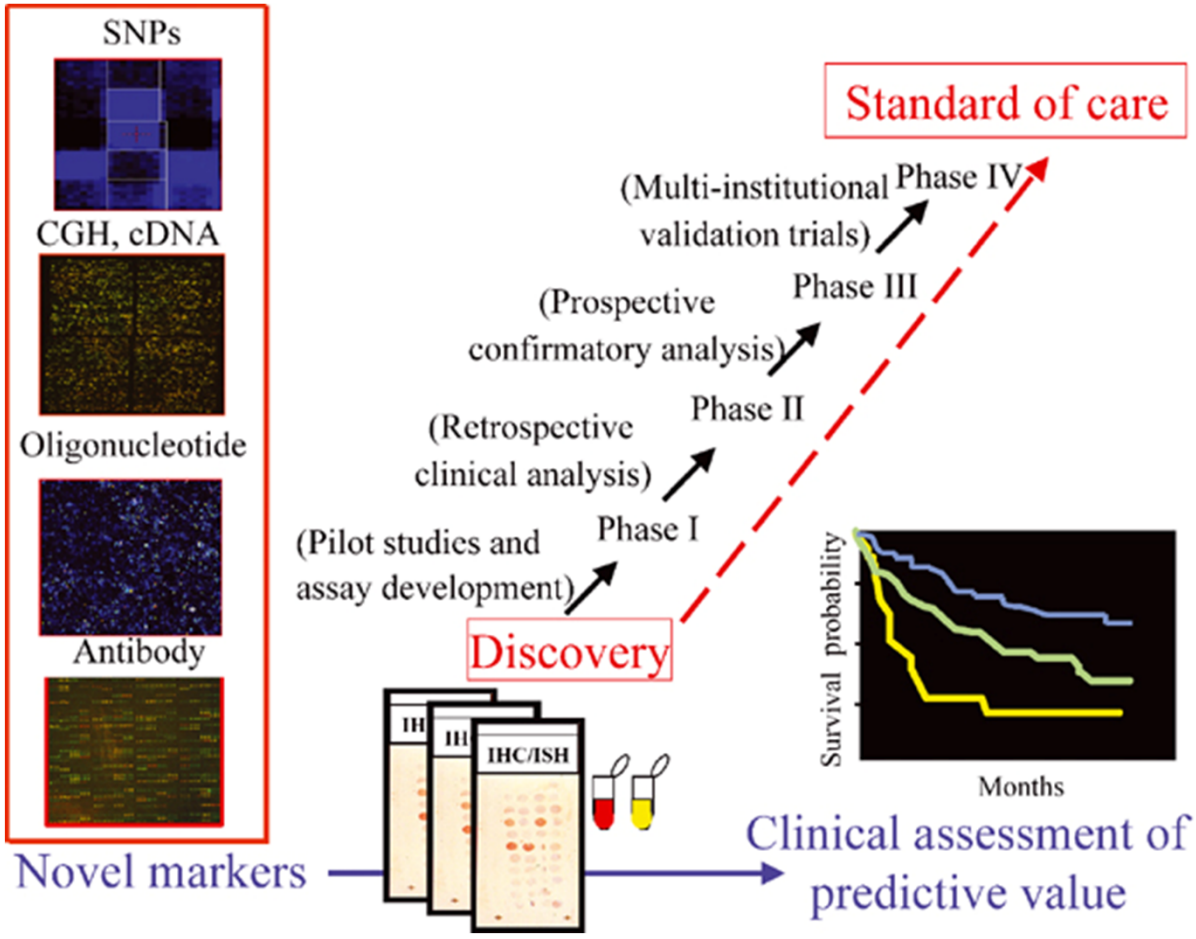
1. Scanning information
   - Scan parameters, including laser power, spatial resolution, pixel space, PMT voltage
   - Laboratory protocol for scanning, including scanning hardware and software used
2. Image analysis information
   - Image analysis software specification
   - All parameters
3. Summarized information from possible replicates


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
MIDAS
(Normalization)
Normalized .tav File
MeV
(Data Analysis)
Interpretation...
Data Entry Pages
Probe Source
Study
Probe
Slide
Hybridization
Scan
Expression
Analysis
MADAM
(Data Manager)
Expression Data
Miner
(.tav File Creator)
Raw .tav File
GenePix Converter
Query Window
Reports
Study
Experiment
Slidetype
Slide
Probe
MAGE-ML
MUSAGE
Data
MAD
Database
Others...
tabase

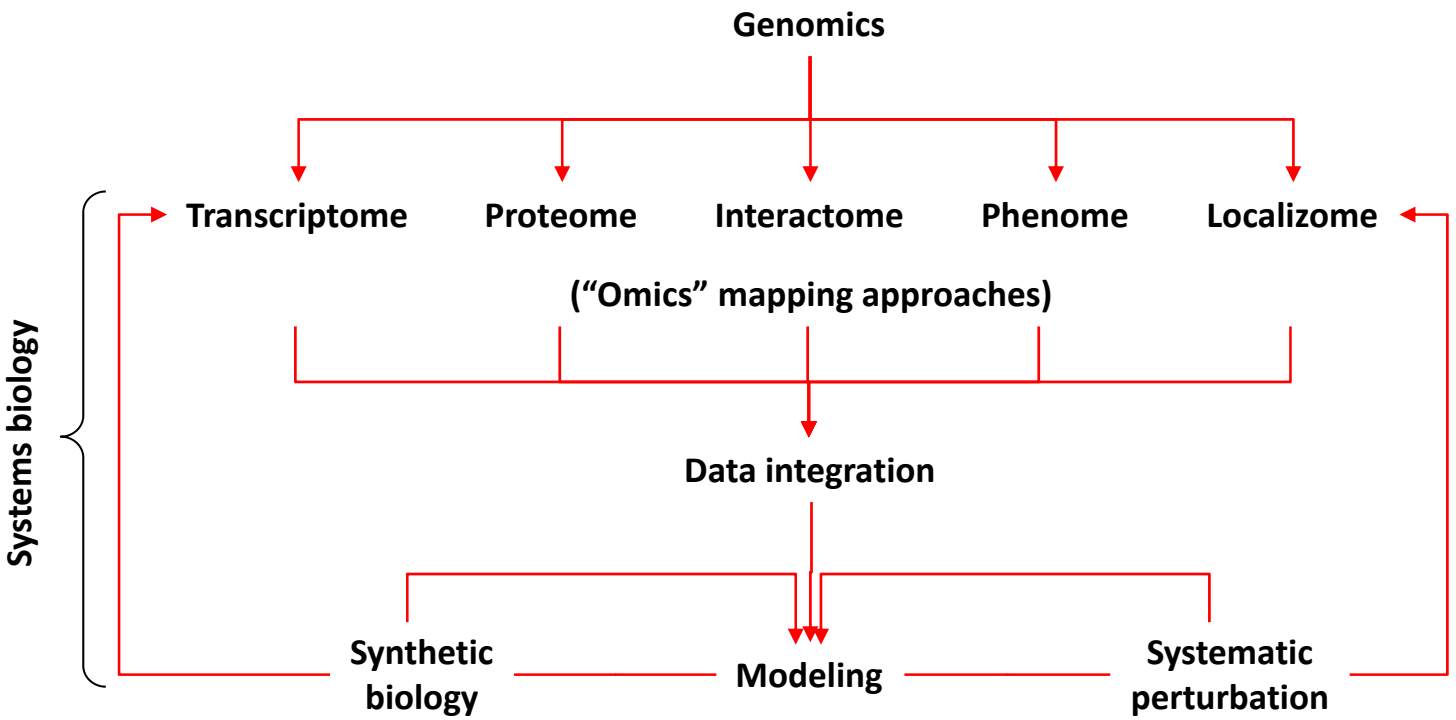
General scheme of the procedure used in tumour expression profiling for target identification and validation. RNA isolation from cell lines, tumour biopsy, and control samples is followed by labeling of the probe, hybridisation with the DNA microarray, data acquisition, and analysis. Verification of the results can be performed using different approaches, such as tissue microarray analysis.



The use of high-throughput technologies provides novel targets that might be developed into biomarkers of diagnostic and prognostic utility. The process of development of a tumour marker requires several steps before acceptance into clinical routine.

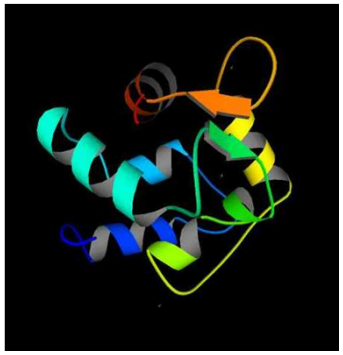


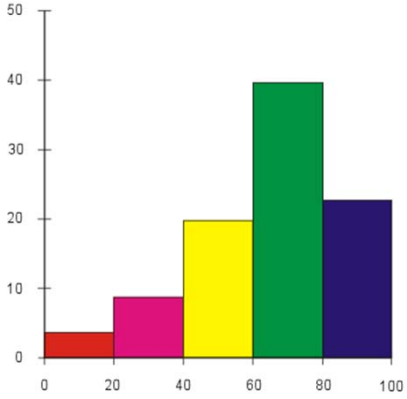
So how can we meaningfully integrate the data?

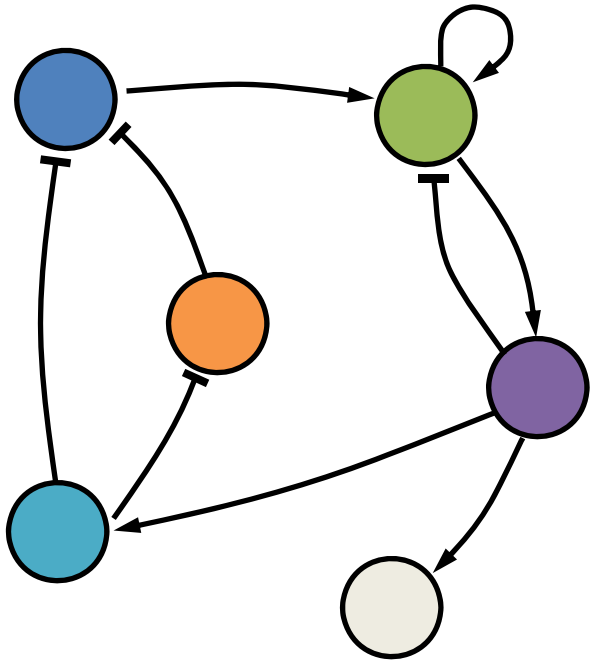


Comparing Genetic Interaction to Protein - Protein Interaction and Gene Regulatory Networks

- PPI (protein-protein interaction) = binding
- GRN (gene regulatory network) = expression
- SGA (synthetic gene array) = function
- Network of functional interactions
- SGA complements PPIs and GRNs

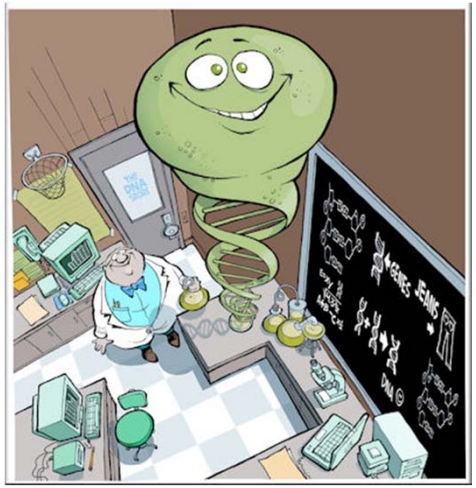






Two ways of looking a problem

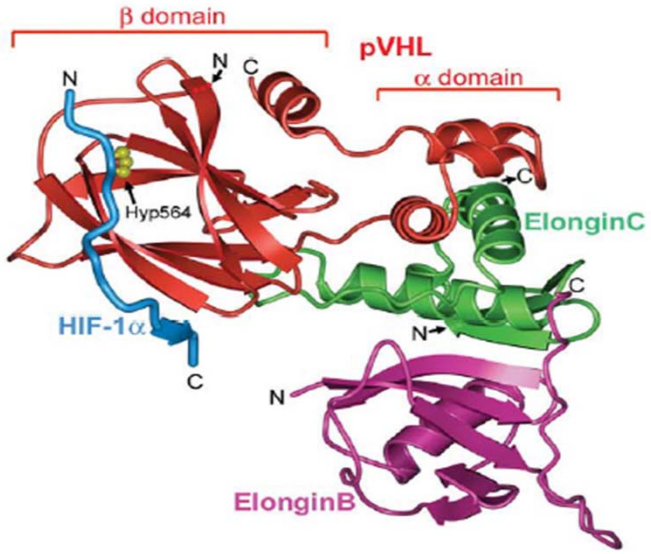
- *Top down* or *bottom up*
  - Either look at the whole organism and abstract large portions of it
  - Or try to understand each small piece and then after understanding every small piece assemble into the whole
  - Both are used, valid and complement each other



Interactions → Networks → Pathways

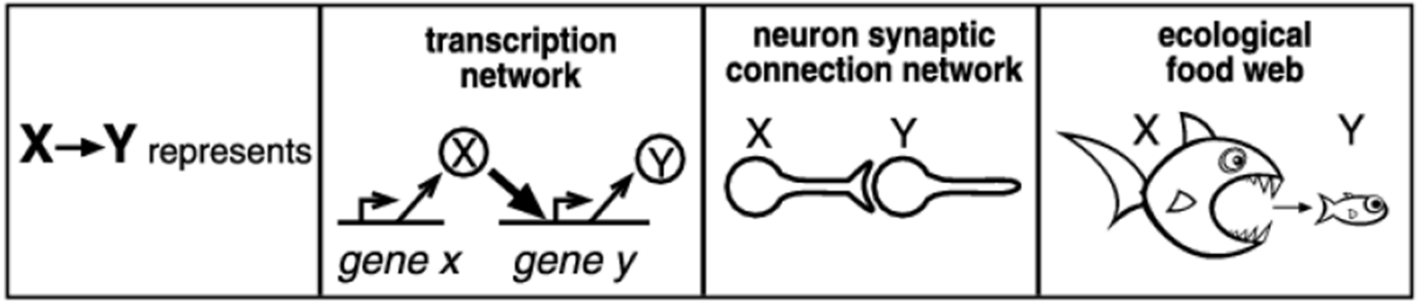
Why do we need protein interaction information

- Learning protein functions:
  - If two proteins interact, there is a very high possibility that their *functions* are related as well
- Cellular operations are largely endured by interactions among proteins
- From protein pathways to understanding cells, tissues, … to life and evolution



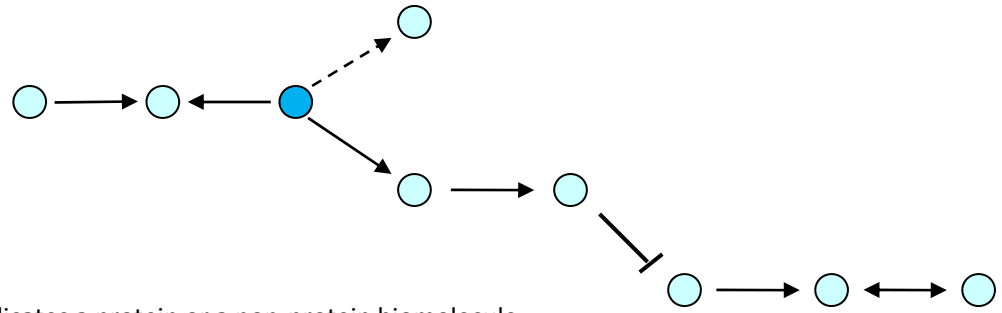
Interactions → **Networks** → Pathways

- A network is the *backbone* of a complex system.
- A group or system of (electric) components and *connecting circuitry* designed to function in a specific manner.
- Such motifs are found in networks from:
  - *Engineering* : Electronic circuits, World Wide Web *et al*.,
  - *Biochemistry* : Transcriptional regulation networks
  - *Neurobiology* : Neuron connectivity
  - *Ecology* : Food webs

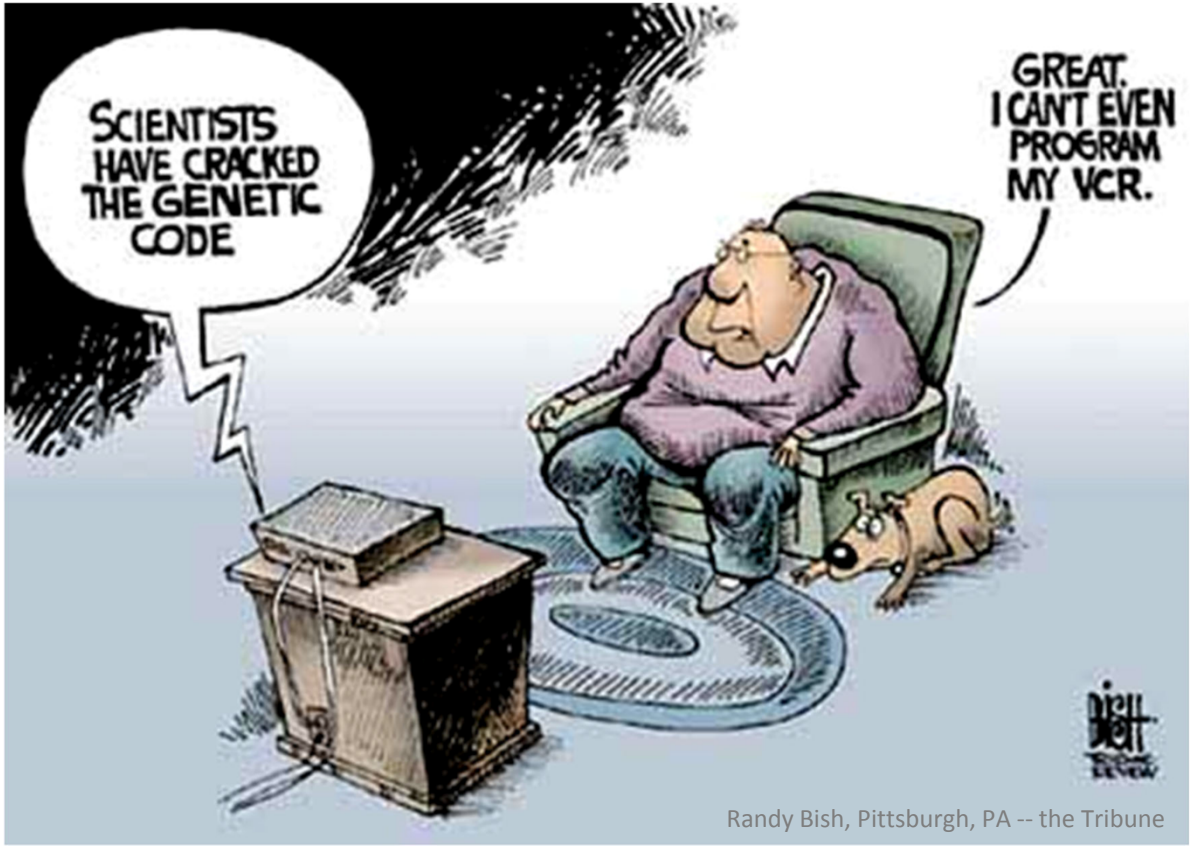


Interactions → Networks → **Pathways**

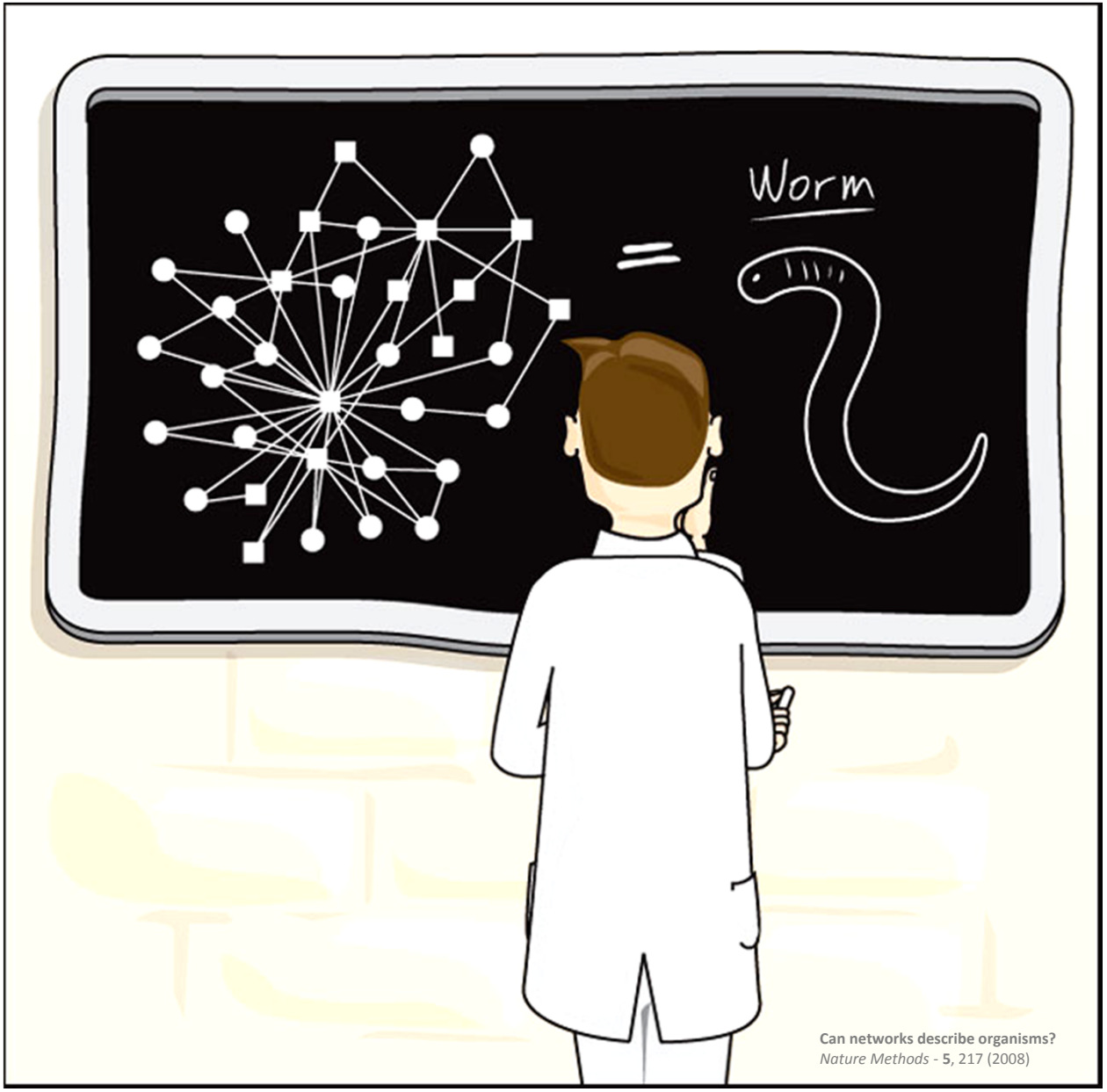
- It can be defined ad a modular unit of interacting molecules to fulfill a cellular function.
- It is usually represented by a *2-D diagram with characteristic symbols* linking the protein and non-protein entities.



A circle indicates a protein or a non-protein biomolecule.
An symbol in between indicates the nature of molecule-molecule interaction.



STORY UNFINISHED
TO BE CONTINUED … …
BUT NOT ME ^^…



Can networks describe organisms?
Nature Methods - **5**, 217 (2008)