Statistical Methods and Software for the Analysis of DNA Microarray Experiments

Sandrine Dudoit
Division of Biostatistics, University of California, Berkeley
Rafael Irizarry
Department of Biostatistics, Johns Hopkins University
www.bioconductor.org

ENAR Spring Meeting, Tampa, FL
March 30, 2003

Outline

- Introduction to the biology and technology of DNA microarrays
- Overview of the Bioconductor project
- Annotation
- Visualization
- Pre-processing: spotted and Affymetrix arrays
- Differential gene expression
- Software demo

Acknowledgments

Bioconductor core team

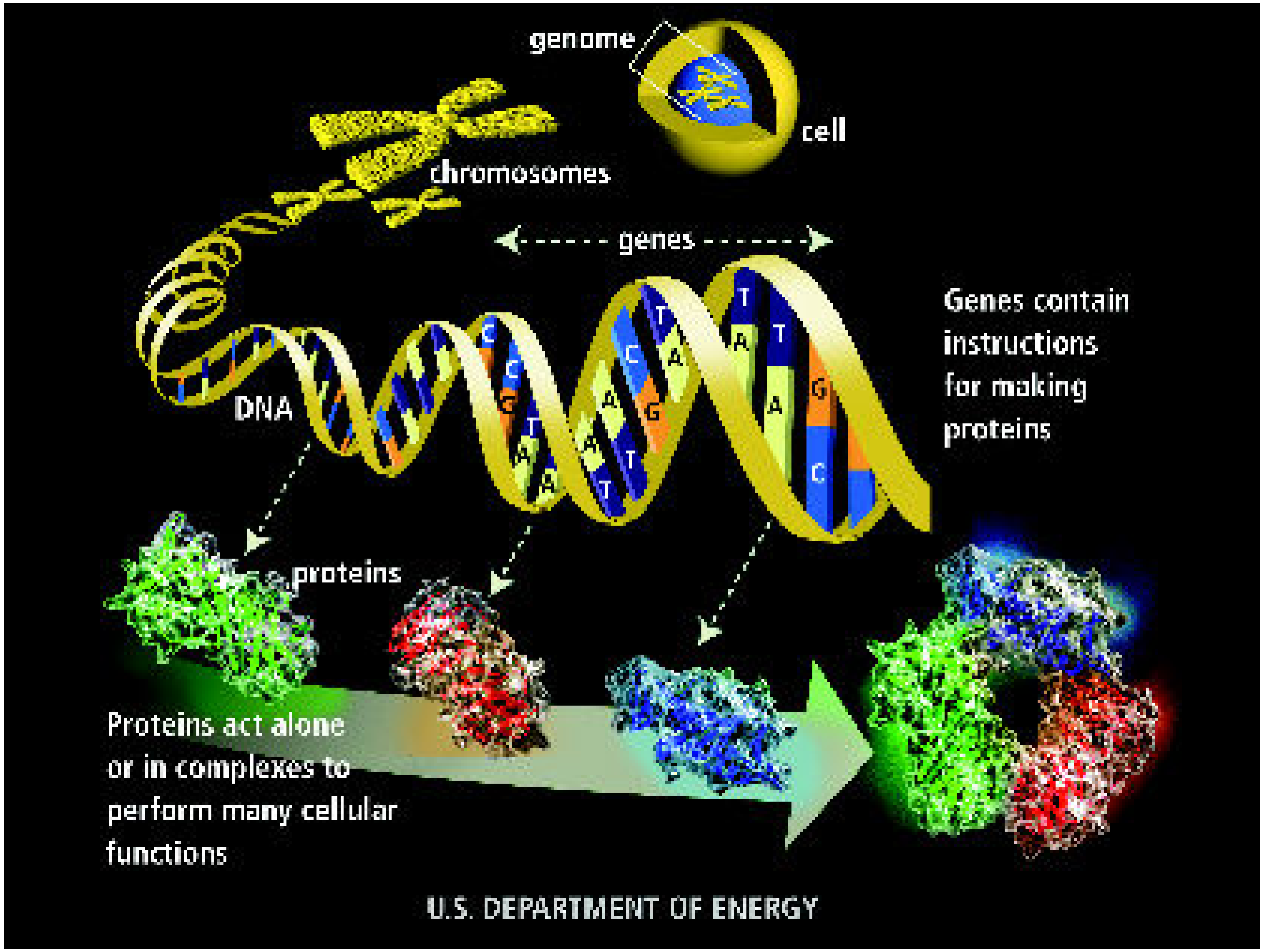
- **Ben Bolstad**, Biostatistics, UC Berkeley
- **Vince Carey**, Biostatistics, Harvard
- **Laurent Gautier**, Technical University of Denmark
- **Yongchao Ge**, Statistics, UC Berkeley
- **Robert Gentleman**, Biostatistics, Harvard
- **Jeff Gentry**, Dana-Farber Cancer Institute
- **Yee Hwa (Jean) Yang**, Biostatistics, UCSF
- **Jianhua (John) Zhang**, Dana-Farber Cancer Institute

References

- **Personal webpages**
  www.stat.berkeley.edu/~sandrine
  biosun01.biostat.jhsph.edu/~ririzarr

  articles and talks on: image analysis; normalization; identification of differentially expressed genes; cluster analysis; classification.
- **Bioconductor** www.bioconductor.org
  - software, data, and documentation (vignettes);
  - training materials from short courses;
  - mailing list.
- **R** www.r-project.org
  - software; documentation; RNews.

From chromosomes to proteins



Genes contain instructions for making proteins

Proteins act alone or in complexes to perform many cellular functions

U.S. DEPARTMENT OF ENERGY

www.ornl.gov/hgmis/graphics/slides/images1.html

Cells

- Cells: the fundamental working units of every living organism.

- Metazoa: multicellular organisms.
  E.g. humans: trillions of cells.
- Protozoa: unicellular organisms.
  E.g. yeast, bacteria.

Cells

- Each cell contains a complete copy of an organism’s genome, or blueprint for all cellular structures and activities.

- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.

The genome

- The human genome is distributed along 23 pairs of chromosomes
  - 22 autosomal pairs;
  - the sex chromosome pair, **XX** for females and **XY** for males.

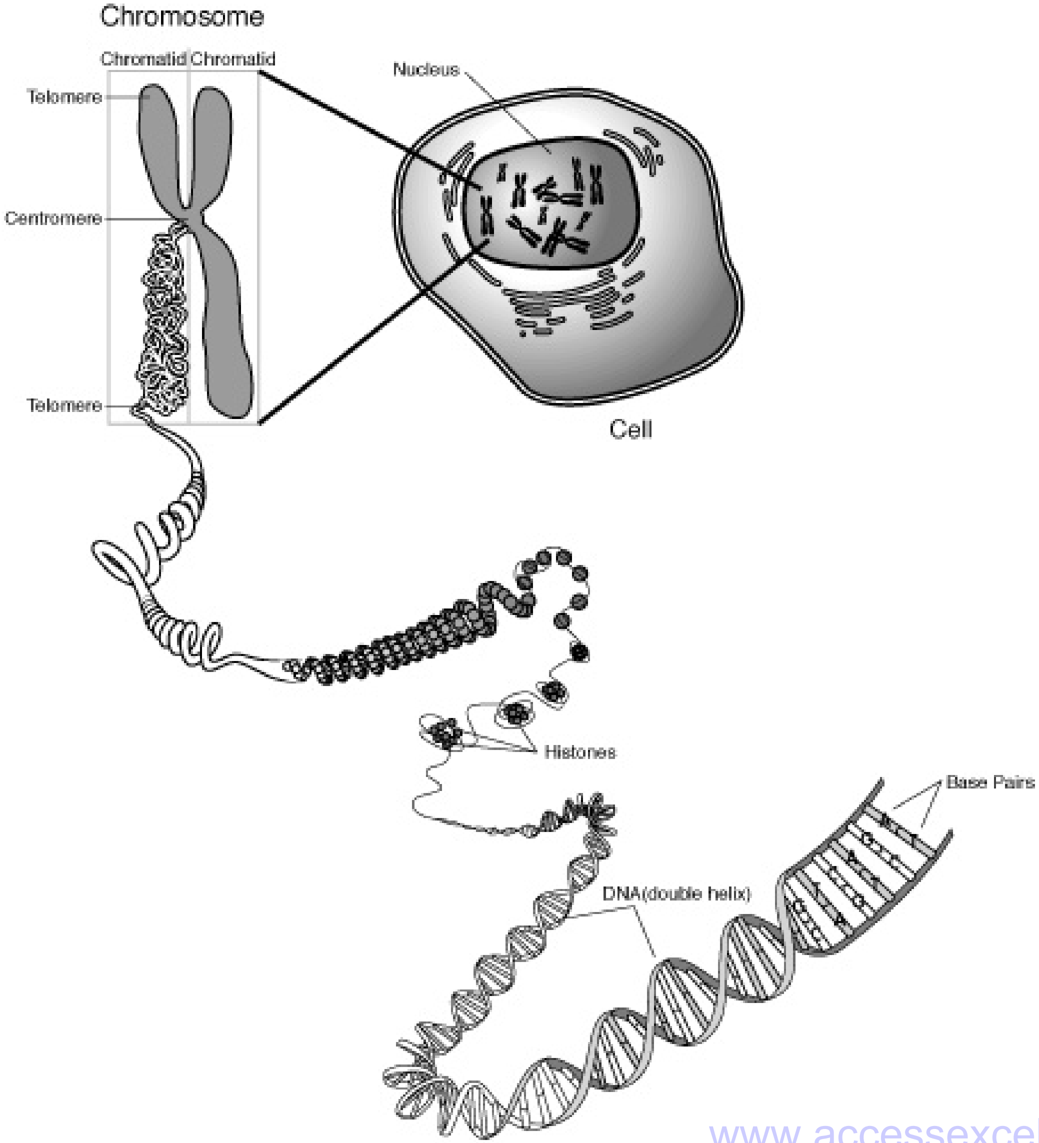
- In each pair, one chromosome is paternally inherited, the other maternally inherited (cf. meiosis).

The genome

- Chromosomes are made of compressed and entwined DNA.

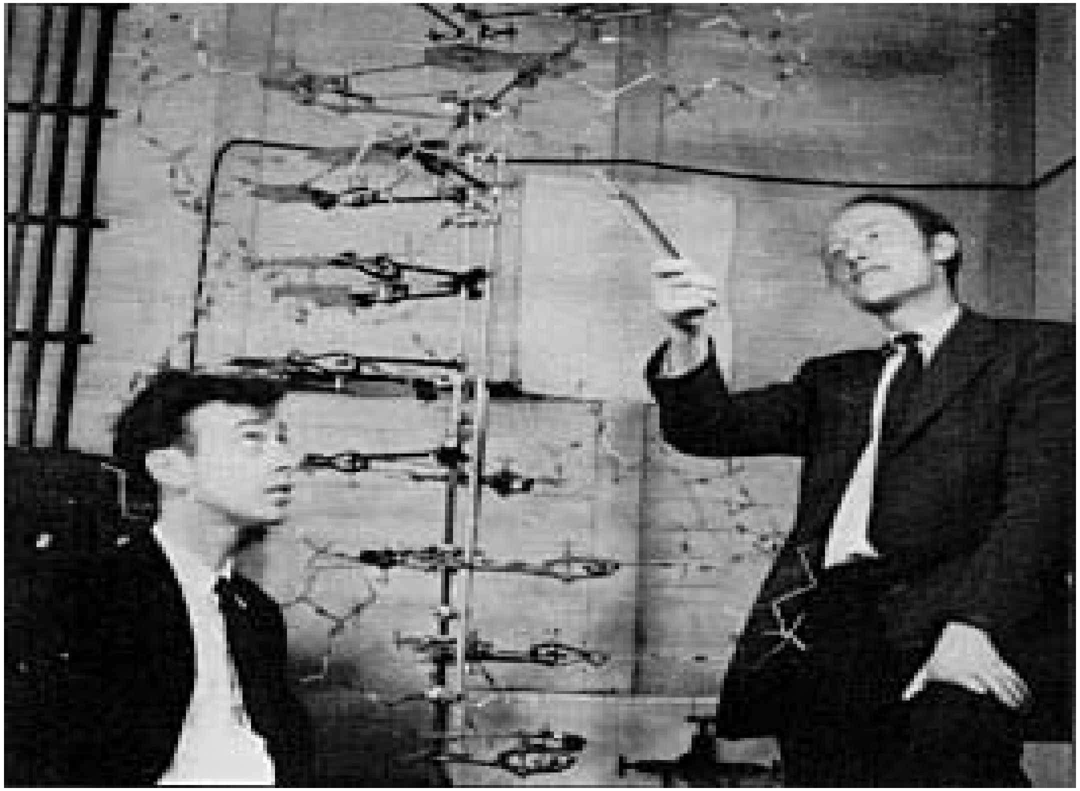
- A (protein-coding) gene is a segment of chromosomal DNA that directs the synthesis of a protein.

Chromosomes and DNA



www.accessexcellence.com/AB/GG/

DNA



"We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest."

J.D. Watson & F. H. C. Crick. (1953). Molecular structure of Nucleic Acids. *Nature.* **171**: 737-738.

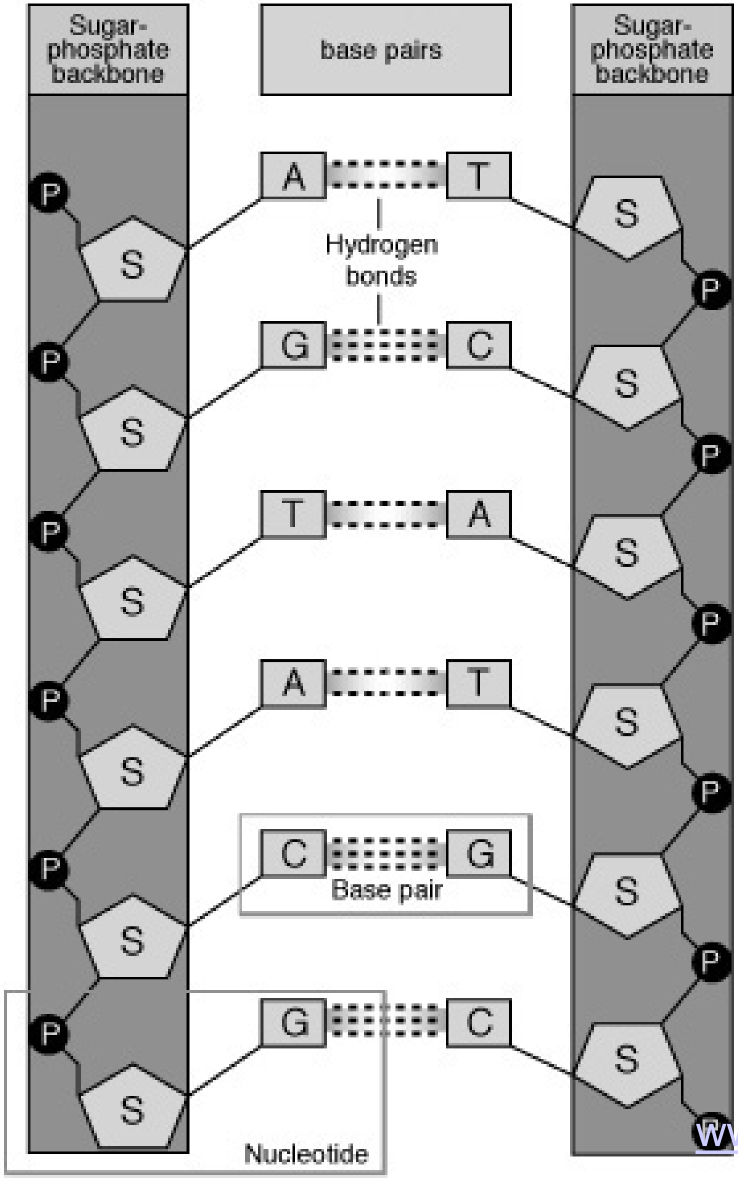
DNA

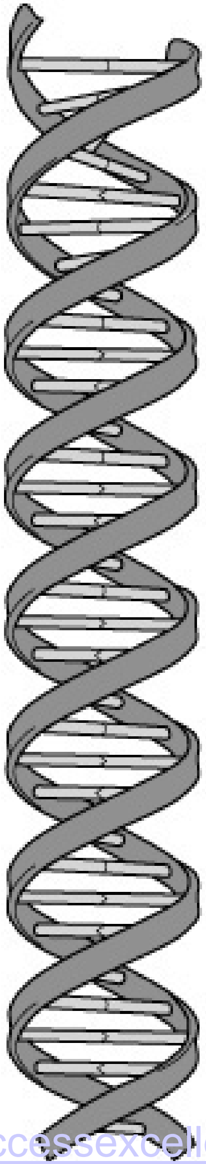
- A deoxyribonucleic acid or DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
- Each nucleotide comprises
  - a phosphate group;
  - a deoxyribose sugar;
  - one of four nitrogen bases:
    - purines: adenine (A) and guanine (G),
    - pyrimidines: cytosine (C) and thymine (T).

DNA

- Base-pairing occurs according to the following rule:
  - **C pairs with G**,
  - **A pairs with T**.
- The two chains are held together by hydrogen bonds between nitrogen bases.

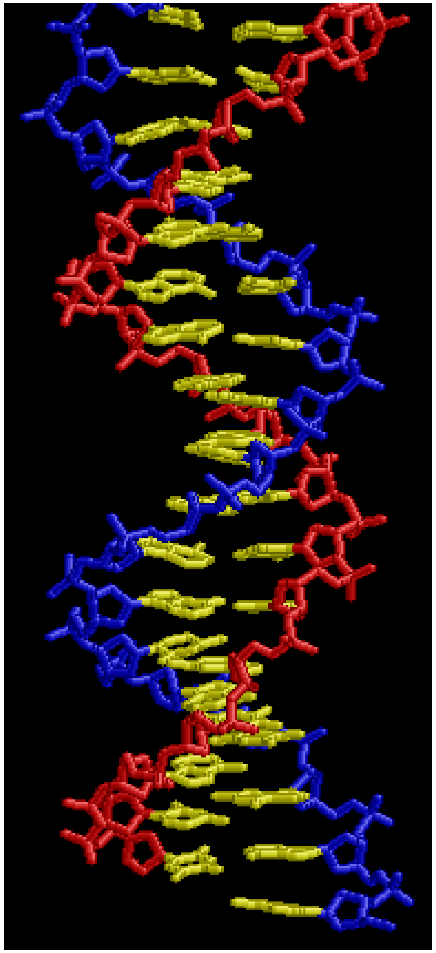
DNA

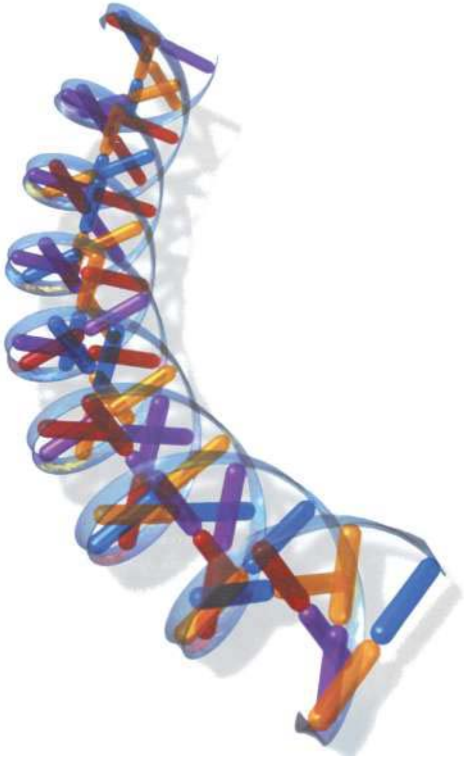


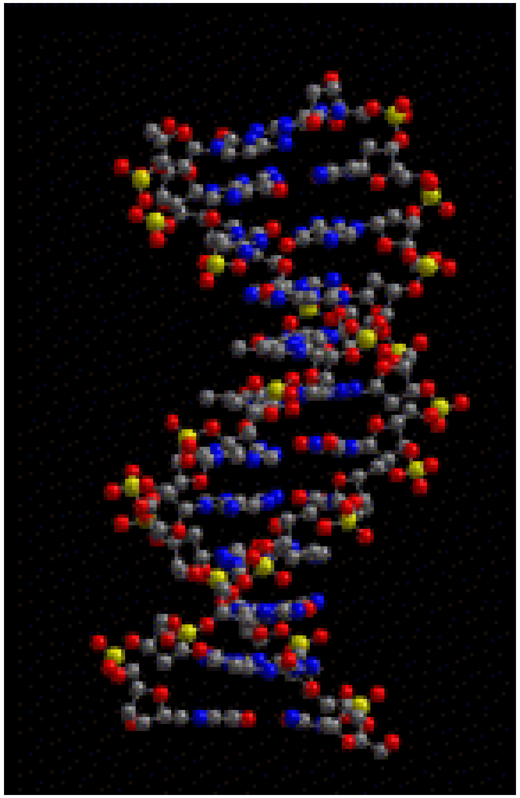


www.accessexcellence.com/AB/GG/

DNA

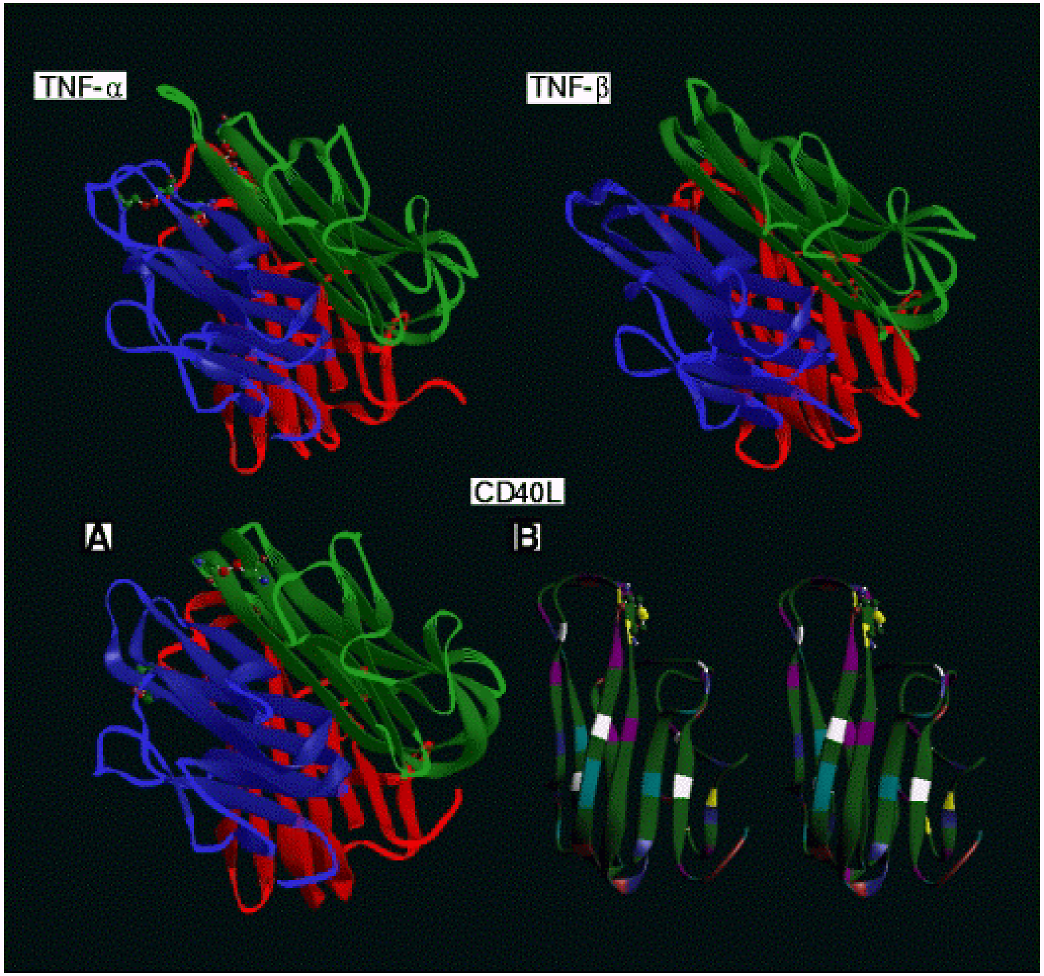






academy.d20.co.edu/kadets/lundberg/dnapic.html

Proteins

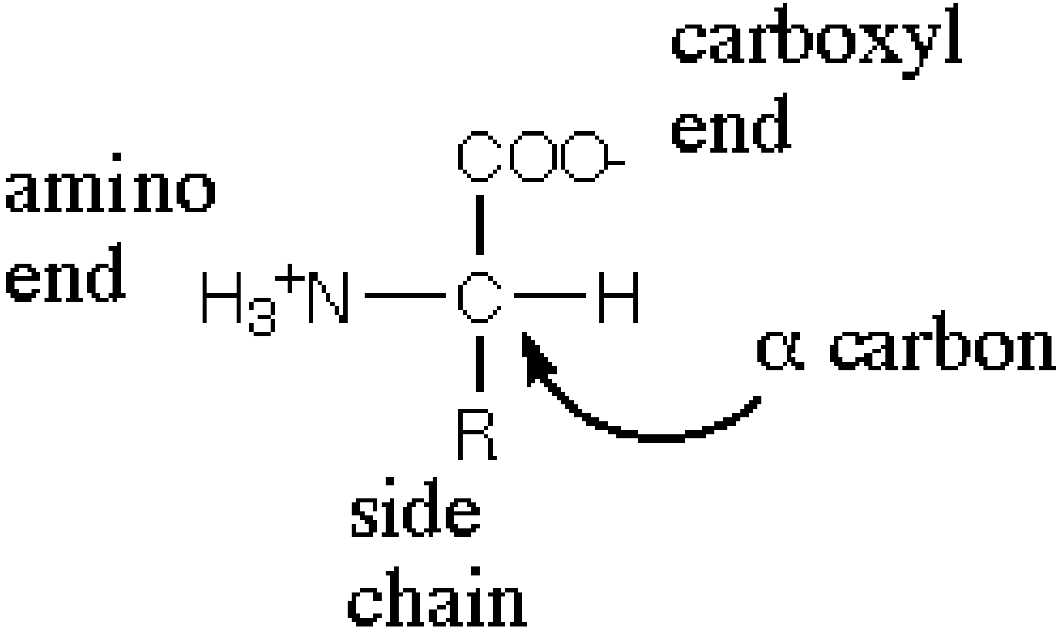


www.biochem.szote.u-szeged.hu/astrojan/protein1.htm

Proteins

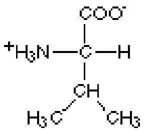
- Proteins: large molecules composed of one or more chains of amino acids, polypeptides.
- Amino acids: class of 20 different organic compounds containing a basic amino group ($-NH_2$) and an acidic carboxyl group (-COOH).
- The order of the amino acids is determined by the base sequence of nucleotides in the gene coding for the protein.
- E.g. hormones, enzymes, antibodies.

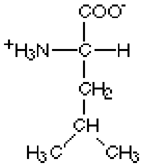
Amino acids

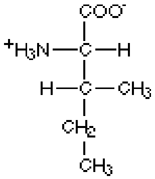


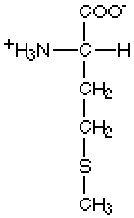
Amino acids

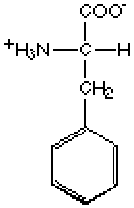
Amino acids with hydrophobic side groups


Valine
(val)

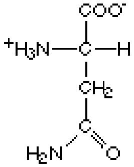

Leucine
(leu)

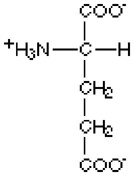

Isoleucine
(ile)

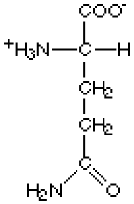

Methionine
(met)

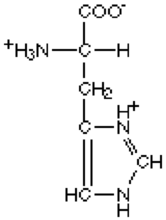

Phenylalanine
(phe)

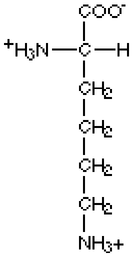
Amino acids with hydrophilic side groups

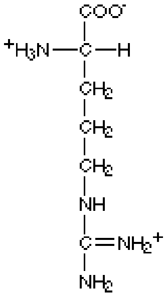

Asparagine
(asn)

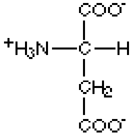

Glutamic acid
(glu)


Glutamine
(gln)

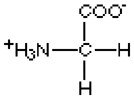

Histidine
(his)

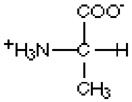

Lysine
(lys)

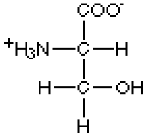

Arginine
(arg)

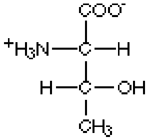

Aspartic acid
(asp)

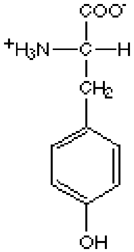
Amino acids that are in between

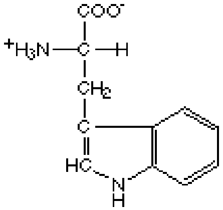

Glycine
(gly)

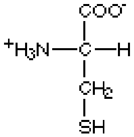

Alanine
(ala)

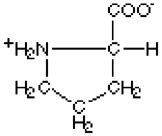

Serine
(ser)


Threonine
(thr)

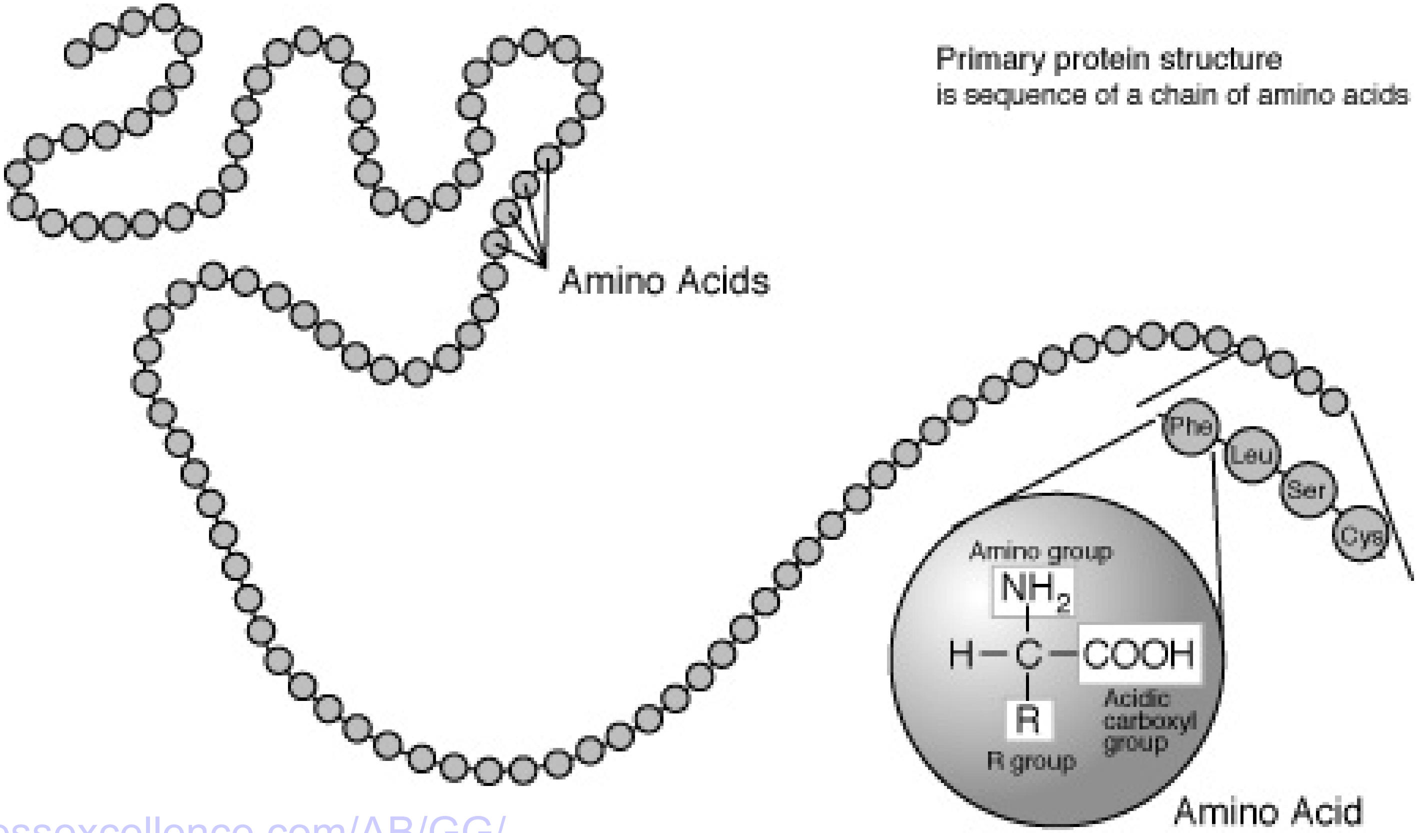

Tyrosine
(tyr)


Tryptophan
(trp)


Cysteine
(cys)


Proline
(pro)

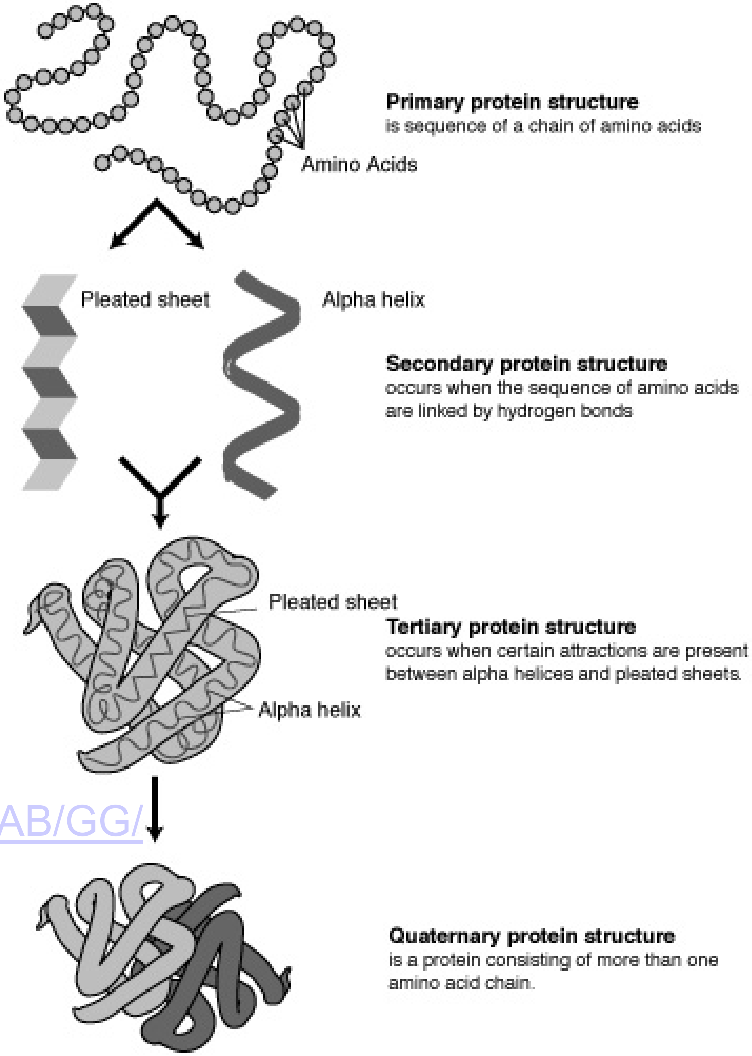
Proteins



Primary protein structure
is sequence of a chain of amino acids

www.accessexcellence.com/AB/GG/

Proteins



Primary protein structure
is sequence of a chain of amino acids

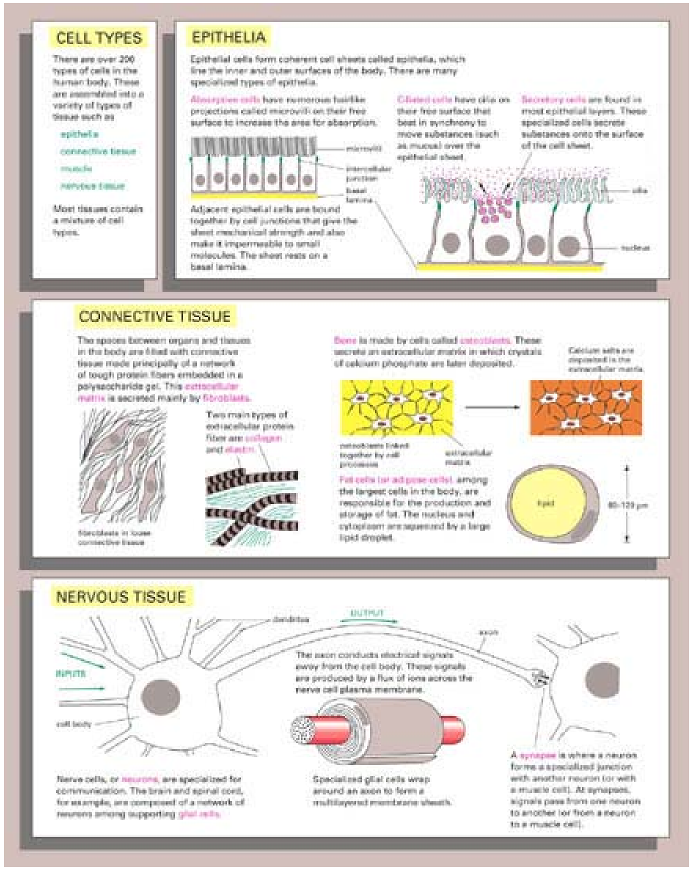
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

Quaternary protein structure
is a protein consisting of more than one amino acid chain.

www.accessexcellence.com/AB/GG/

Cell types



CELL TYPES

There are over 200 types of cells in the human body. These are assembled into a variety of types of tissue such as

- epithelia
- connective tissue
- muscle
- nervous tissue

Most tissues contain a mixture of cell types.

EPITHELIA

Epithelial cells form coherent cell sheets called epithelia, which line the inner and outer surfaces of the body. There are many specialized types of epithelia.

Absorptive cells have numerous hairlike projections called microvilli on their free surface to increase the area for absorption.

Ciliated cells have cilia on their free surface that beat in synchrony to move substances (such as mucus) over the epithelial sheet.

Secretory cells are found in most epithelial layers. These specialized cells secrete substances onto the surface of the cell sheet.

Adjacent epithelial cells are bound together by cell junctions that give the sheet mechanical strength and also make it impermeable to small molecules. The sheet rests on a basal lamina.

CONNECTIVE TISSUE

The spaces between organs and tissues in the body are filled with connective tissue made principally of a network of tough protein fibers embedded in a polysaccharide gel. This extracellular matrix is secreted mainly by fibroblasts.

Two main types of extracellular protein fiber are collagen and elastin.

fibroblasts in loose connective tissue

Bone is made by cells called osteoblasts. These secrete an extracellular matrix in which crystals of calcium phosphate are later deposited.

Calcium salts are deposited in the extracellular matrix.

Fat cells (or adipose cells), among the largest cells in the body, are responsible for the production and storage of fat. The nucleus and cytoplasm are squeezed by a large lipid droplet.

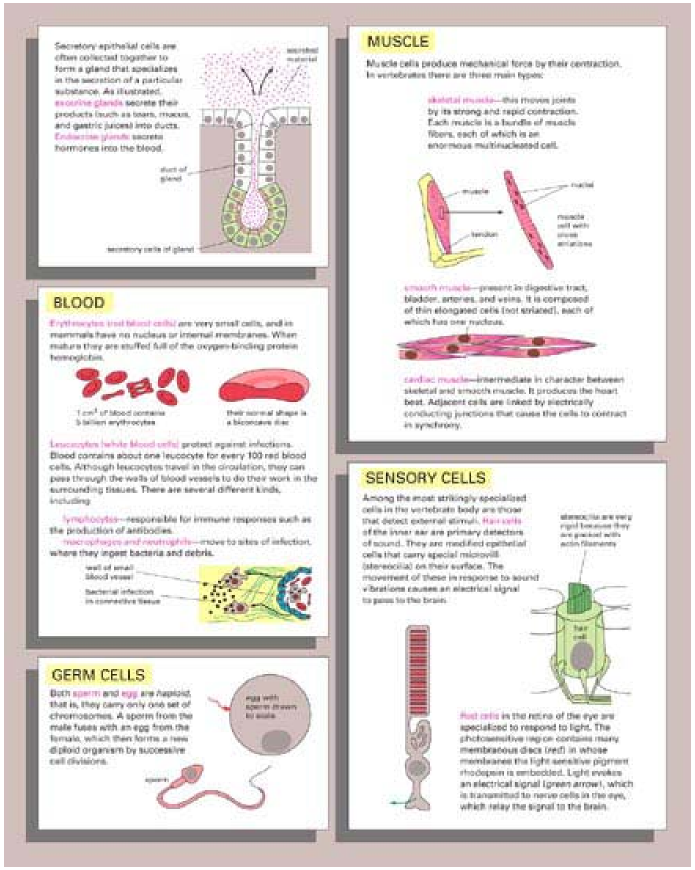
NERVOUS TISSUE

The axon conducts electrical signals away from the cell body. These signals are produced by a flux of ions across the nerve cell plasma membrane.

Nerve cells, or neurons, are specialized for communication. The brain and spinal cord, for example, are composed of a network of neurons among supporting glial cells.

Specialized glial cells wrap around an axon to form a multilayered membrane sheath.

A synapse is where a neuron forms a specialized junction with another neuron (or with a muscle cell). At synapses, signals pass from one neuron to another (or from a neuron to a muscle cell).



Secretory epithelial cells are often collected together to form a gland that specializes in the secretion of a particular substance. As illustrated, exocrine glands secrete their products (such as tears, mucus, and gastric juices) into ducts. Endocrine glands secrete hormones into the blood.

secreted material
duct of gland
secretory cells of gland

BLOOD

Erythrocytes (red blood cells) are very small cells, and in mammals have no nucleus or internal membranes. When mature they are stuffed full of the oxygen-binding protein hemoglobin.

$1\ cm^3$ of blood contains 5 billion erythrocytes

their normal shape is a biconcave disc

Leucocytes (white blood cells) protect against infections. Blood contains about one leucocyte for every 100 red blood cells. Although leucocytes travel in the circulation, they can pass through the walls of blood vessels to do their work in the surrounding tissues. There are several different kinds, including

lymphocytes—responsible for immune responses such as the production of antibodies.

macrophages and neutrophils—move to sites of infection, where they ingest bacteria and debris.

wall of small blood vessel
bacterial infection in connective tissue

GERM CELLS

Both sperm and egg are haploid, that is, they carry only one set of chromosomes. A sperm from the male fuses with an egg from the female, which then forms a new diploid organism by successive cell divisions.

MUSCLE

Muscle cells produce mechanical force by their contraction. In vertebrates there are three main types:

skeletal muscle—this moves joints by its strong and rapid contraction. Each muscle is a bundle of muscle fibers, each of which is an enormous multinucleated cell.

smooth muscle—present in digestive tract, bladder, arteries, and veins. It is composed of thin elongated cells (not striated), each of which has one nucleus.

cardiac muscle—intermediate in character between skeletal and smooth muscle. It produces the heart beat. Adjacent cells are linked by electrically conducting junctions that cause the cells to contract in synchrony.

SENSORY CELLS

Among the most strikingly specialized cells in the vertebrate body are those that detect external stimuli. Hair cells of the inner ear are primary detectors of sound. They are modified epithelial cells that carry special microvilli (stereocilia) on their surface. The movement of these in response to sound vibrations causes an electrical signal to pass to the brain.

stereocilia are very rigid because they are packed with actin filaments

Rod cells in the retina of the eye are specialized to respond to light. The photosensitive region contains many membranous discs (red) in whose membranes the light-sensitive pigment rhodopsin is embedded. Light evokes an electrical signal (green arrow), which is transmitted to nerve cells in the eye, which relay the signal to the brain.

www.accessexcellence.com/AB/GG/

Differential expression

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states E.g. blood, nerve, and skin cells, dividing cells, cancerous cells, etc.
- What makes the cells different?
- Differential gene expression, i.e., when, where, and how much each gene is expressed.
- On average, 40% of our genes are expressed at any given time.

Central dogma

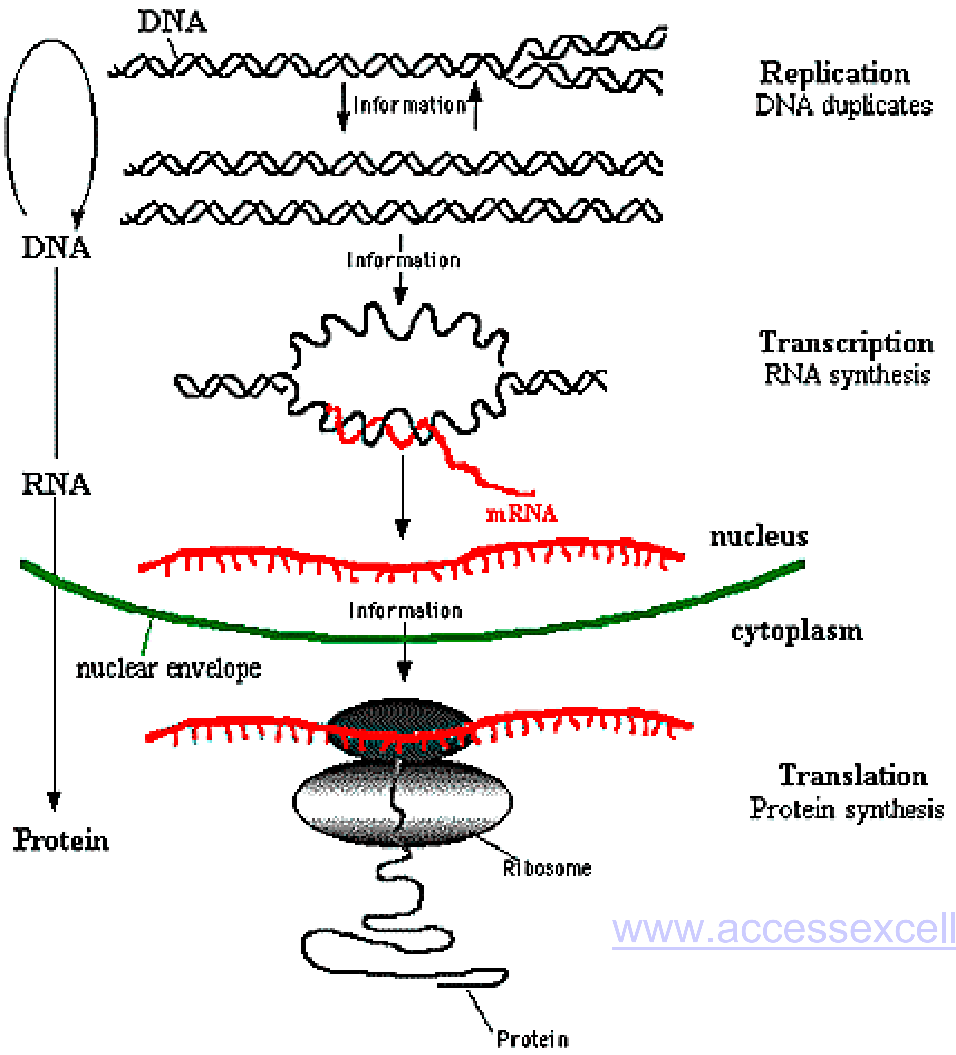
The expression of the genetic information stored in the DNA molecule occurs in two stages:

- (i) transcription, during which DNA is transcribed into mRNA;
- (ii) translation, during which mRNA is translated to produce a protein.

DNA → mRNA → protein

Other important aspects of regulation: methylation, alternative splicing, etc.

Central dogma



www.accessexcellence.com/AB/GG/

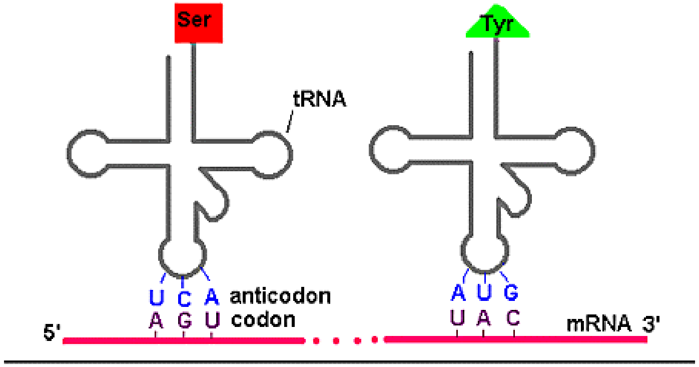
RNA

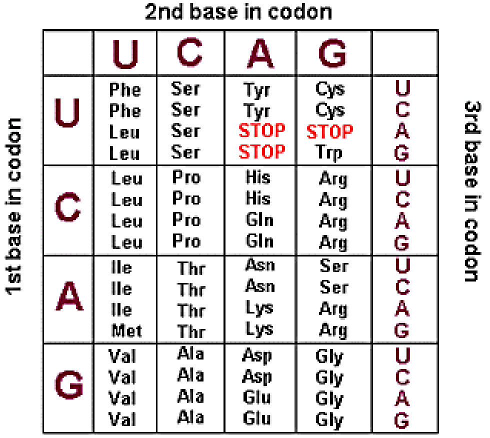
- A ribonucleic acid or RNA molecule is a nucleic acid similar to DNA, but
  - single-stranded;
  - ribose sugar rather than deoxyribose sugar;
  - uracil (U) replaces thymine (T) as one of the bases.
- RNA plays an important role in protein synthesis and other chemical activities of the cell.
- Several classes of RNA molecules, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs.

The genetic code

- DNA: sequence of four different nucleotides.
- Proteins: sequence of twenty different amino acids.
- The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the genetic code, which relates nucleotide triplets or codons to amino acids.

The genetic code





2nd base in codon

1st base in codon

3rd base in codon

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

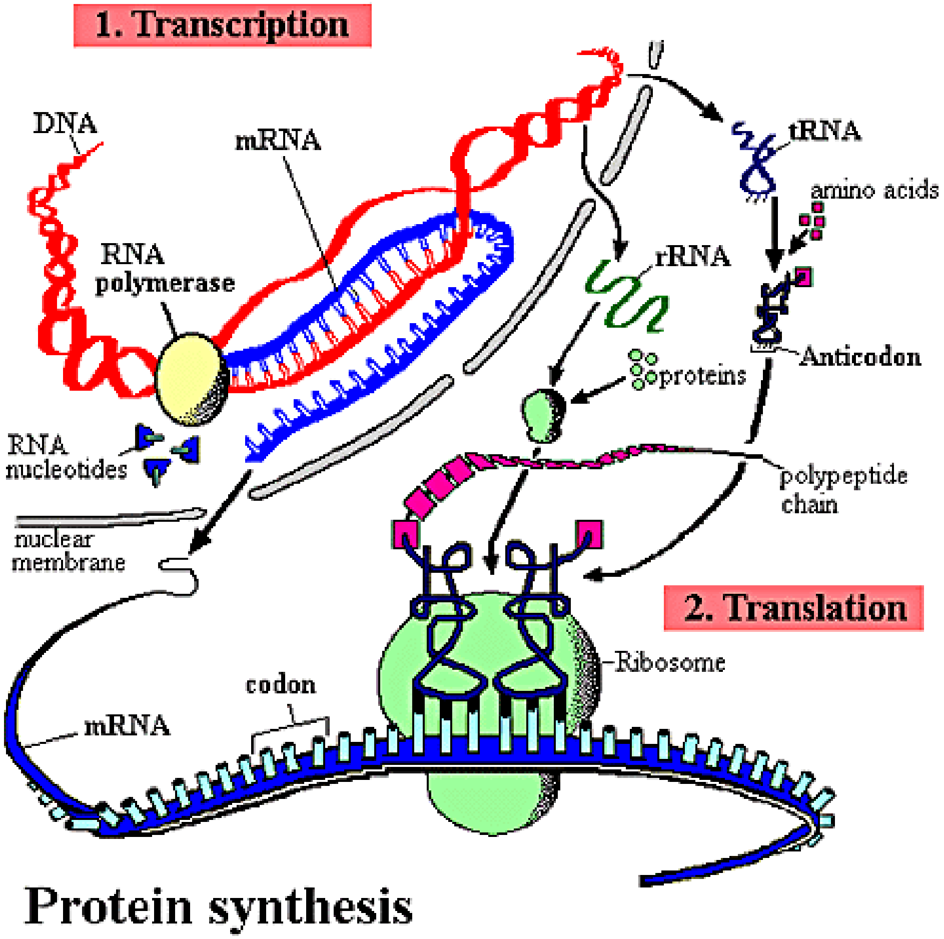
The Genetic Code

www.accessexcellence.com/AB/GG/

Start codon: initiation of translation (AUG, Met).
Stop codons: termination of translation.

Mapping between codons and amino acids is **many-to-one**: 64 codons but only 20 a.a.. Third base in codon is often redundant,
e.g., stop codons.

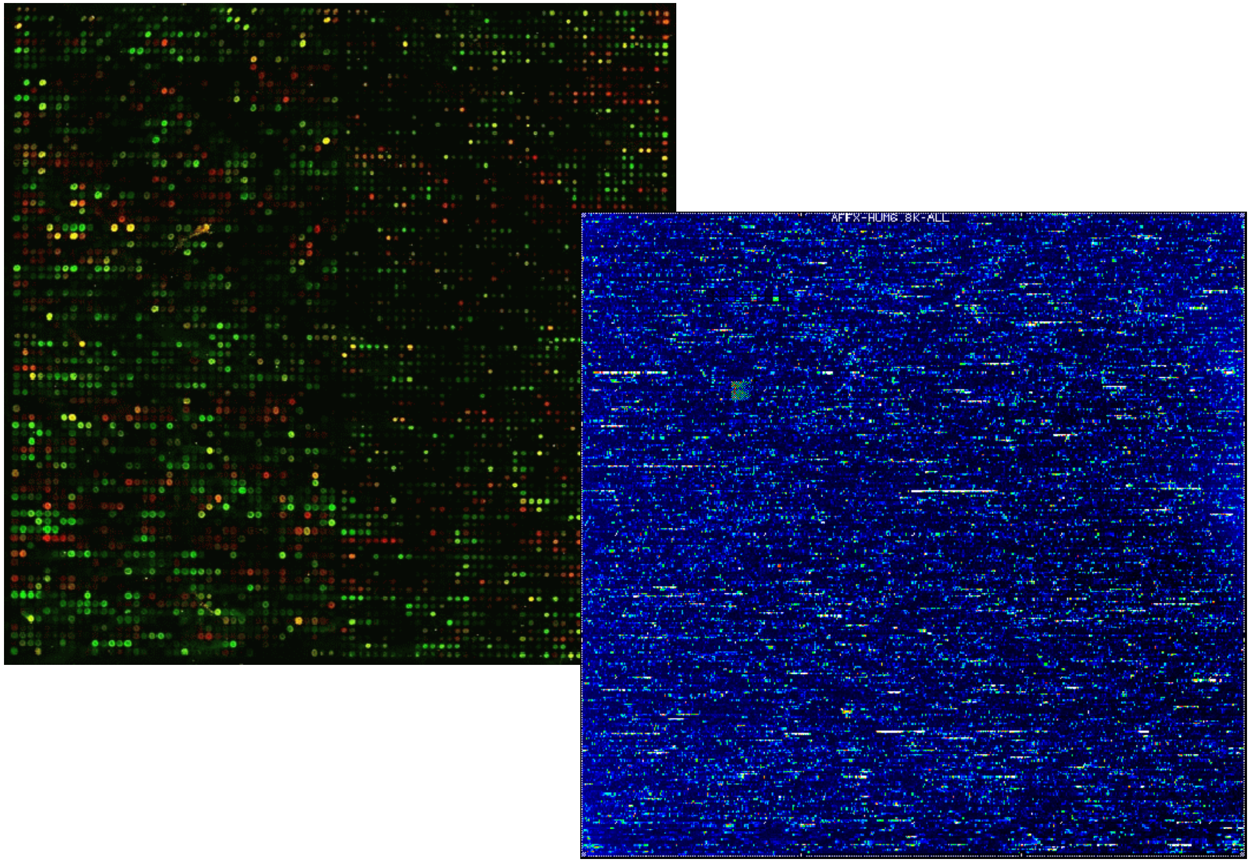
Protein synthesis



Functional genomics

- The various genome projects have yielded the complete DNA sequences of many organisms.
  - E.g. human, mouse, yeast, fruitfly, etc.
  - Human: 3 billion base-pairs, 30-40 thousand genes.
- Challenge: **go from sequence to function**, i.e., define the role of each gene and understand how the genome functions as a whole.

DNA microarrays



DNA microarrays

- Basic principles

- Spotted DNA microarrays

- Affymetrix oligonucleotide chips

DNA microarrays

- DNA microarray experiments are high-throughput biological assays for measuring the abundance of DNA or RNA sequences in different types of cell samples for thousands of sequences simultaneously.
- DNA microarray experiments exploit the availability of sequence data to get information on gene expression in different types of cells.

DNA microarrays

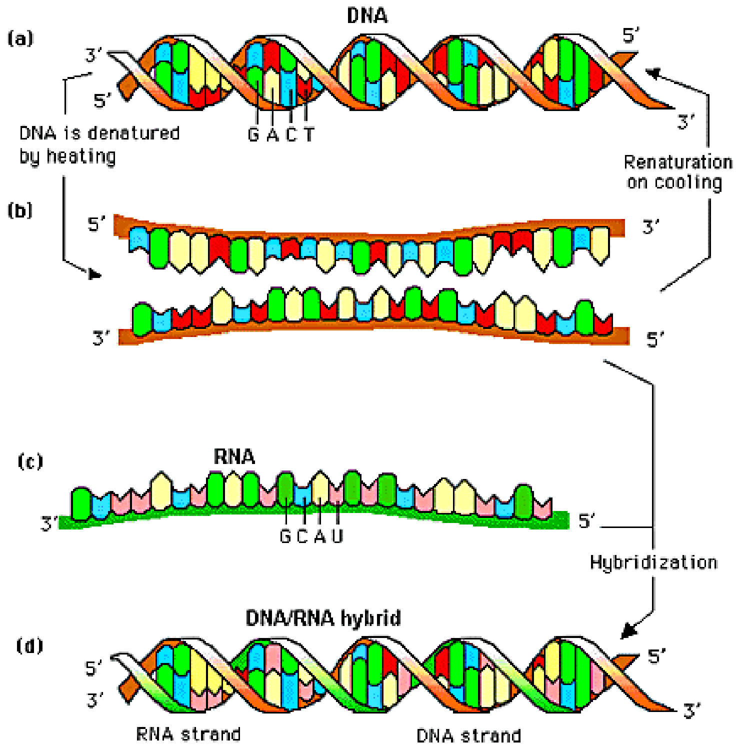
- DNA microarrays rely on the hybridization properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.
- The ancestor of cDNA microarrays: the Northern blot.

Hybridization

- Hybridization refers to the annealing of two nucleic acid strands following the base-pairing rules.

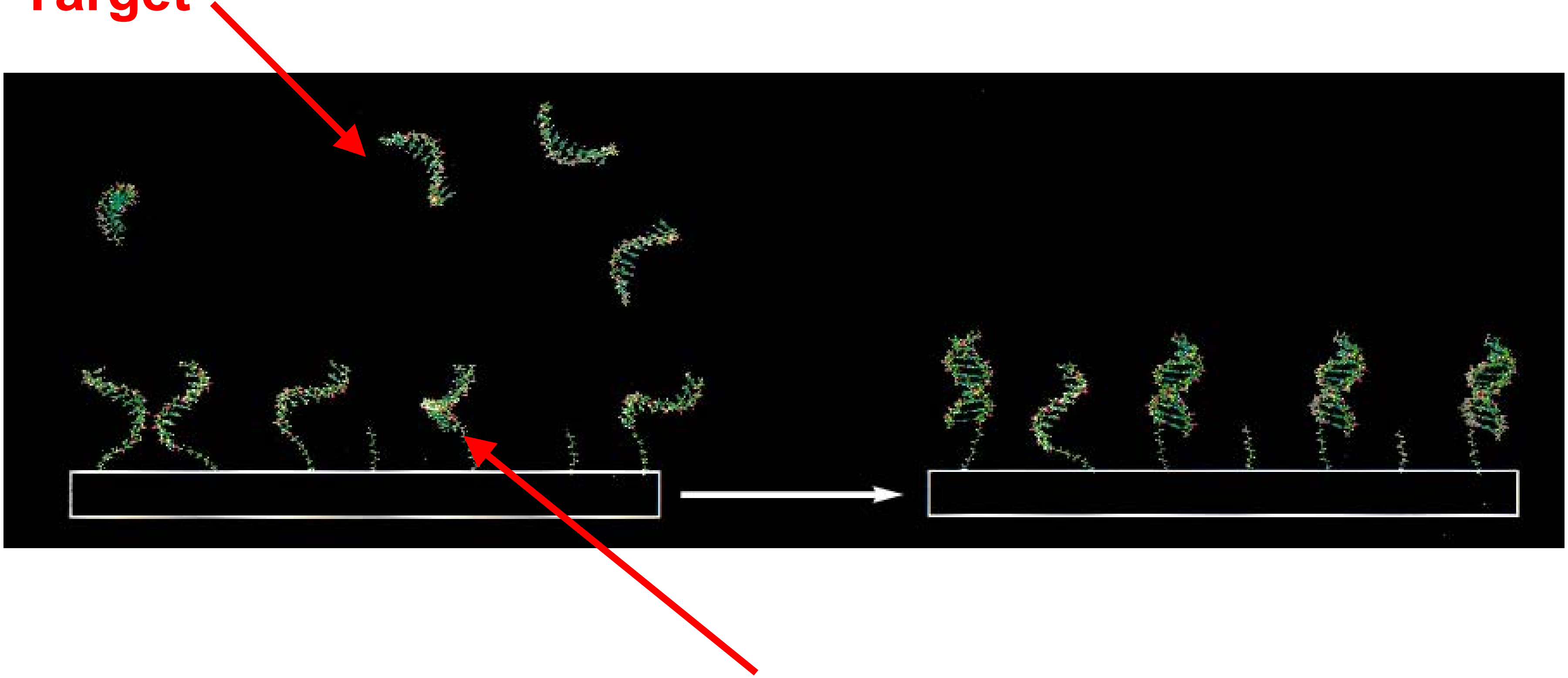
- Nucleic acid strands in a duplex can be separated, or denatured, by heating to destroy the hydrogen bonds.

Hybridization



Nucleic Acid Hybridization

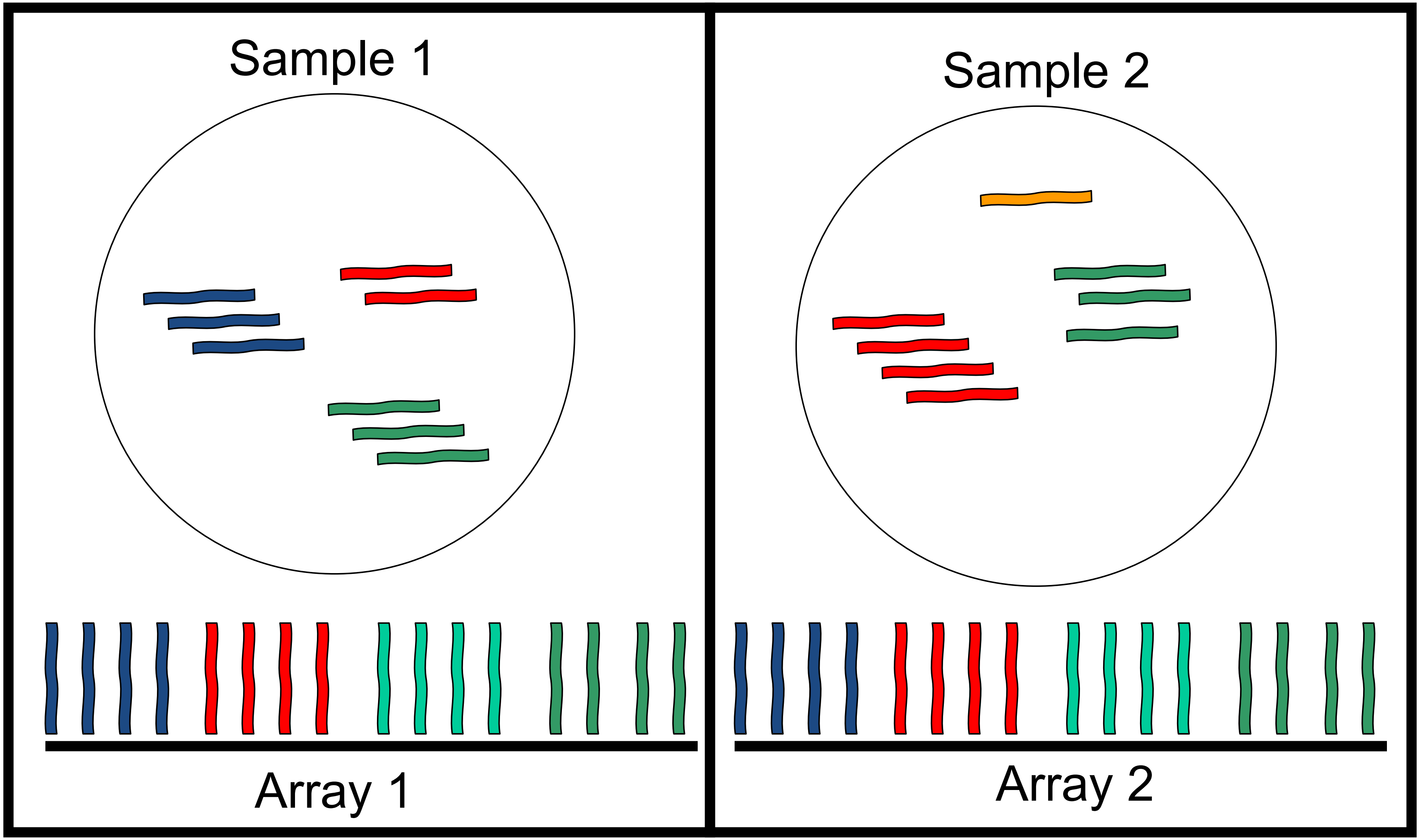
DNA microarrays



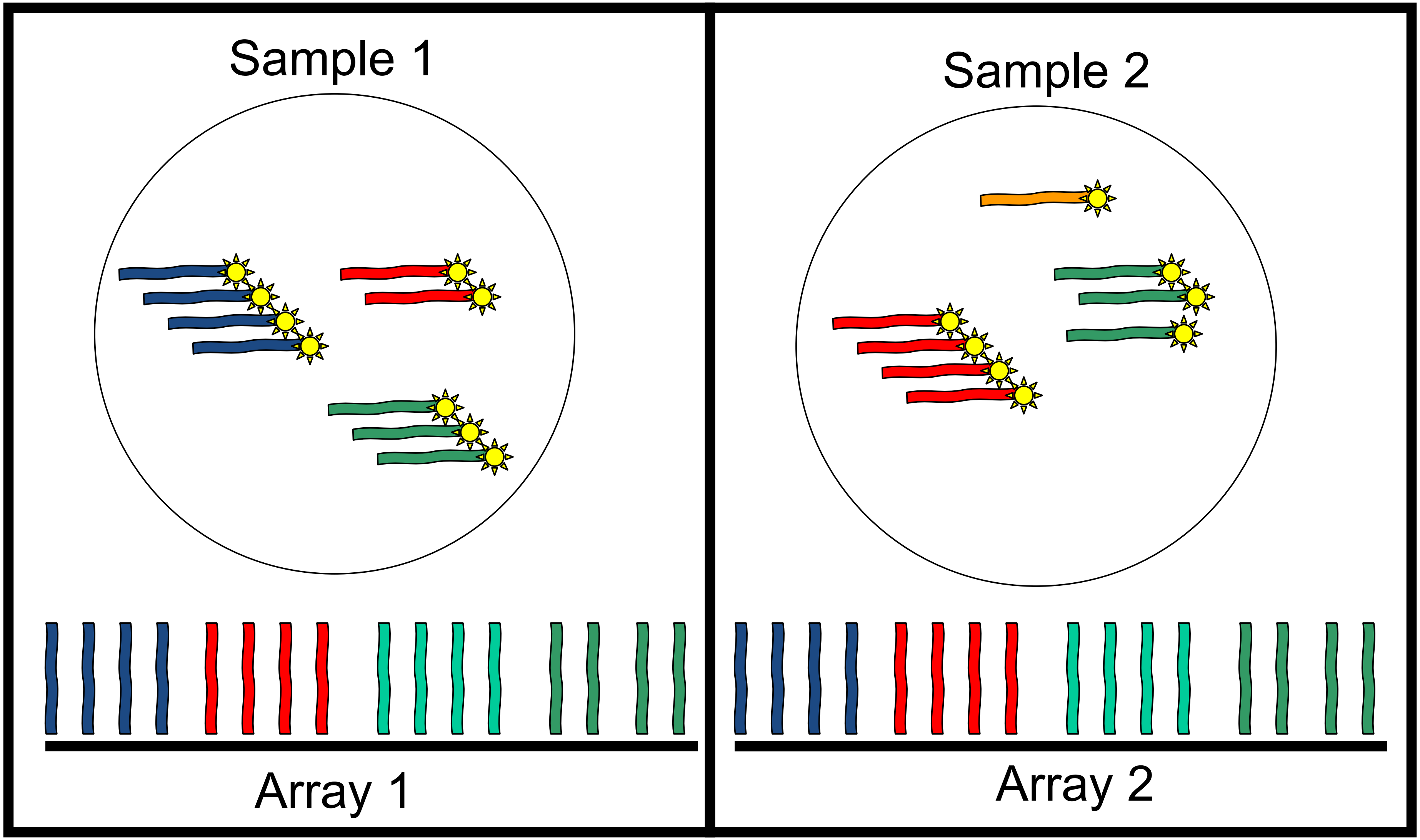
DNA microarrays

- The extent of hybridization of DNA sequences in the target sample to probe sequences on the array reflects the abundance of the probe sequences in the target sample.
- To quantify the extent of hybridization, the target sequences are fluorescently labeled.
- The hybridized arrays are scanned and the measured fluorescence intensities are used as measures of DNA/RNA abundance.

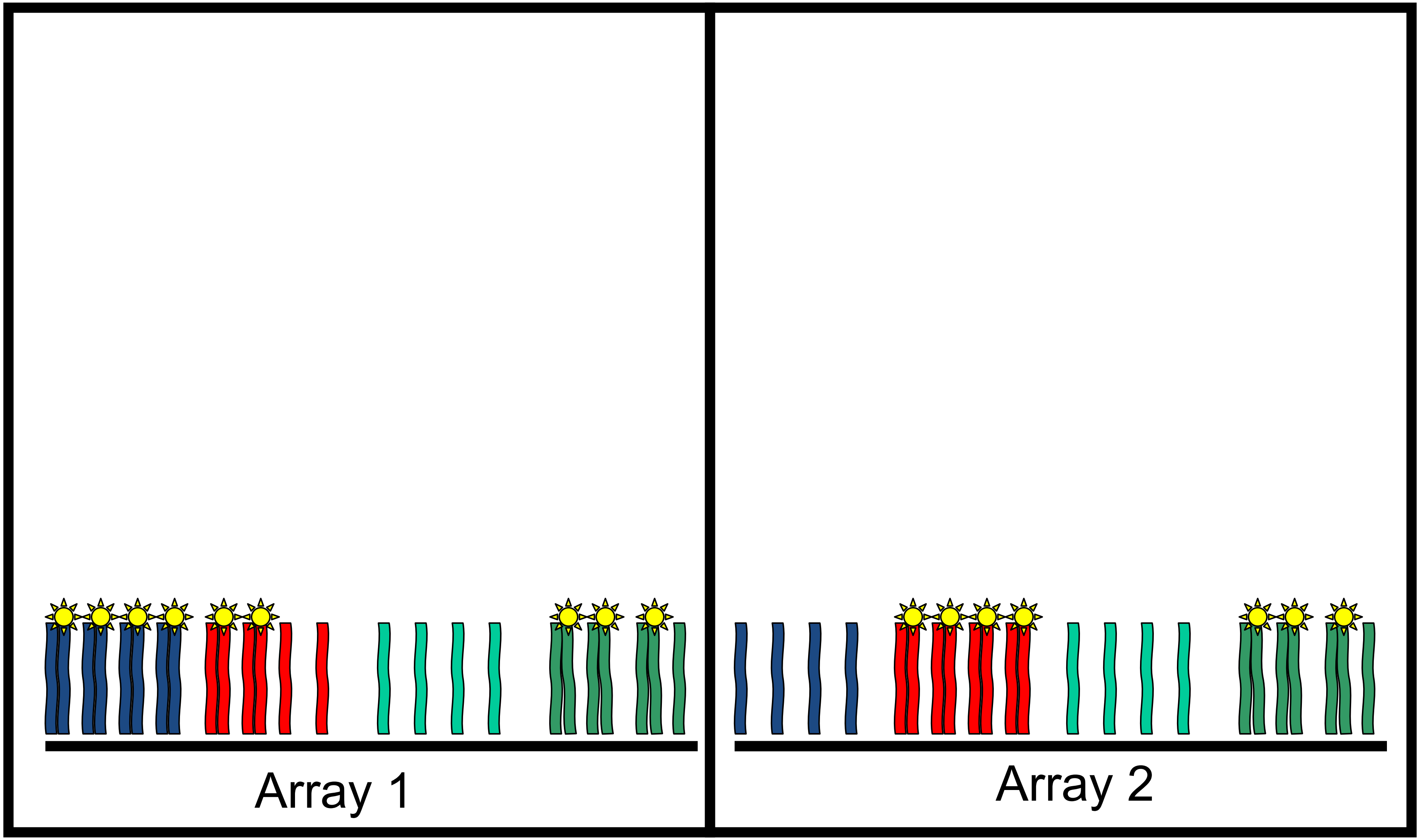
Before labeling



Before hybridization



After hybridization



Scanner image

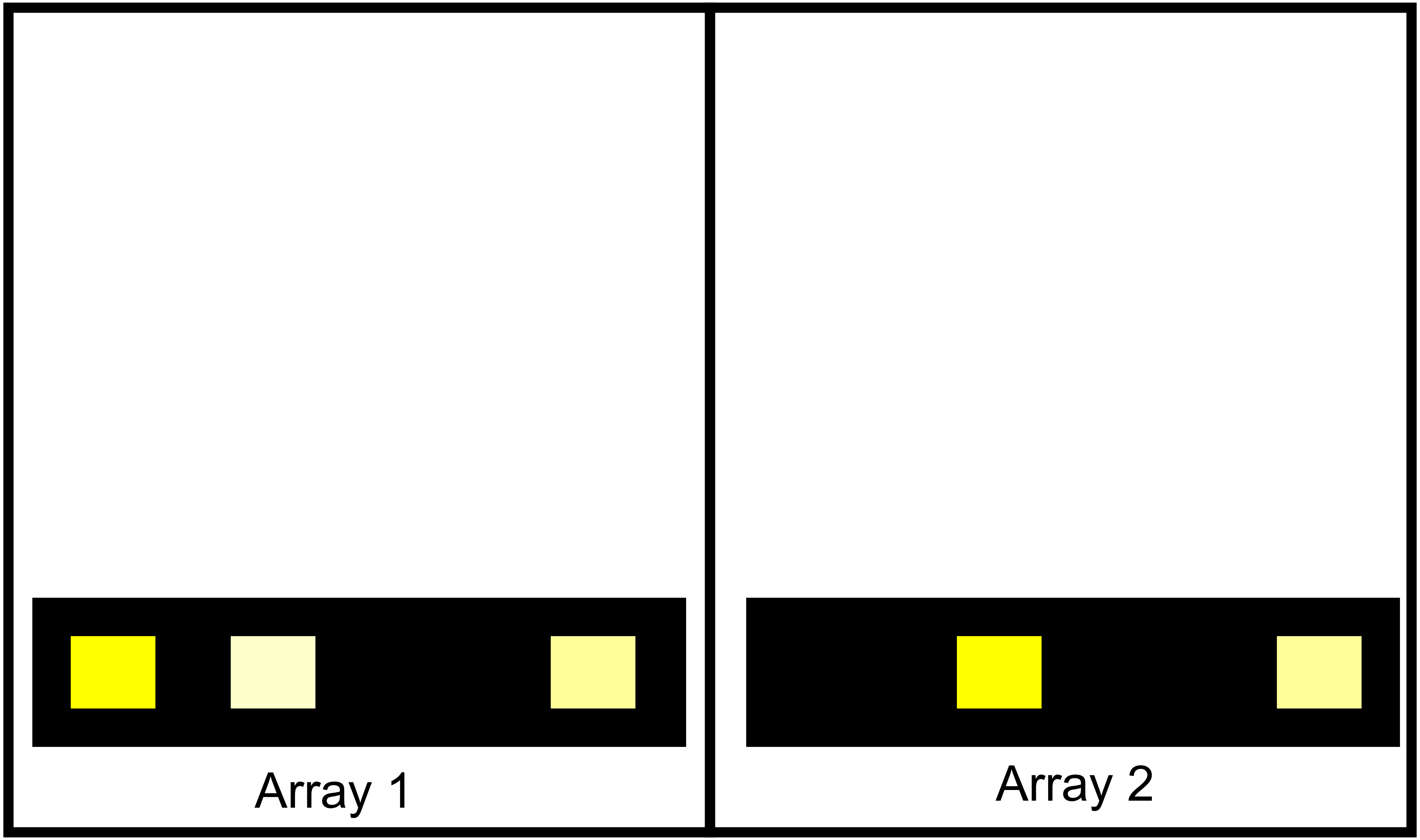
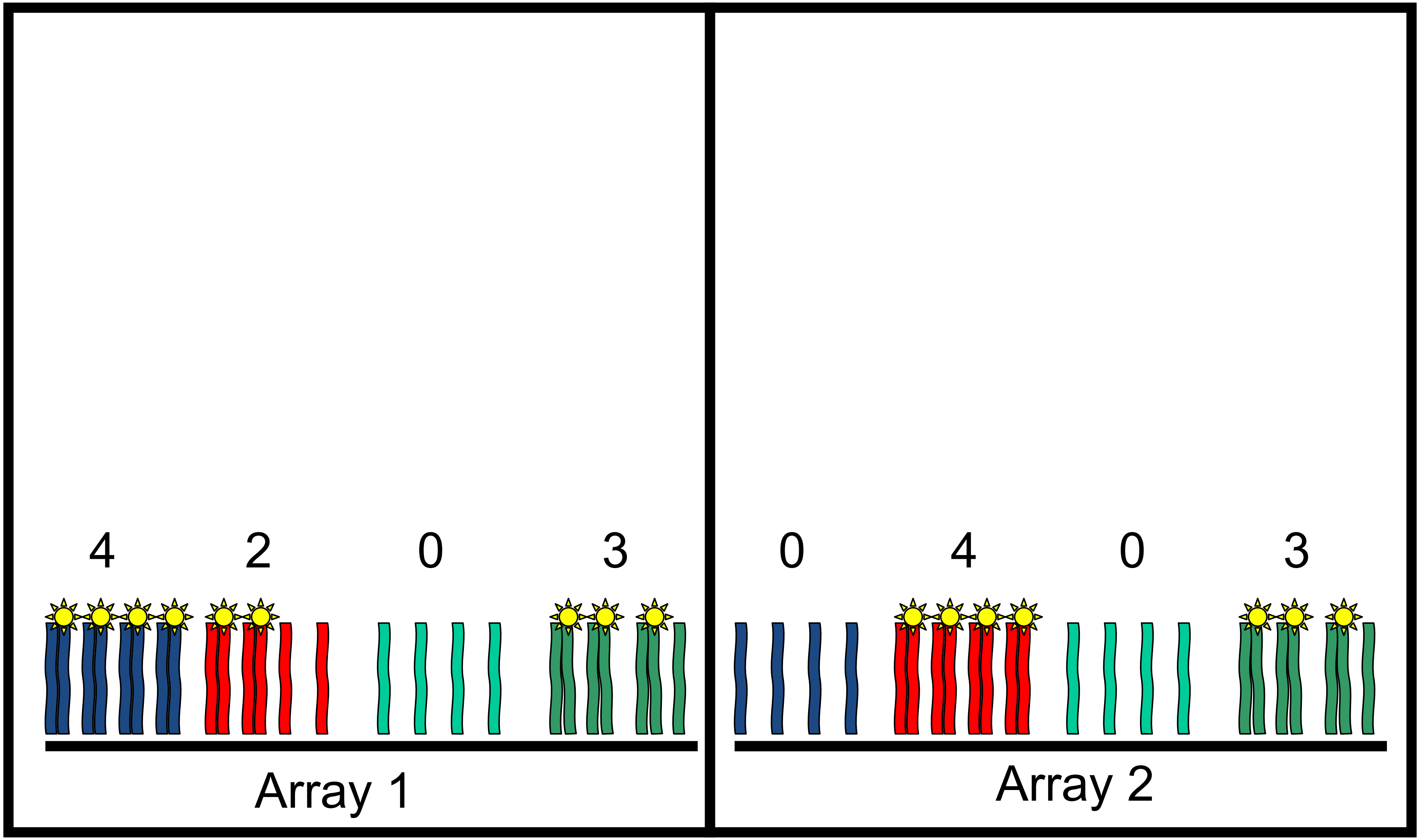


Image quantification



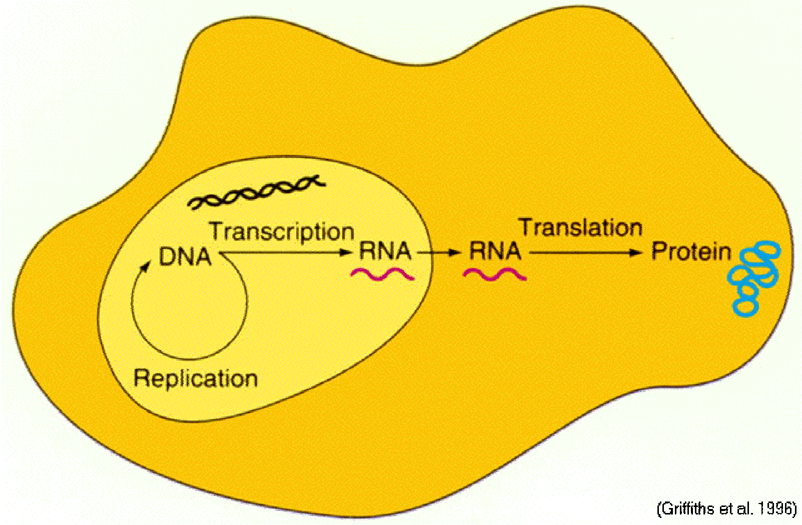
Gene expression assays

- Spotted cDNA arrays (Brown/Botstein);
- Short oligonucleotide arrays (Affymetrix);
- Long oligonucleotide arrays (Agilent Inkjet);
- Fibre optic arrays (Illumina);
- Serial analysis of gene expression (SAGE);
- …

Applications of microarrays

- Measuring transcript abundance (cDNA arrays);
- Genotyping;
- Estimating DNA copy number (CGH);
- Determining identity by descent (GMS);
- Measuring mRNA decay rates;
- Identifying protein binding sites;
- Determining sub-cellular localization of gene products;
- …

Transcriptome



(Griffiths et al. 1996)

- mRNA or transcript levels sensitively reflect the state of a cell.
- Measuring protein levels (translation) would be more direct but more difficult.

Transcriptome

- The transcriptome reflects
  - Tissue source: cell type, organ.
  - Tissue activity and state:
    - Stage of development, growth, death.
    - Cell cycle.
    - Disease vs. healthy.
    - Response to therapy, stress.

Applications of microarrays

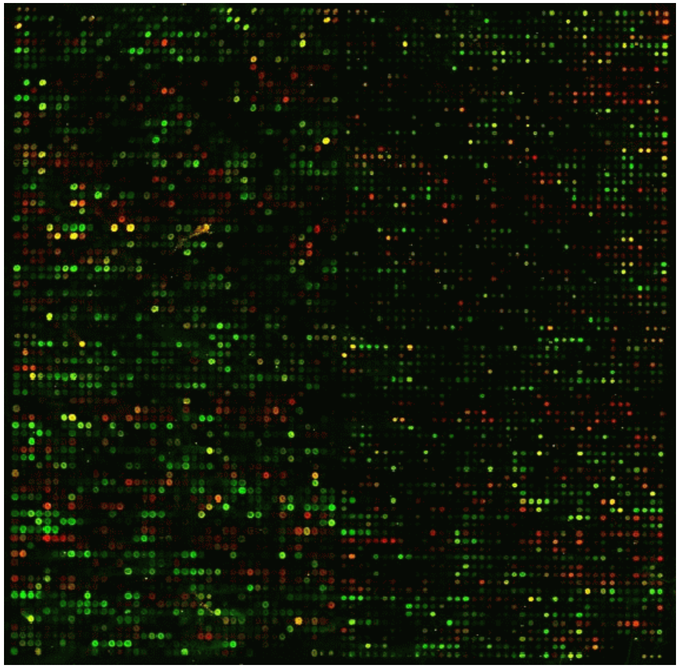
- **Cancer research:** Molecular characterization of tumors on a genomic scale

  → more reliable diagnosis and effective treatment of cancer.
- **Immunology:** Study of host genomic responses to bacterial infections.
- …

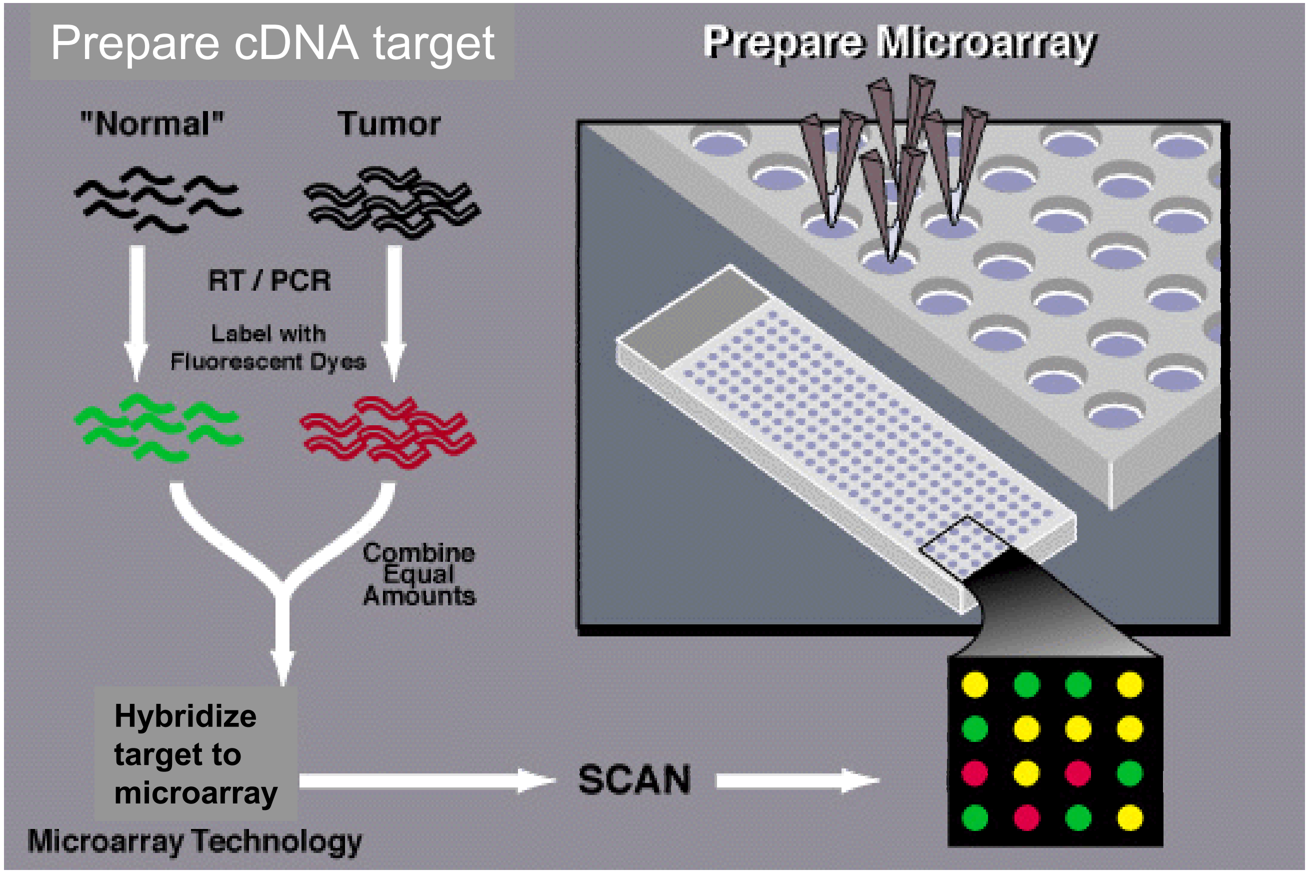
Applications of microarrays

- Compare mRNA (transcript) levels in different types of cells, i.e., vary
  - Tissue: liver vs. brain;
  - Treatment: drugs A, B, and C;
  - State: tumor vs. non-tumor, development;
  - Organism: different yeast strains;
  - Timepoint;
  - etc.

Spotted DNA microarrays



Spotted DNA microarrays



Spotted DNA microarrays

- The relative abundance of a spotted DNA sequence in two DNA or RNA samples may be assessed by monitoring the differential hybridization of these two samples to the sequence on the array.
- Probes: DNA sequences spotted on the array, immobile substrate.
- Targets: Nucleic acid samples hybridized to the array, mobile substrate.

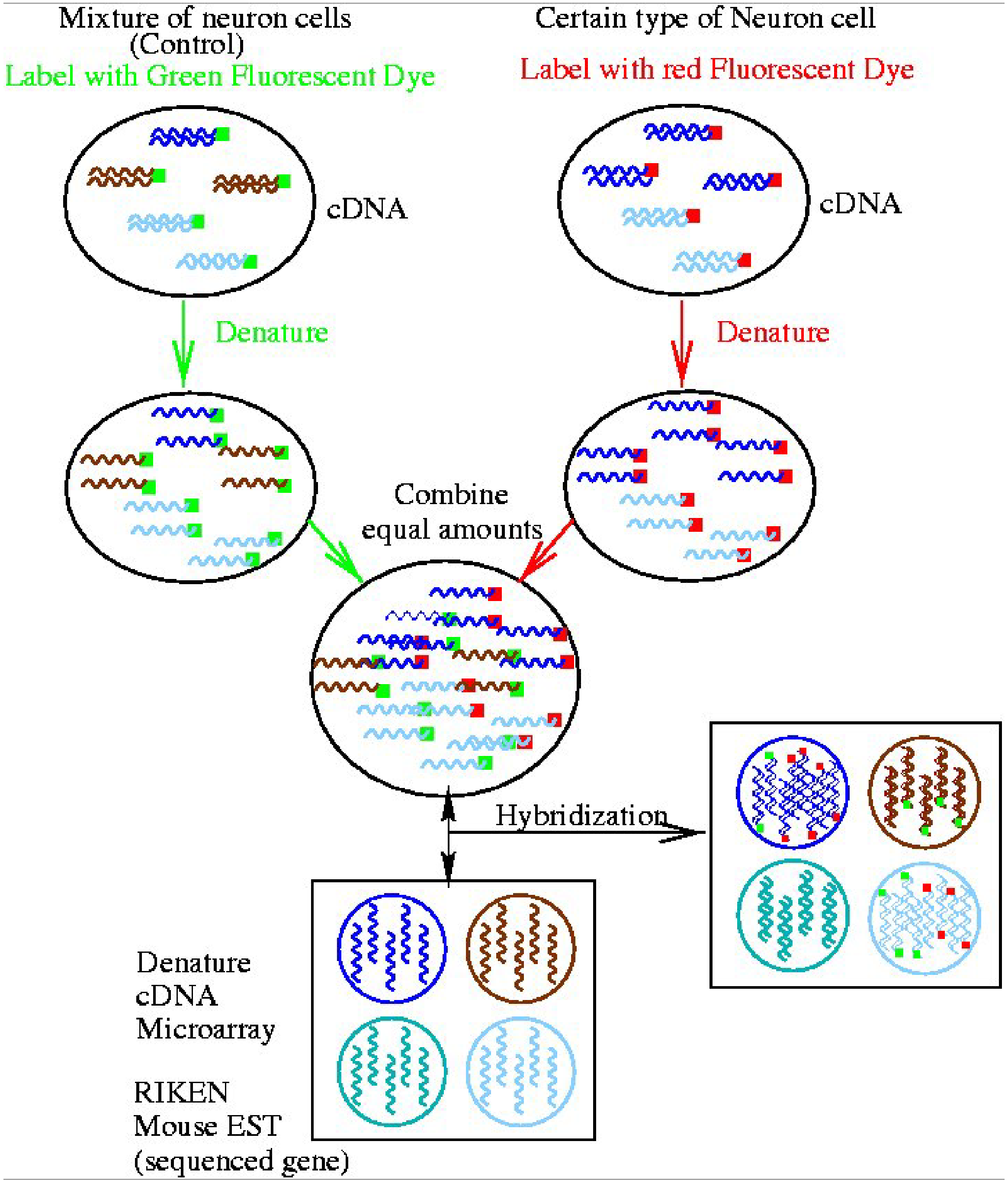
Spotted DNA microarrays

- The ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples.

Spotted DNA microarrays

$$M = \log_2 R/G = \log_2 R - \log_2 G$$

- **$M < 0$**, gene is over-expressed in green-labeled sample compared to red-labeled sample.
- **$M = 0$**, gene is equally expressed in both samples.
- **$M > 0$**, gene is over-expressed in red-labeled sample compared to green-labeled sample.



Mixture of neuron cells
(Control)
Label with Green Fluorescent Dye

Certain type of Neuron cell
Label with red Fluorescent Dye

cDNA

cDNA

Denature

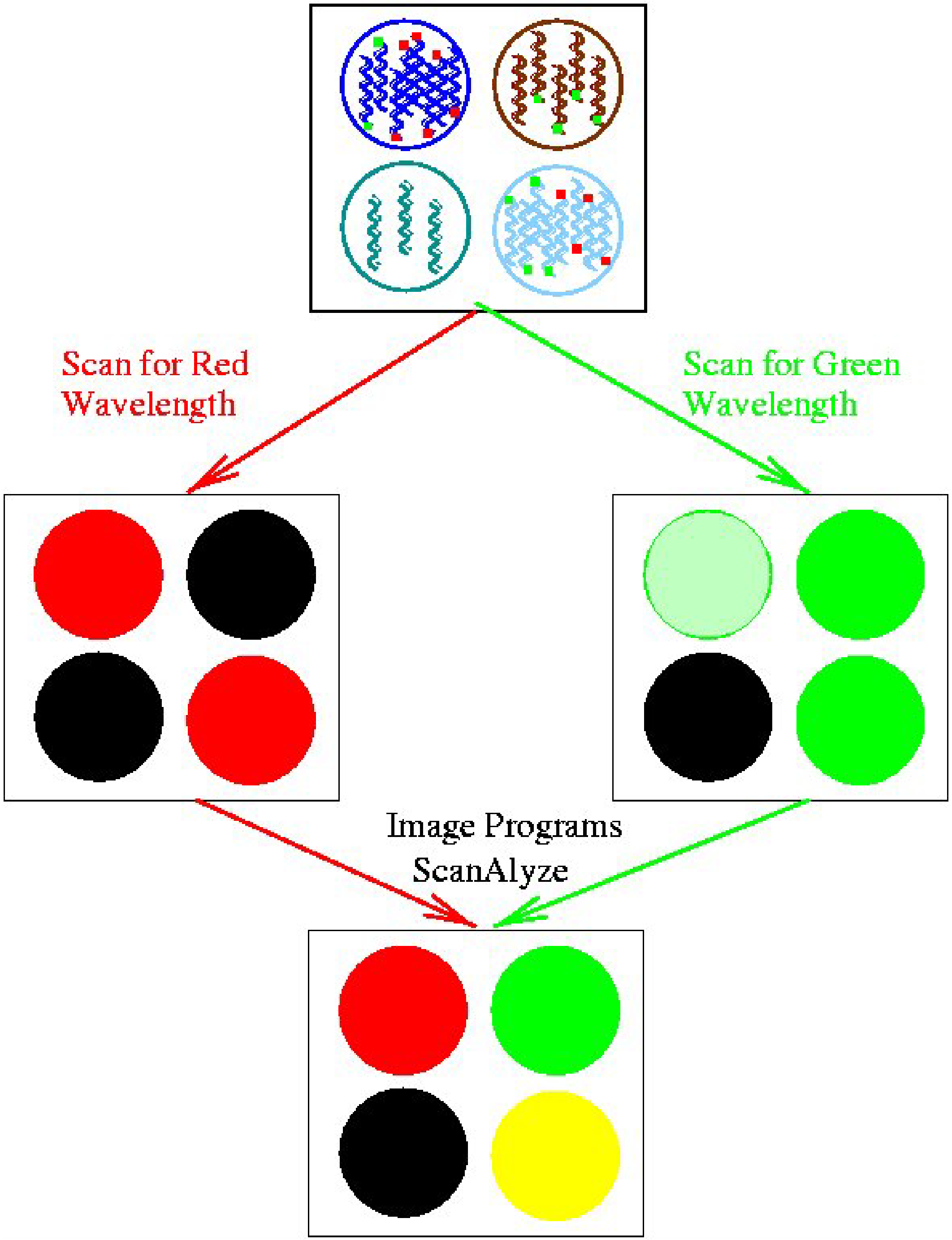
Denature

Combine
equal amounts

Hybridization

Denature
cDNA
Microarray

RIKEN
Mouse EST
(sequenced gene)

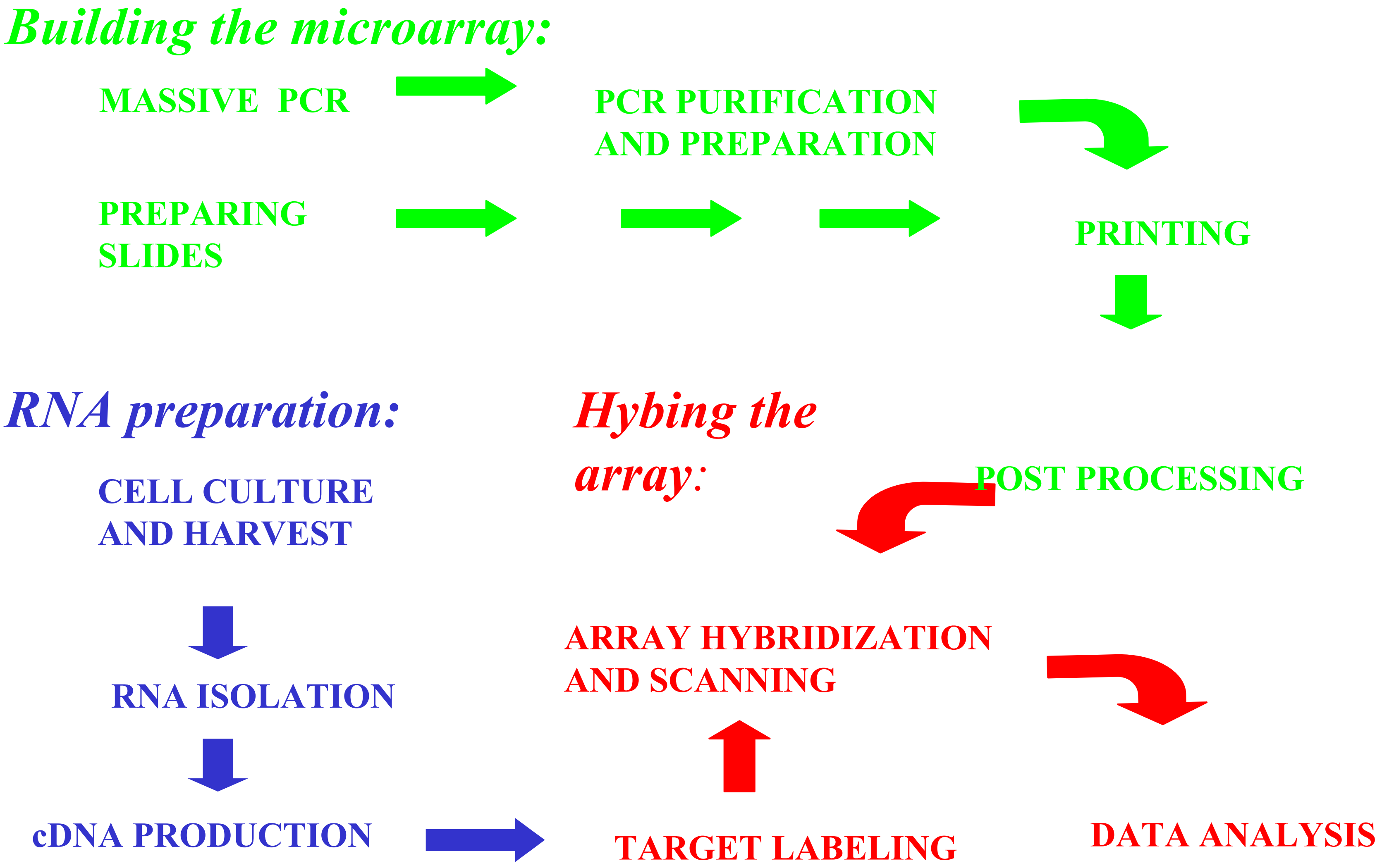


Scan for Red
Wavelength

Scan for Green
Wavelength

Image Programs
ScanAlyze

The process



Building the microarray:

MASSIVE PCR

PCR PURIFICATION AND PREPARATION

PREPARING SLIDES

PRINTING

POST PROCESSING

RNA preparation:

CELL CULTURE AND HARVEST

RNA ISOLATION

cDNA PRODUCTION

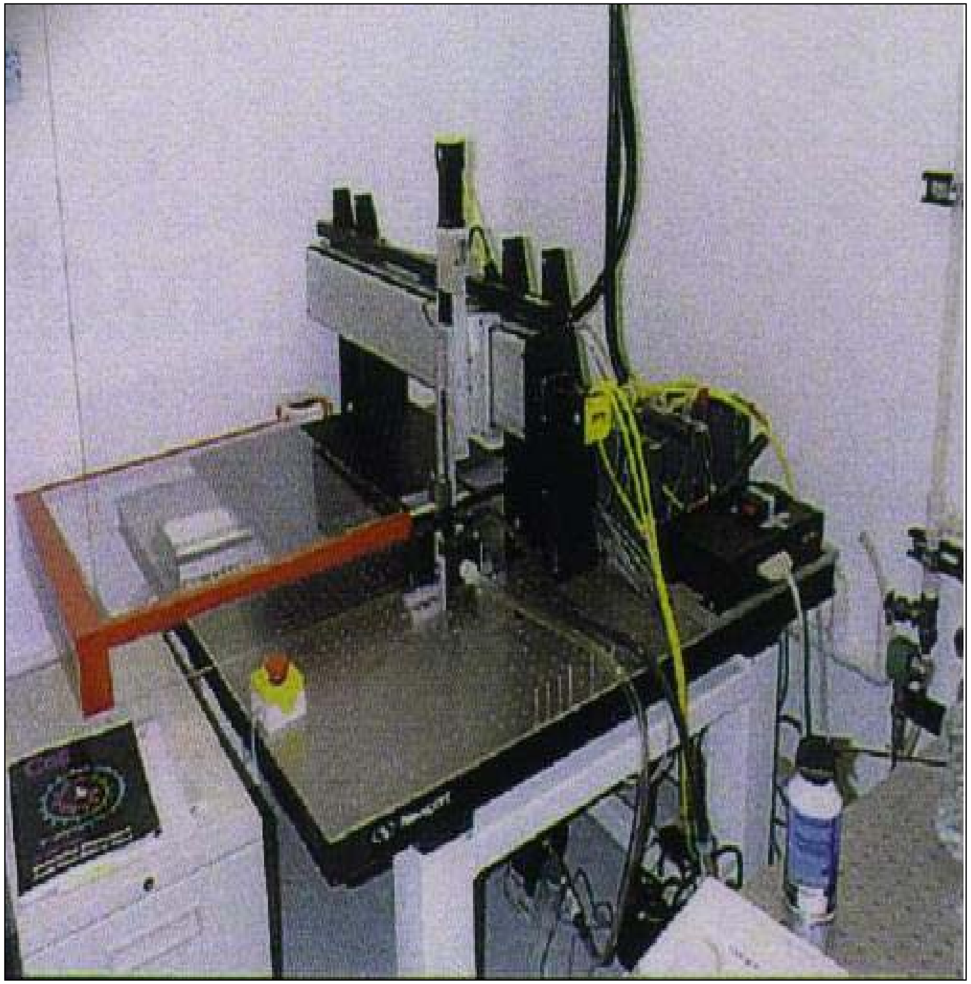
Hybing the array:

TARGET LABELING

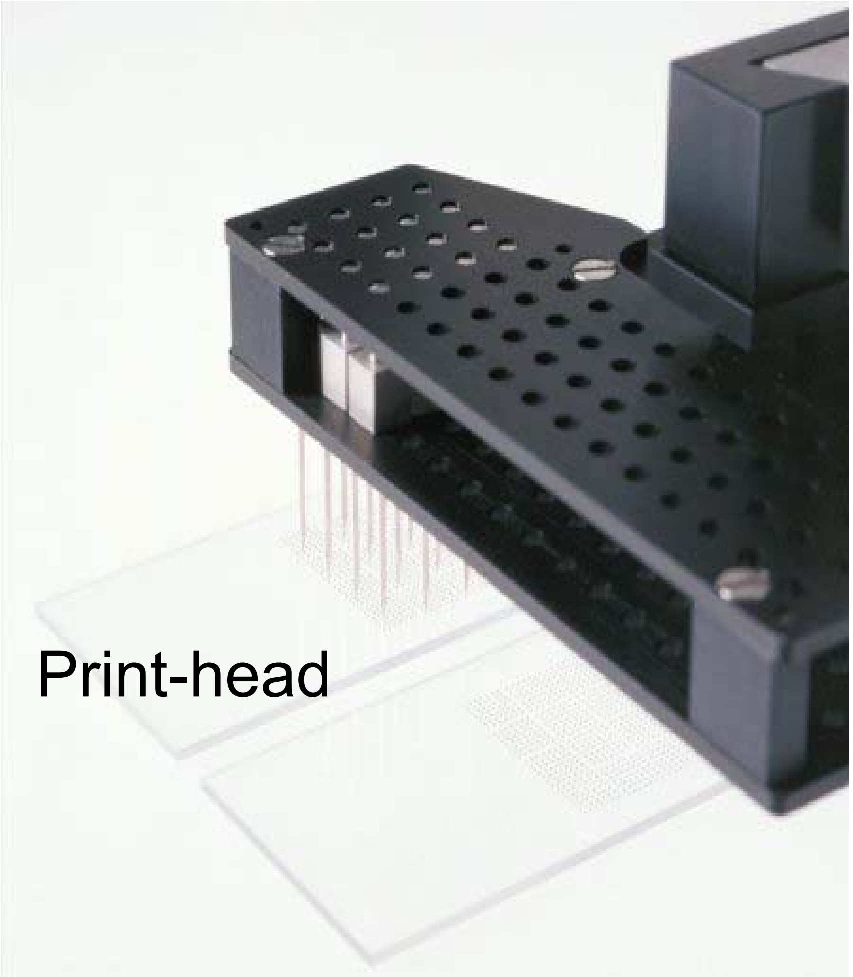
ARRAY HYBRIDIZATION AND SCANNING

DATA ANALYSIS

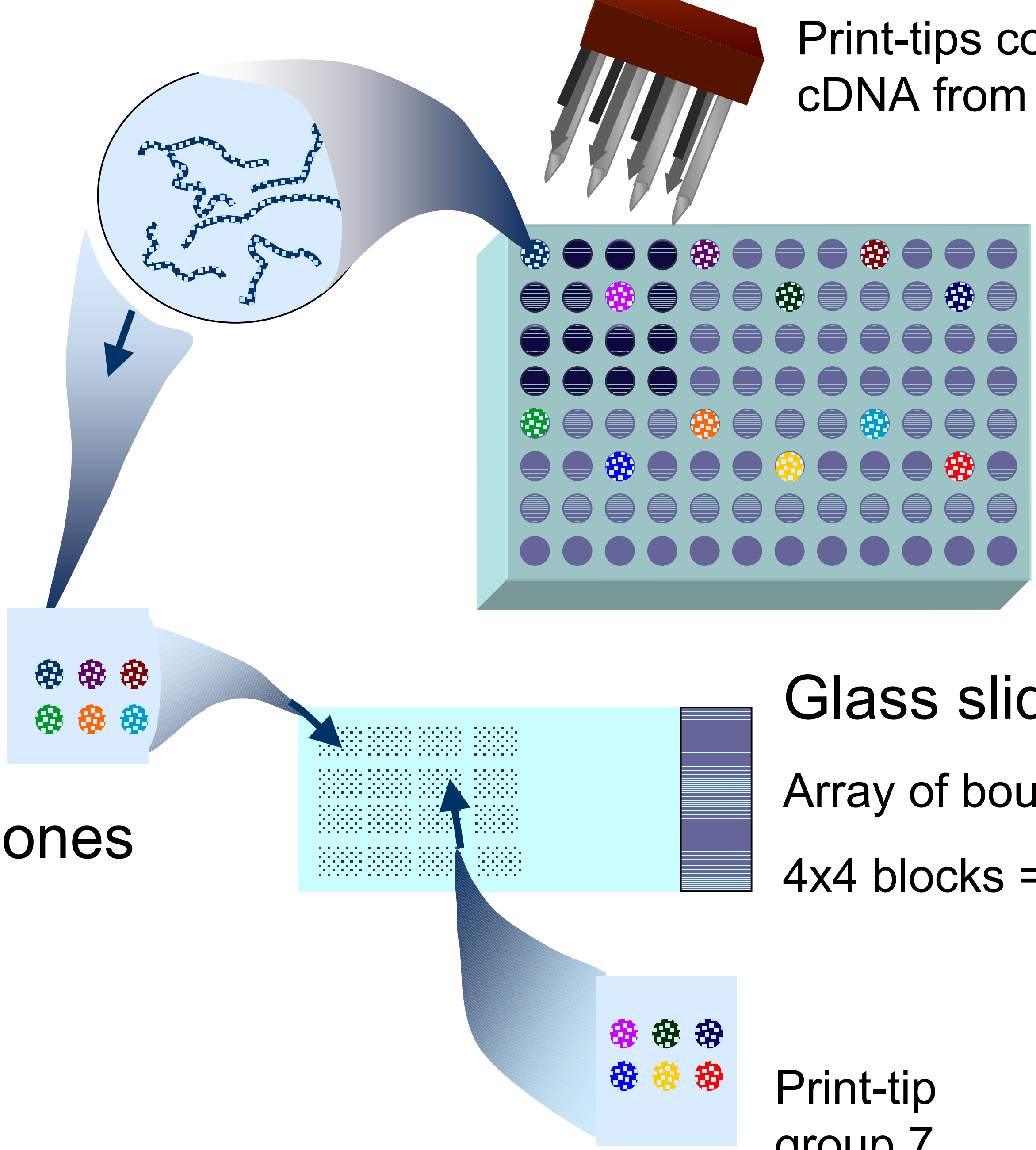
The arrayer



Ngai Lab arrayer, UC Berkeley



Print-head



Print-tips collect cDNA from wells

96-well plate

Contains cDNA probes

Print-tip group 1

cDNA clones

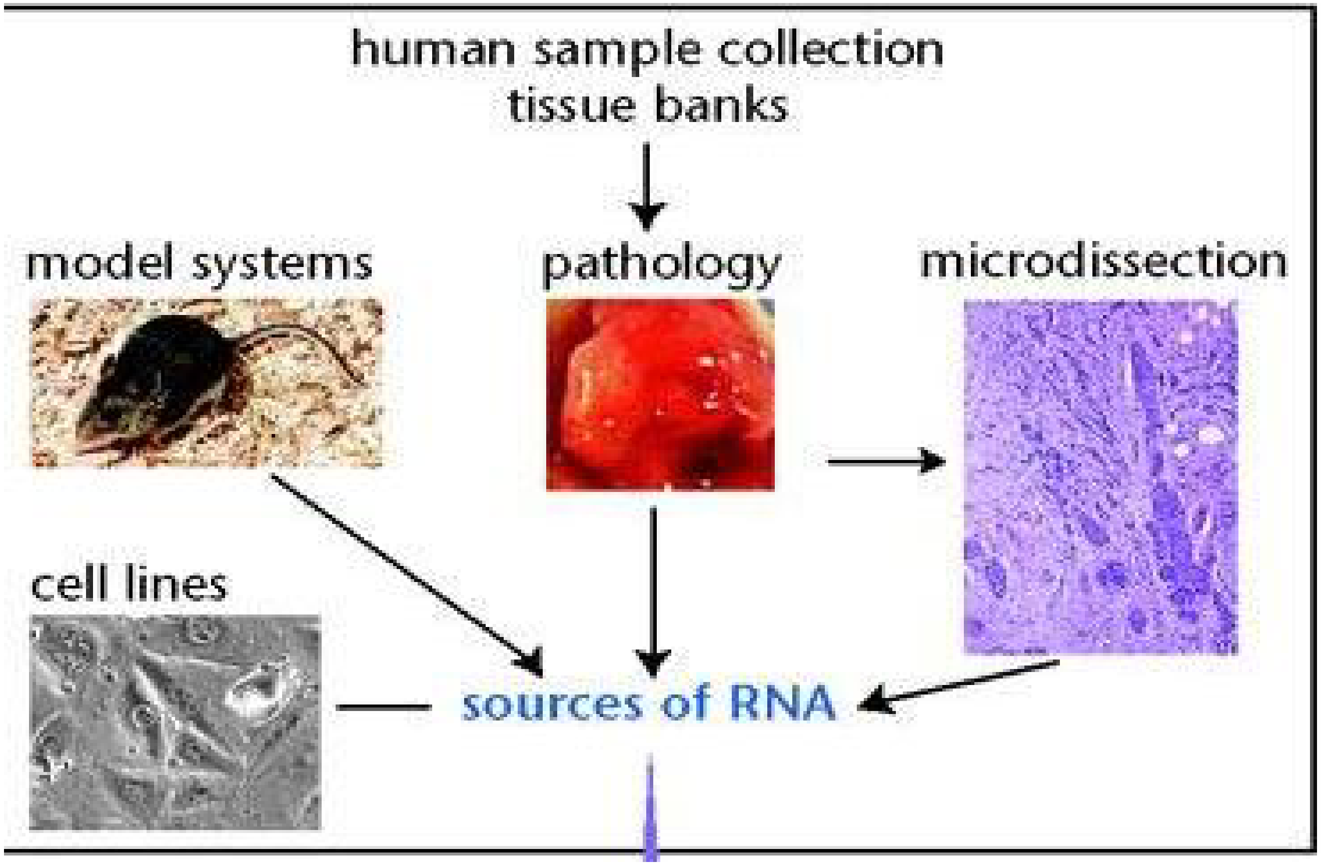
Glass slide

Array of bound cDNA probes

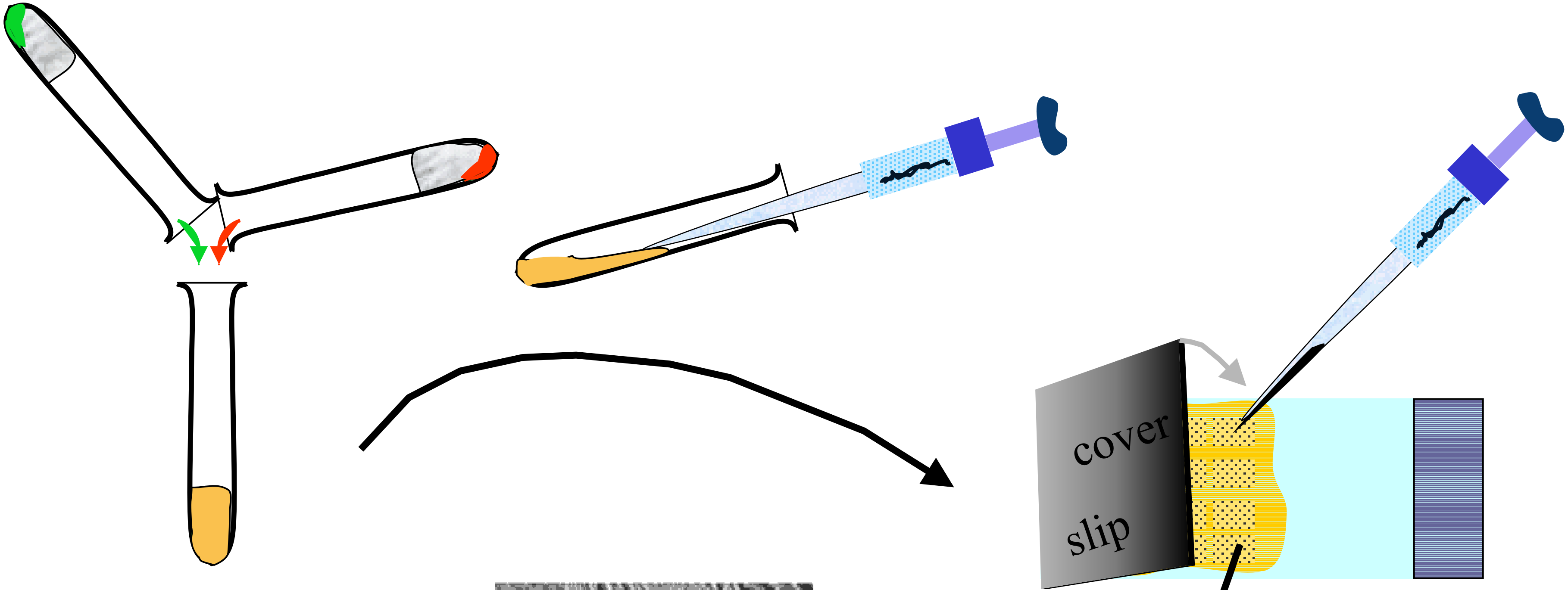
4x4 blocks = 16 print-tip-groups

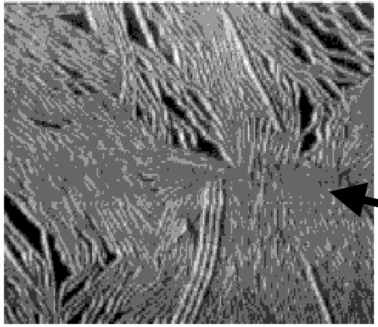
Print-tip group 7

Sample preparation



Hybridization



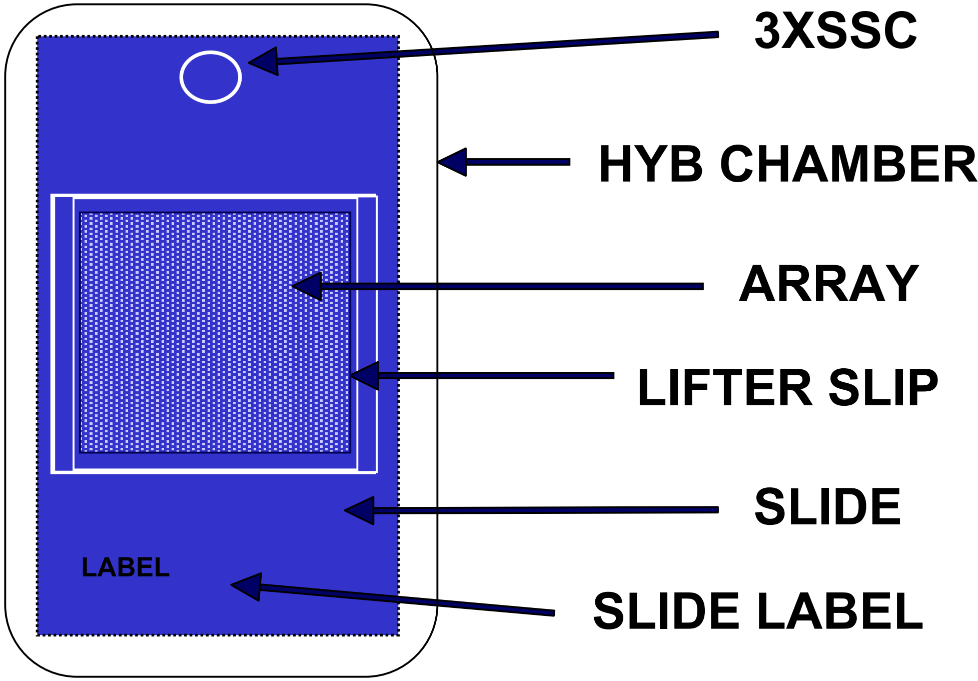


Binding of cDNA target samples to cDNA probes on the slide

Hybridize for

5-12 hours

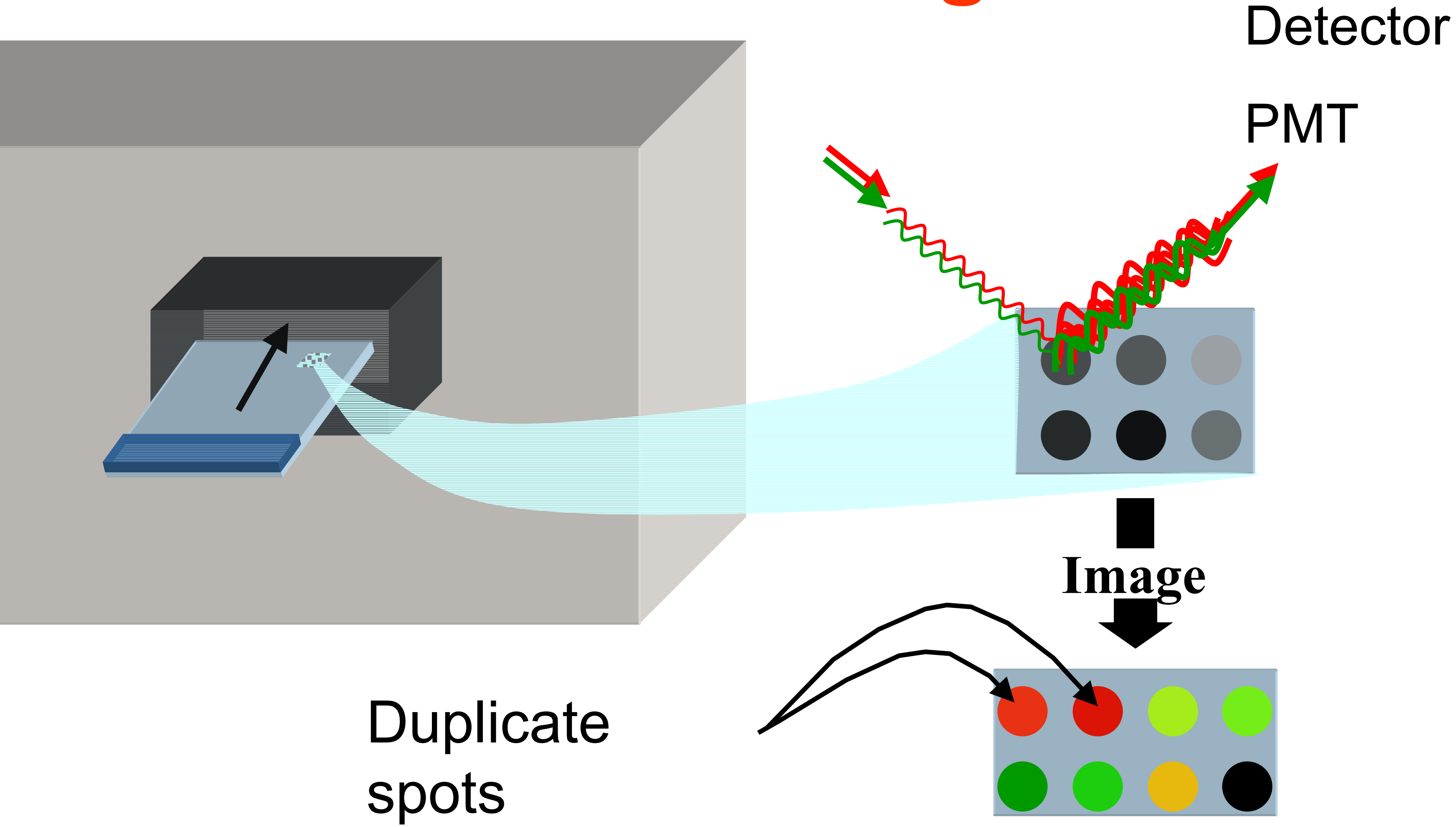
Hybridization chamber



- Humidity
- Temperature
- Formamide

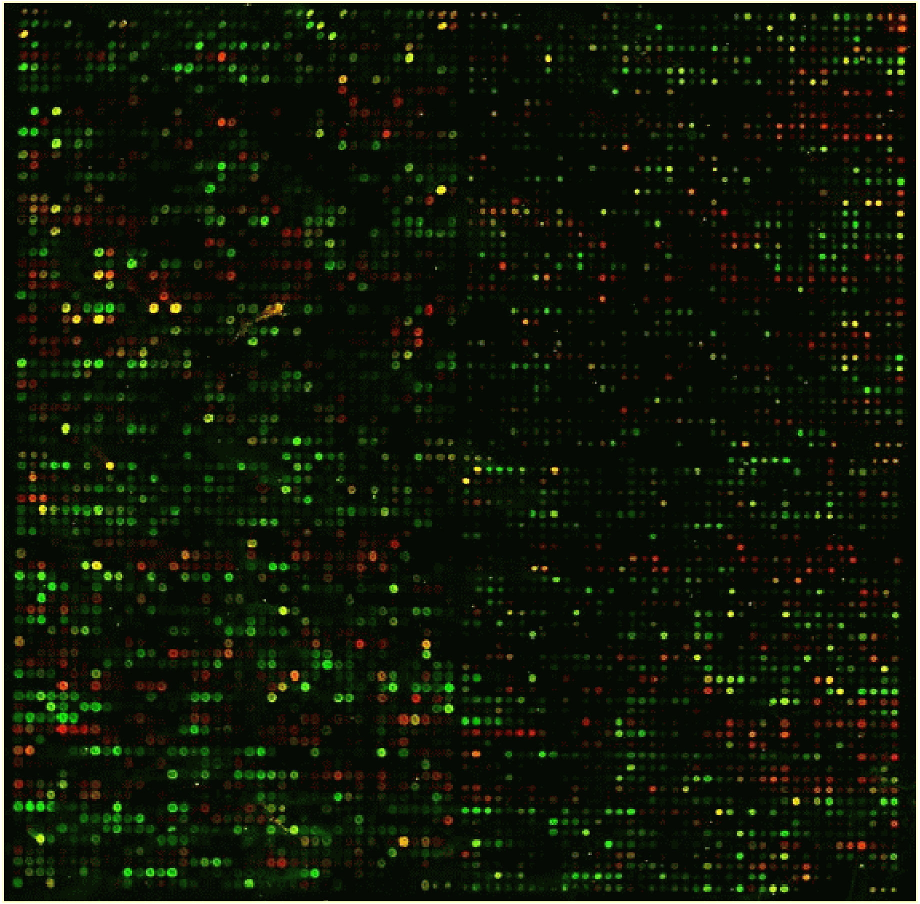
(Lowers the Tmp)

Scanning



Cy5: 635nm
Cy3: 532nm

RGB overlay of Cy3 and Cy5 images



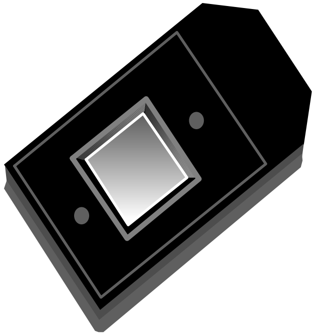
Raw data

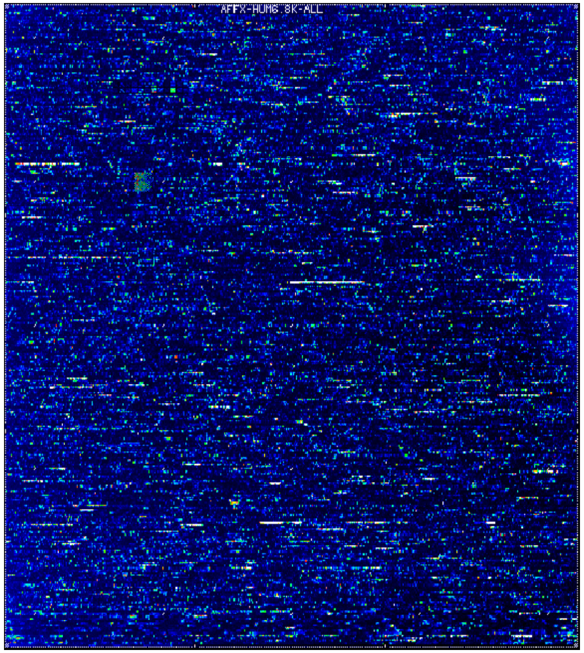
- Pairs of 16–bit TIFFs, one for each dye.
- E.g. Human cDNA arrays:
  - ~43K spots;
  - ~ 20Mb per channel;
  - ~ 2,000 x 5,500 pixels per image;
  - spot separation: ~ 136um.
- For a “typical” array, the spot area has
  - mean = 43 pixels,
  - med = 32 pixels,
  - SD = 26 pixels.

Animation

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

Oligonucleotide chips





Terminology

- Each gene or portion of a gene is represented by 16 to 20 oligonucleotides of 25 base-pairs.

- Probe: an oligonucleotide of 25 base-pairs, i.e., a 25-mer.
- Perfect match (PM): A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
- Mismatch (MM): same as PM but with a single homomeric base change for the middle (13th) base (transversion purine <-> pyrimidine, G <->C, A <->T) .
- Probe-pair: a (PM,MM) pair.
- Probe-pair set: a collection of probe-pairs (16 to 20) related to a common gene or fraction of a gene.
- Affy ID: an identifier for a probe-pair set.
- The purpose of the MM probe design is to measure non-specific binding and background noise.

Probe-pair set

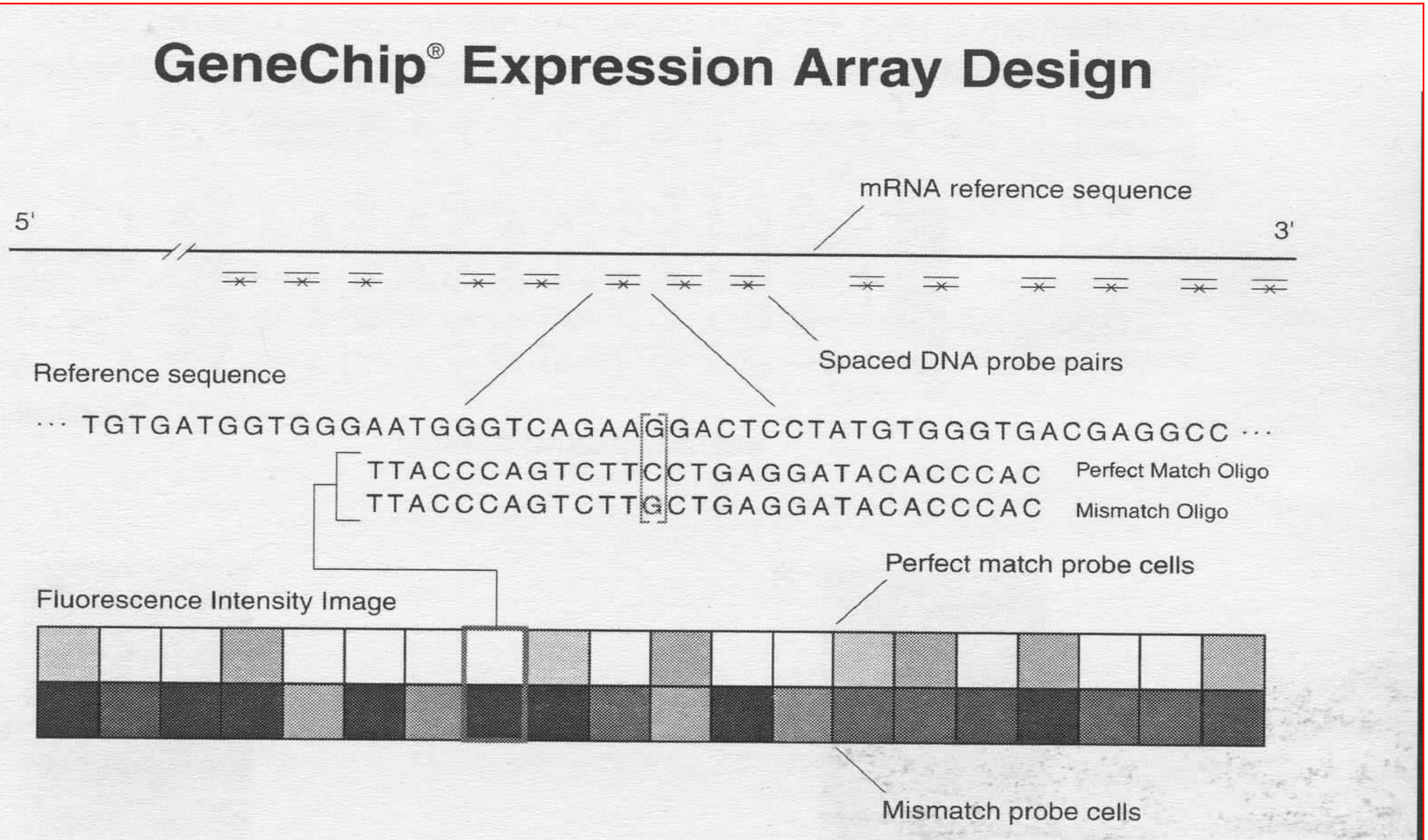
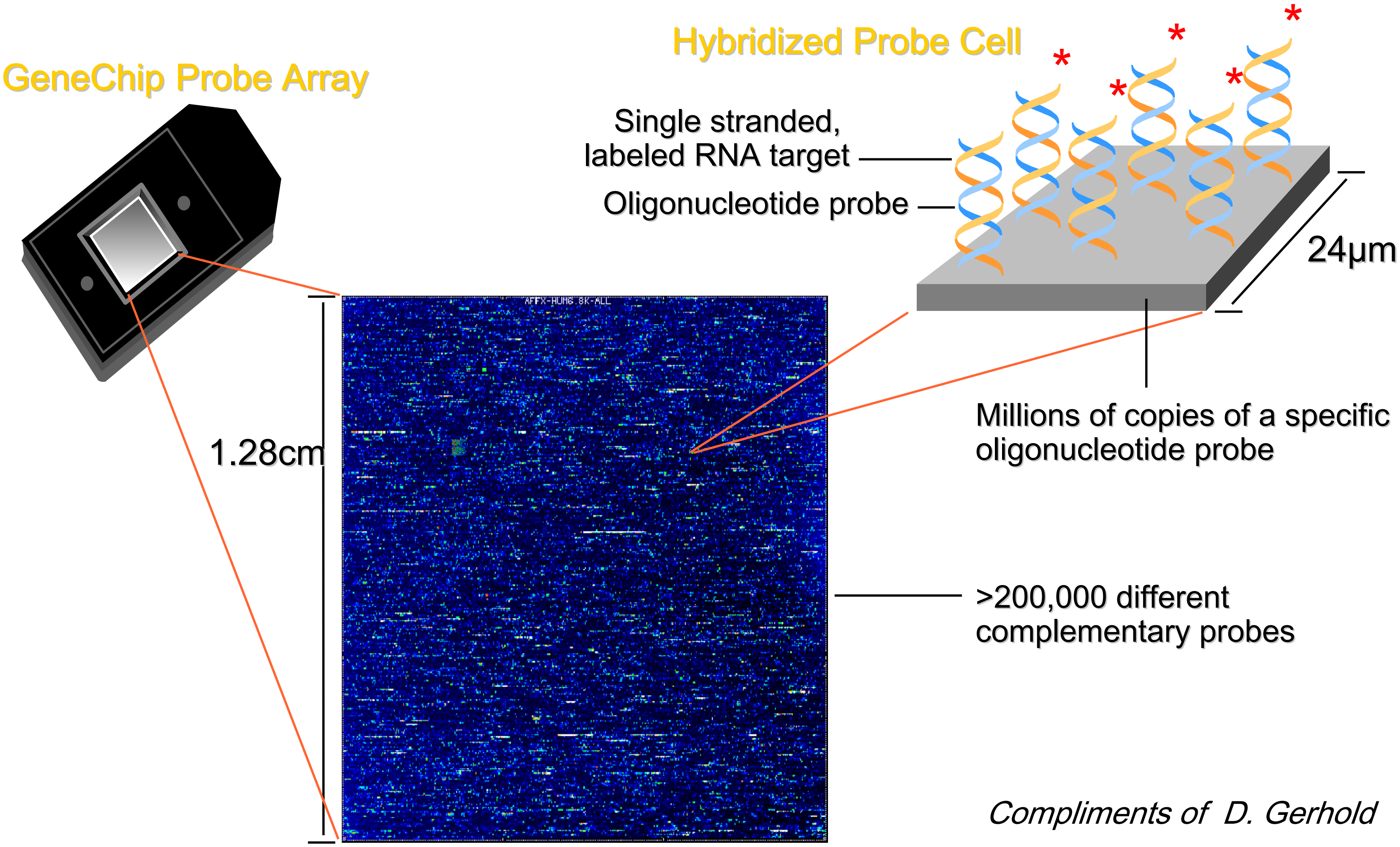

GeneChip® Expression Array Design
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
TTACCCAGTCTTCCTGAGGATACACCCAC
Perfect Match Oligo
TTACCCAGTCTTGCTGAGGATACACCCAC
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 1-3 Expression tiling strategy

Spotted vs. Affymetrix arrays

Spotted arrays	Affymetrix arrays
One probe per gene	16 – 20 probe-pairs per gene
Probes of varying length	Probes are 25-mers
Two target samples per array	One target sample per array

Oligonucleotide chips



GeneChip Probe Array

Hybridized Probe Cell

Single stranded, labeled RNA target

Oligonucleotide probe

24µm

1.28cm

Millions of copies of a specific oligonucleotide probe

>200,000 different complementary probes

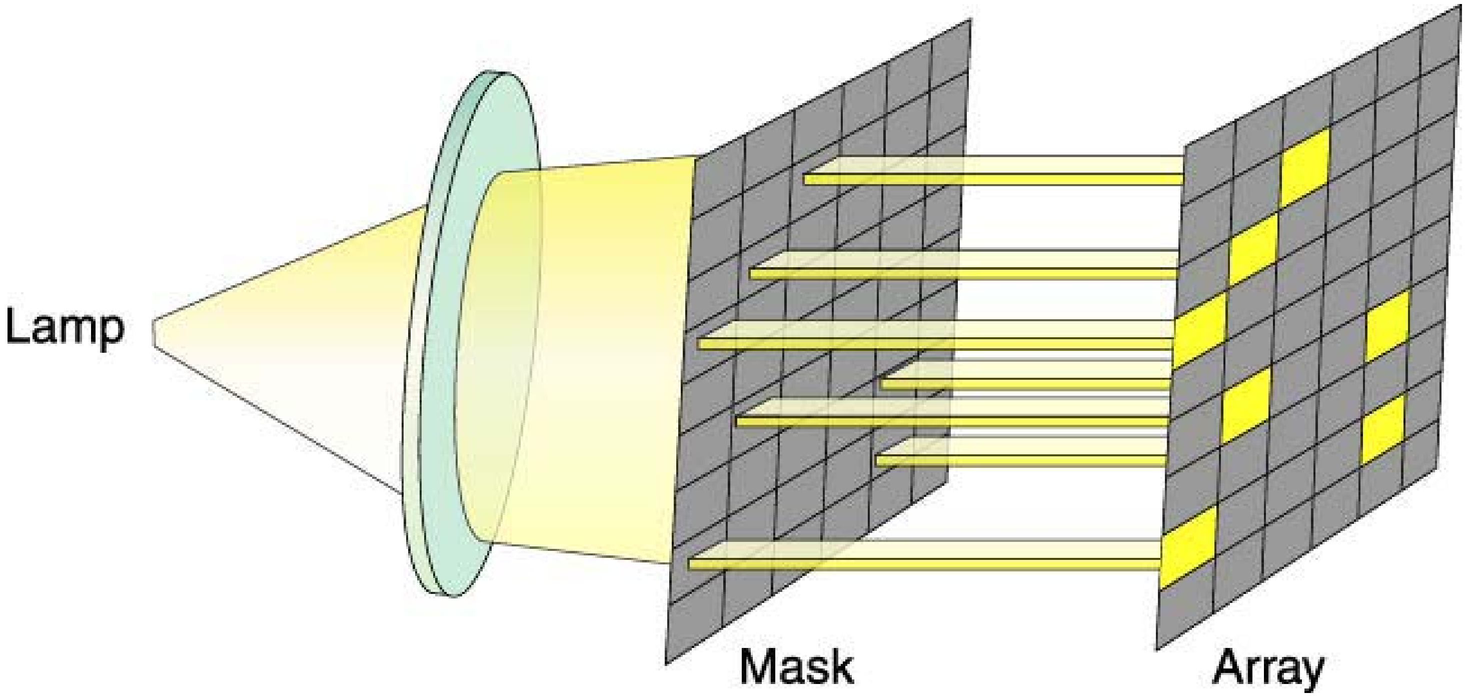
Image of Hybridized Probe Array

Compliments of D. Gerhold

Oligonucleotide chips

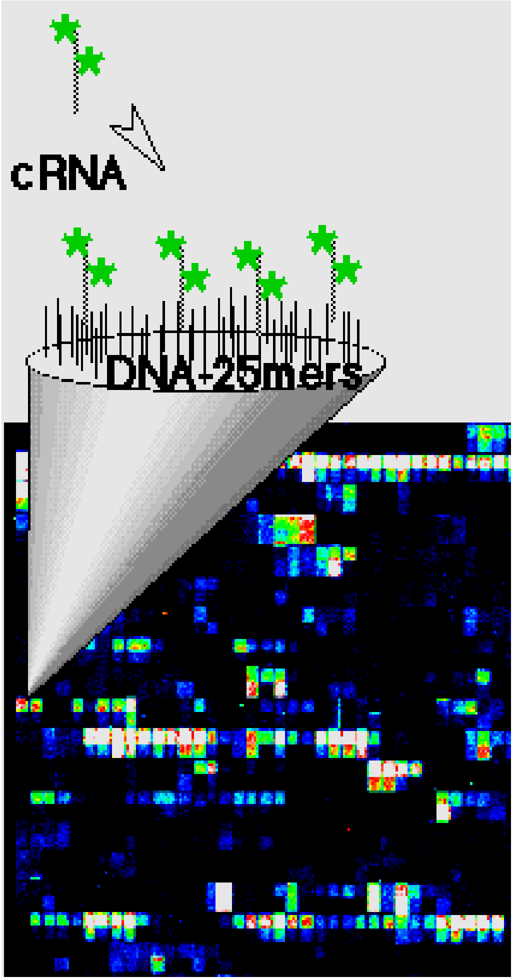
- The probes are synthesized *in situ*, using combinatorial chemistry and photolithography.
- Probe cells are square-shaped features on the chip containing millions of copies of a single 25-mer probe. Sides are 18-50 microns.

Oligonucleotide chips



The manufacturing of GeneChip® probe arrays is a combination of photolithography and combinatorial chemistry.

Image analysis



- About 100 pixels per probe cell.
- These intensities are combined to form one number representing the expression level for the probe cell oligo.
- → CEL file with PM or MM intensity for each cell.

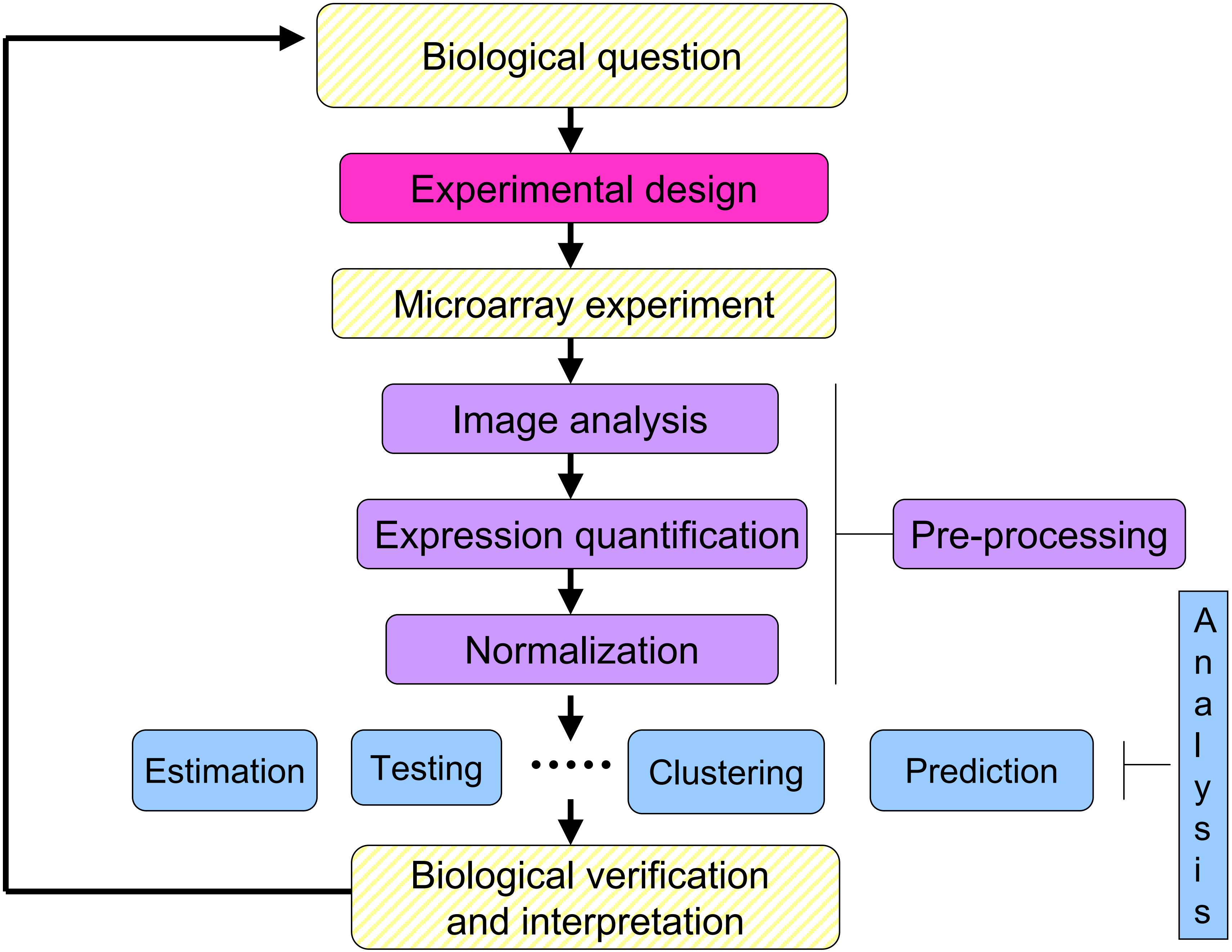
Expression measures

- Most expression measures are based on differences of **PM-MM**.
- The intention is to correct for background and non-specific binding.
- E.g. *MarrayArray Suite*® (MAS) v. 4.0 uses Average Difference Intensity (ADI) or

  AvDiff = average of PM-MM.
- Problem: MM may also measure signal.
- More on this in lecture *Pre-processing DNA Microarray Data.*

WWW resources

- **Complete guide to “microarraying”**
  http://cmgm.stanford.edu/pbrown/mguide/
  http://www.microarrays.org
  - Parts and assembly instructions for printer and scanner;
  - Protocols for sample prep;
  - Software;
  - Forum, etc.
- **cDNA microarray animation**
  http://www.bio.davidson.edu/courses/genomics/chip/chip.html
- **Affymetrix**
  http://www.affymetrix.com



Biological question

Experimental design

Microarray experiment

Image analysis

Expression quantification

Normalization

Pre-processing

Estimation

Testing

……

Clustering

Prediction

Analysis

Biological verification and interpretation

Statistical computing

Everywhere …

- Statistical design and analysis:
  - image analysis, normalization, estimation, testing, clustering, prediction, etc.

- Integration of experimental data with biological metadata from WWW-resources
  - gene annotation (GenBank, LocusLink);
  - literature (PubMed);
  - graphical (pathways, chromosome maps).

Outline

- Introduction to the biology and technology of DNA microarrays
- Overview of the Bioconductor project
- Annotation
- Visualization
- Pre-processing: spotted and Affymetrix arrays
- Differential gene expression
- Software demo

Overview of the Bioconductor Project

Bioconductor

- Bioconductor is an open source and open development software project for the analysis and comprehension of biomedical and genomic data.

- Software, data, and documentation are available from www.bioconductor.org.

Bioconductor

- The project was started in the Fall of 2001 by Robert Gentleman, at the Biostatistics Unit of the Dana Farber Cancer Institute.
- There are currently 21 core developers, at various institutions in the US and Europe.
- R and the R package system are used to design and distribute software (www.r-project.org).
- First release (v 1.0): May 2nd, 2002, 15 packages.
- Second release (v 1.1): November 18th, 2002, 5 new packages.

Bioconductor

There are two main classes of packages

- End-user packages:
  - aimed at users unfamiliar with R or computer programming;
  - polished and easy to use interfaces to a wide variety of computational and statistical methods for the analysis of genomic data.
- Developer packages: aimed at software developers, in the sense that they provide ``software to write software''.

Bioconductor packages

Release 1.1, November 18th, 2002

- General infrastructure:
  `Biobase`, **`reposTools`**, **`rhdf5`**, **`tkWidgets`**.
- Annotation:
  `annotate`, **`AnnBuilder`** → data packages.
- Graphics:
  `geneplotter`, **`hexbin`**.
- Pre-processing for Affymetrix oligonucleotide chip data:
  `affy`, **`vsn`**, CDF packages.
- Pre-processing for spotted DNA microarray data:
  `marrayClasses`, **`marrayInput`**, **`marrayNorm`**, **`marrayPlots`**, **`marrayTools`**, **`vsn`**.
- Differential gene expression:
  `edd`, **`genefilter`**, **`multtest`**, **`ROC`**.
- Graphs:
  `graph`.

Ongoing efforts

- Variable (feature) selection;
- Prediction;
- Cluster analysis;
- Cross-validation;
- Multiple testing;
- Quality measures for microarray data;
- Interactions with MAGE-ML;
- Biological sequence analysis;
- Etc.

Computing needs

- Mechanisms for facilitating the design and deployment of portable, extensible, and scalable software.
- Support for interoperability with software written in other languages.
- Tools for integrating biological metadata from the WWW in the analysis of experimental metadata.
- Access to a broad range of statistical and numerical methods.
- High-quality visualization and graphics tools that support interactivity.
- An effective, extensible user interface.
- Tools for producing innovative, high-quality documentation and training materials.
- Methodology that supports the creation, testing, and distribution of software and data modules.

Bioconductor

- Interactive tools for linking experimental data in real time, to biological metadata from WWW resources.

  E.g. PubMed, GenBank, LocusLink.

- Scenario. Normalize spotted array data with **`marrayNorm`**, obtain list of differentially expressed genes from **`multtest`** or **`genefilter`**, use the **`annotate`** package
  - to retrieve and search PubMed abstracts for these genes;
  - to generate an HTML report with links to LocusLink for each gene.

Bioconductor

- Widgets. Small-scale graphical user interfaces (GUI), providing point & click access for specific tasks (**`tkWidgets`**).
- E.g. File browsing and selection for data input, basic analyses.

- Object-oriented class/method design. Allows efficient representation and manipulation of large and complex biological datasets of multiple types (cf. MIAME standards).

Object-oriented programming

- The Bioconductor project has adopted the object-oriented programming – OOP – paradigm presented in

  J. M. Chambers (1998). *Programming with Data*.
- Tools for programming using the class/method mechanism are provided in the R `methods` package.
- Tutorial: www.omegahat.org/RSMethods/index.html

OOP

- A class provides a software abstraction of a real world object. It reflects how we think of certain objects and what information these objects should contain.
- Classes are defined in terms of slots which contain the relevant data.
- An object is an instance of a class.
- A class defines the structure, inheritance, and initialization of objects.

OOP

- A method is a function that performs an action on data (objects).
- Methods define how a particular function should behave depending on the class of its arguments.
- Methods allow computations to be adapted to particular data types, i.e., classes.
- A generic function is a dispatcher, it examines its arguments and determines the appropriate method to invoke.
- Examples of generic functions include **`plot`**, **`summary`**, **`print`**.

Data

- Issues:
  - complexity;
  - size;
  - evolution.

- We distinguish between biological metadata and experimental metadata.

Experimental metadata

- Gene expression measures
  - scanned images, i.e., raw data;
  - image quantitation data, i.e., output from image analysis;
  - normalized expression measures, i.e., log ratios M or Affy measures.
- Reliability information for the expression measures.
- Information on the probe sequences printed on the arrays (array layout).
- Information on the target samples hybridized to the arrays.
- See *Minimum Information About a Microarray Experiment – MIAME* – standards.

Biological metadata

- Biological attributes that can be applied to the experimental data.
- E.g. for genes
  - chromosomal location;
  - gene annotation (LocusLink, GO);
  - relevant literature (PubMed).
- Biological metadata sources are large, complex, evolving rapidly, and typically distributed via the WWW.

`exprSet` class

`exprs`	Matrix of expression measures, genes x samples
`se.exprs`	Matrix of SEs for expression measures, genes x samples
`phenoData`	Sample level covariates, instance of class **`phenoData`**
`annotation`	Name of annotation data
`description`	MIAME information
`notes`	Any notes

marrayRaw class

Pre-normalization intensity data for a batch of arrays

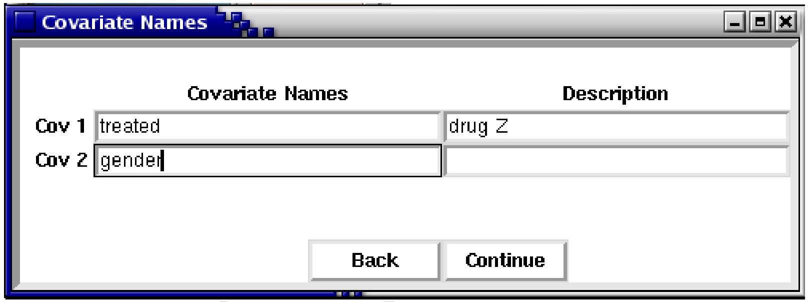
maRf	**maGf**	Matrix of red and green foreground intensities
maRb	**maGb**	Matrix of red and green background intensities
maW		Matrix of spot quality weights
maLayout		Array layout parameters - **marrayLayout**
maGnames		Description of spotted probe sequences - **marrayInfo**
maTargets		Description of target samples - **marrayInfo**
maNotes		Any notes

`AffyBatch` class

Probe-level intensity data for a batch of arrays (same `CDF`)

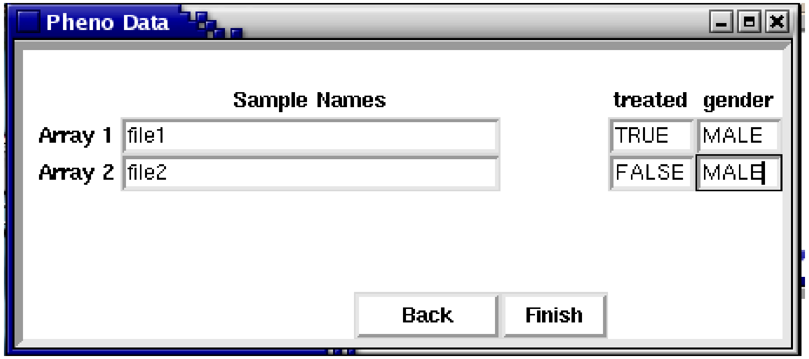
`cdfName`		Name of **`CDF`** file for arrays in the batch
`nrow`	`ncol`	Dimensions of the array
`exprs`	`se.exprs`	Matrices of probe-level intensities and SEs rows → probe cells, columns → arrays.
`phenoData`		Sample level covariates, instance of class **`phenoData`**
`annotation`		Name of annotation data
`description`		MIAME information
`notes`		Any notes

Reading in phenoData



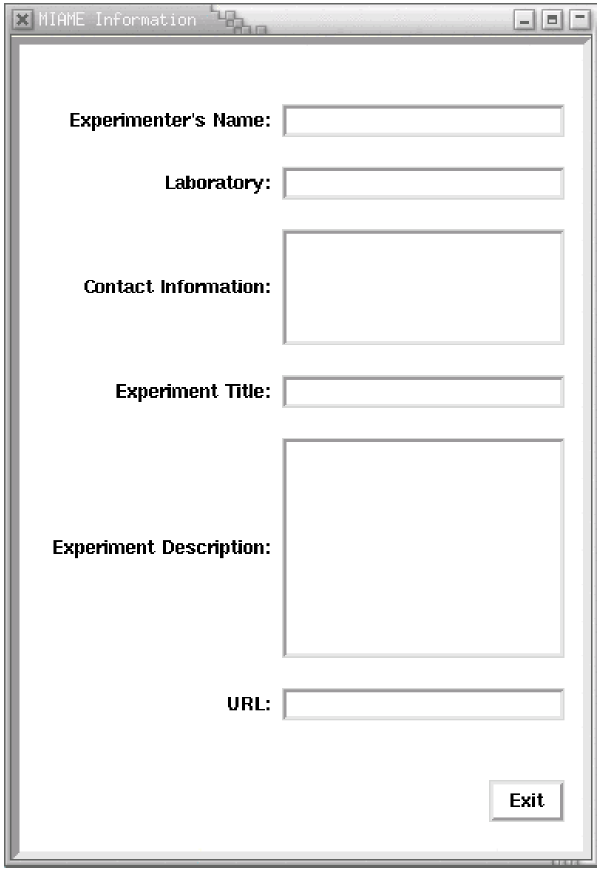
	Covariate Names	Description
Cov 1	treated	drug Z
Cov 2	gender	

tkSampleNames



	Sample Names	treated	gender
Array 1	file1	TRUE	MALE
Array 2	file2	FALSE	MALE

tkphenoData



tkMIAME

Pedagogy

Extensive documentation and training resources for R and Bioconductor are available on the WWW.

- R manuals and tutorials are available from the R website.
- R help system
  - detailed on-line documentation, available in text, HTML, PDF, and LaTeX formats;
  - e.g. **`help(genefilter), ?pubmed`**.
- R demo system
  - user-friendly interface for running demonstrations of R scripts;
  - e.g. **`demo(marrayPlots)`**, **`demo(affy)`**.
- Bioconductor short courses
  - modular training segments on software and statistical methodology;
  - lectures and computer labs available on WWW for self-instruction.

Vignettes

- Bioconductor has adopted a new documentation paradigm, the vignette.
- A vignette is an executable document consisting of a collection of documentation text and code chunks.
- Vignettes form dynamic, integrated, and reproducible statistical documents that can be automatically updated if either data or analyses are changed.
- Vignettes can be generated using the **`Sweave`** function from the R `tools` package.

Vignettes

- Each Bioconductor package contains at least one vignette, located in the `doc` subdirectory of an installed package and accessible from the help browser.
- Vignettes provide task-oriented descriptions of the package's functionality and can be used interactively.
- Vignettes are available separately from the Bioconductor website or as part of the packages.

Vignettes

- Tools are being developed for managing and using this repository of step-by-step tutorials
  - **Biobase**: **openVignette** – Menu of available vignettes and interface for viewing vignettes (PDF).
  - **tkWidgets**: **vExplorer** – Interactive use of vignettes.
  - **reposTools**.

Sweave

- The Sweave system allows the generation of integrated statistical documents intermixing text, code, and code output (textual and graphical).
- Functions are available in the R `tools` package.
- See `? Sweave` and manual www.ci.tuwien.ac.at/~leisch/Sweave/

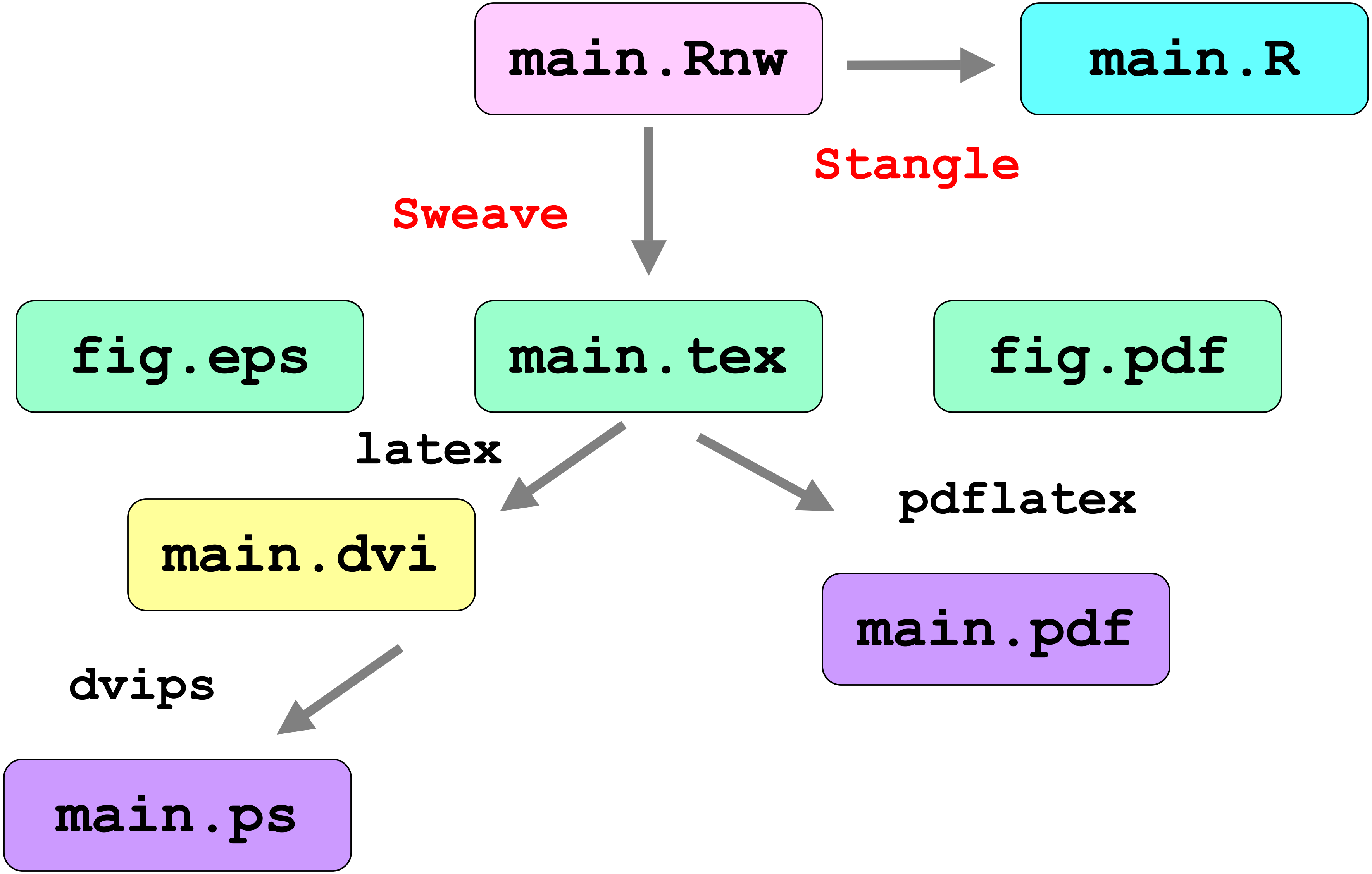
Sweave input

- Input: a text file which consists of a sequence of code and documentation chunks, or segments (noweb file).
  - Documentation chunks
    - start with @
    - can be text in a markup language like LaTeX.
  - Code chunks
    - start with <<*name*>>=
    - can be R or S-Plus code.
  - File extension: **`.rnw`**, **`.Rnw`**, **`.snw`**, **`.Snw`**.

Sweave output

- Output: a single document, e.g., **`.tex`** file or **`.pdf`** file containing
  - the documentation text,
  - the R code,
  - the code output: text and graphs.
- The document can be automatically regenerated whenever the data, code, or documentation text change.
- **`Stangle`** or **`tangleToR`**: extract only the code.

Sweave


main.Rnw
main.R
Stangle
Sweave
fig.eps
main.tex
fig.pdf
latex
pdflatex
main.dvi
main.pdf
dvips
main.ps

Annotation

`annotate` package

- One of the largest challenges in analyzing genomic data is associating the experimental data with the available biological metadata, e.g., sequence, gene annotation, chromosomal maps, literature.

- Bioconductor provides two main packages for this purpose:
  - **`annotate`** (end-user);
  - **`AnnBuilder`** (developer).

WWW resources

- Nucleotide databases: e.g. GenBank.
- Gene databases: e.g. LocusLink, UniGene.
- Protein sequence and structure databases: e.g. SwissProt, Protein DataBank (PDB).
- Literature databases: e.g. PubMed, OMIM.
- Chromosome maps: e.g. NCBI Map Viewer.
- Pathways: e.g. KEGG.
- Entrez is a search and retrieval system that integrates information from databases at NCBI (National Center for Biotechnology Information).

`annotate`: matching IDs

Important tasks

- Associate manufacturers or in-house probe identifiers to other available identifiers.

  E.g.

  Affymetrix IDs → LocusLink LocusID

  Affymetrix IDs → GenBank accession number.
- Associate probes with biological data such as chromosomal position, pathways.
- Associate probes with published literature data via PubMed (need PMID).

`annotate`: matching IDs

Affymetrix identifier HGU95A chips	“41046_s_at”
LocusLink, LocusID	“9203”
GenBank accession #	“X95808”
Gene symbol	“ZNF261”
PubMed, PMID	“10486218” “9205841” “8817323”
Chromosomal location	“X”, “Xq13.1”

Annotation data packages

- The Bioconductor project provides packages that contain only data.
- These data packages are built using **`AnnBuilder`**.
- They can be downloaded from the Bioconductor website and also using **`update.packages`**.*** installDataPackage.
- Data packages contain many different mappings to interesting data.
  - Mappings between Affy IDs and other probe IDs: **`hgu95a`** for HGU95A GeneChip series, also, **`hgu133a`**, **`hu6800`**, **`mgu74a`**, **`rgu34a`**.
  - Affy CDF data packages.
- The packages are updated and expanded regularly as updated and new data become available.

`annotate`: matching IDs

- Much of what `annotate` does relies on matching symbols.
- This is basically the role of a hash table in most programming languages.
- In R, we rely on environments.
- The annotation data packages provide R environment objects containing key and value pairs for the mappings between two sets of probe identifiers.
- Keys can be accessed using the R `ls` function.
- Matching values in different environments can be accessed using the `get` or `multiget` functions.

annotate: matching IDs

```
> library(hgu95a)
> get("41046_s_at", env = hgu95aACCNUM)
[1] "X95808”
> get("41046_s_at", env = hgu95aLOCUSID)
[1] "9203”
> get("41046_s_at", env = hgu95aSYMBOL)
[1] "ZNF261"
> get("41046_s_at", env = hgu95aGENENAME)
[1] "zinc finger protein 261"
> get("41046_s_at", env = hgu95aSUMFUNC)
[1] "Contains a putative zinc-binding
  motif (MYM)|Proteome"
> get("41046_s_at", env = hgu95aUNIGENE)
[1] "Hs.9568"
```

annotate: matching IDs

```
> get("41046_s_at", env = hgu95aCHR)
[1] "X"
> get("41046_s_at", env = hgu95aCHRLOC)
[1] "66457019@X"
> get("41046_s_at", env = hgu95aCHRORI)
[1] "-@X"
> get("41046_s_at", env = hgu95aMAP)
[1] "Xq13.1”
> get("41046_s_at", env = hgu95aPMID)
[1] "10486218" "9205841"  "8817323"
> get("41046_s_at", env = hgu95aGO)
[1] "GO:0003677" "GO:0007275"
```

`annotate`: matching IDs

- Instead of relying on the general R functions for environments, new user-friendly functions have been written for accessing and working with specific identifiers.

- E.g. **`getGO`**, **`getGOdesc`**, **`getLL`**, **`getPMID`**, **`getSYMBOL`**.

annotate: matching IDs

```
> getSYMBOL("41046_s_at",data="hgu95a")
  41046_s_at
    "ZNF261"
> gg<- getGO("41046_s_at",data="hgu95a")
> getGOdesc(gg, "MF")
  $"c("GO:0003677", "GO:0007275")"
  [1] "DNA binding"
> getLL("41046_s_at",data="hgu95a")
  41046_s_at
        9203
> getPMID("41046_s_at",data="hgu95a")
  $"41046_s_at"
  [1] 10486218  9205841  8817323
```

`annotate`: querying databases

The `annotate` package provides tools for

- Searching and processing information from various WWW biological databases
  - GenBank,
  - LocusLink,
  - PubMed.
- Regular expression searching of PubMed abstracts.
- Generating nice HTML reports of analyses, with links to biological databases.

`annotate`: WWW queries

- Functions for querying WWW databases from R rely on the **`browseURL`** function

  ```
  browseURL("www.r-project.org")
  ```

- The **`XML`** package is used to parse query results.

annotate: querying GenBank

www.ncbi.nlm.nih.gov/Genbank/index.html

- Given a vector of GenBank accession numbers or NCBI UIDs, the **genbank** function
  - opens a browser at the URLs for the corresponding GenBank queries;
  - returns an **XMLdoc** object with the same data.

```
genbank("X95808",disp="browser")
```

http://www.ncbi.nih.gov/entrez/query.fcgi?tool=bioconductor&cmd=Search&db=Nucleotide&term=X95808

```
genbank(1430782,disp="data",
    type="uid")
```

annotate: querying LocusLink

www.ncbi.nlm.nih.gov/LocusLink/

- **locuslinkByID**: given one or more LocusIDs, the browser is opened at the URL corresponding to the first gene.

  ```
  locuslinkByID("9203")
  ```

  http://www.ncbi.nih.gov/LocusLink/LocRpt.cgi?l=9203

- **locuslinkQuery**: given a search string, the results of the LocusLink query are displayed in the browser.

  ```
  locuslinkQuery("zinc finger")
  ```

  http://www.ncbi.nih.gov/LocusLink/list.cgi?Q=zinc finger&ORG=Hs&V=0

`annotate`: querying PubMed

www.ncbi.nlm.nih.gov

- For any gene there is often a large amount of data available from PubMed.
- The `annotate` package provides the following tools for interacting with PubMed
  - `pubMedAbst`: a class structure for PubMed abstracts in R.
  - `pubmed`: the basic engine for talking to PubMed.

annotate: pubMedAbst class

Class structure for storing and processing PubMed abstracts in R

- **pmid**
- **authors**
- **abstText**
- **articleTitle**
- **journal**
- **pubDate**
- **abstUrl**

`annotate`: high-level tools for querying PubMed

- `pm.getabst`: download the specified PubMed abstracts (stored in XML) and create a list of `pubMedAbst` objects.
- `pm.titles`: extract the titles from a list of PubMed abstracts.
- `pm.abstGrep`: regular expression matching on the abstracts.

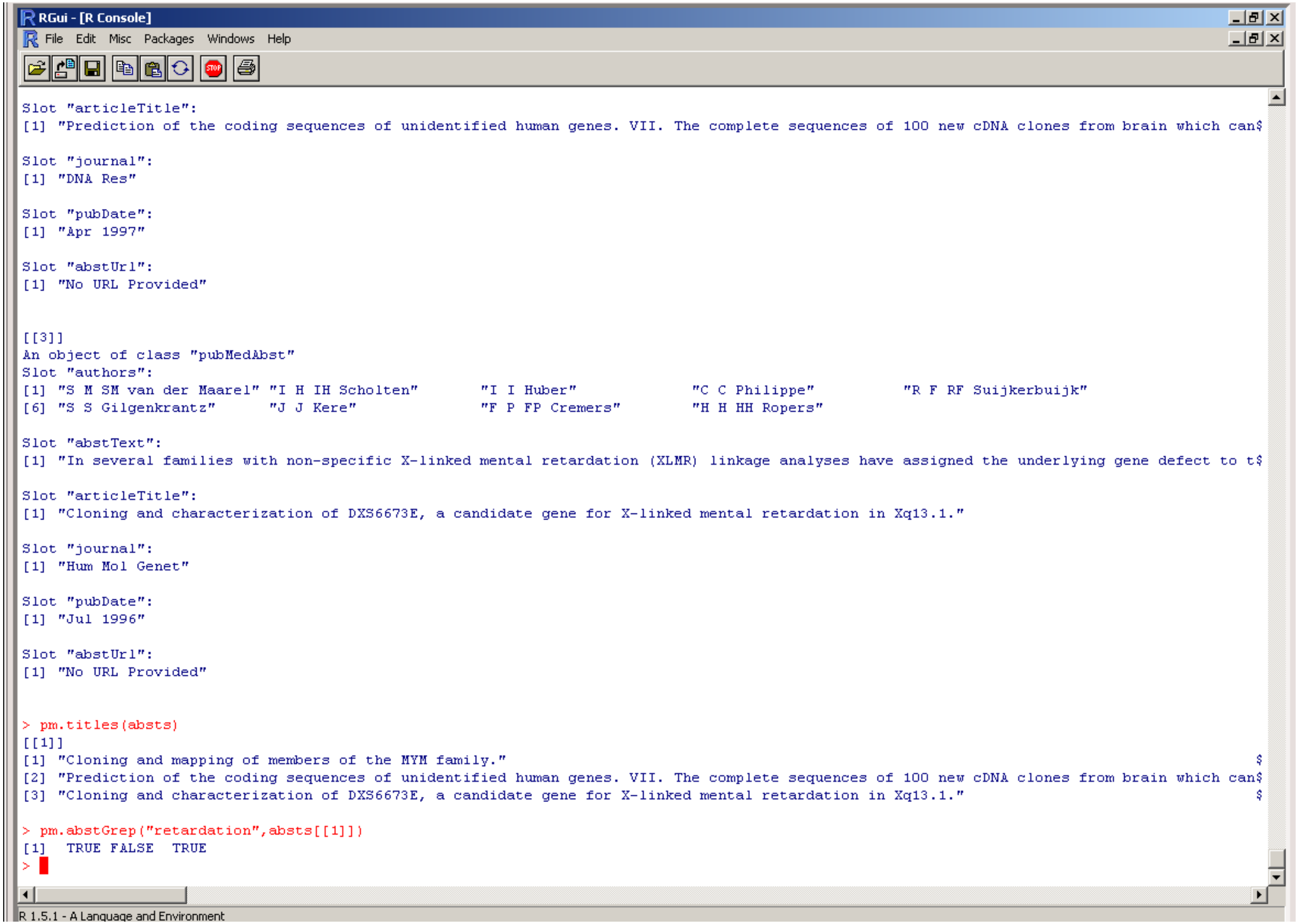
annotate: PubMed example

```
pmid <-get("41046_s_at", env=hgu95aPMID)
pubmed(pmid, disp=“browser”)
```

http://www.ncbi.nih.gov/entrez/query.fcgi?tool=bioconductor&cmd=Retrieve&db=PubMed&list_uids=10486218%2c9205841%2c8817323

```
absts <- pm.getabst(“41046_s_at”,
  base=“hgu95a”)
pm.titles(absts)
pm.abstGrep("retardation",absts[[1]])
```

annotate: PubMed example

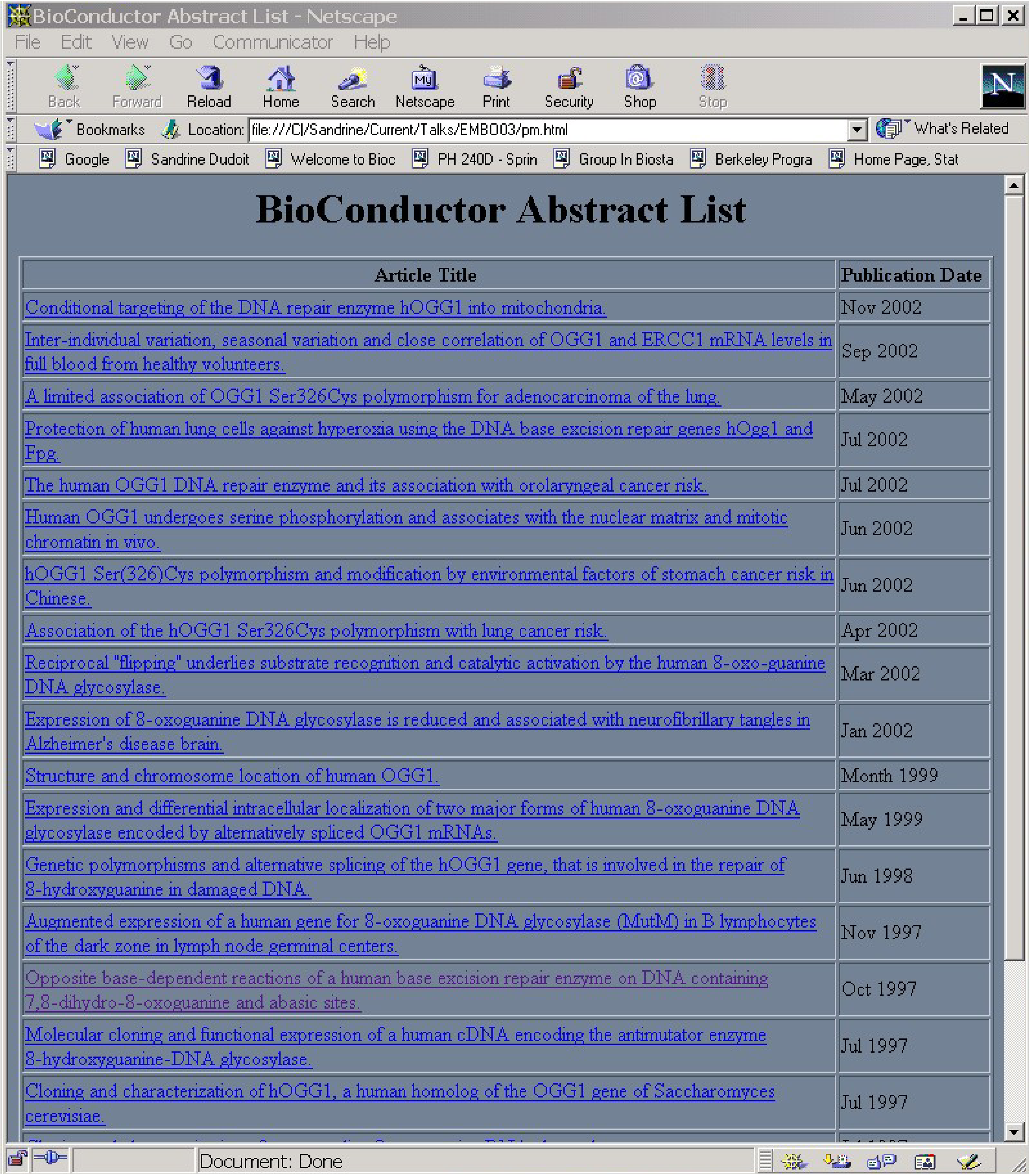


```
Slot "articleTitle":
[1] "Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can$

Slot "journal":
[1] "DNA Res"

Slot "pubDate":
[1] "Apr 1997"

Slot "abstUrl":
[1] "No URL Provided"


[[3]]
An object of class "pubMedAbst"
Slot "authors":
[1] "S M SM van der Maarel" "I H IH Scholten"       "I I Huber"             "C C Philippe"          "R F RF Suijkerbuijk"
[6] "S S Gilgenkrantz"      "J J Kere"              "F P FP Cremers"        "H H HH Ropers"

Slot "abstText":
[1] "In several families with non-specific X-linked mental retardation (XLMR) linkage analyses have assigned the underlying gene defect to t$

Slot "articleTitle":
[1] "Cloning and characterization of DXS6673E, a candidate gene for X-linked mental retardation in Xq13.1."

Slot "journal":
[1] "Hum Mol Genet"

Slot "pubDate":
[1] "Jul 1996"

Slot "abstUrl":
[1] "No URL Provided"


> pm.titles(absts)
[[1]]
[1] "Cloning and mapping of members of the MYM family."                                                                                   $
[2] "Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can$
[3] "Cloning and characterization of DXS6673E, a candidate gene for X-linked mental retardation in Xq13.1."                                 $

> pm.abstGrep("retardation",absts[[1]])
[1]  TRUE FALSE  TRUE
> 
```

`annotate`: PubMed HTML report

- The new function **`pmAbst2HTML`** takes a list of **`pubMedAbst`** objects and generates an HTML report with the titles of the abstracts and links to their full page on PubMed.

```
pmAbst2HTML(absts[[1]],filename="pm.htm
l")
```


BioConductor Abstract List - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks
What's Related
Google Sandrine Dudoit Welcome to Bioc PH 240D - Sprin Group In Biosta Berkeley Progra Home Page, Stat

Location: file:///C|/Sandrine/Current/Talks/EMBO03/pm.html

BioConductor Abstract List

Article Title	Publication Date
Conditional targeting of the DNA repair enzyme hOGG1 into mitochondria.	Nov 2002
Inter-individual variation, seasonal variation and close correlation of OGG1 and ERCC1 mRNA levels in full blood from healthy volunteers.	Sep 2002
A limited association of OGG1 Ser326Cys polymorphism for adenocarcinoma of the lung.	May 2002
Protection of human lung cells against hyperoxia using the DNA base excision repair genes hOgg1 and Fpg.	Jul 2002
The human OGG1 DNA repair enzyme and its association with orolaryngeal cancer risk.	Jul 2002
Human OGG1 undergoes serine phosphorylation and associates with the nuclear matrix and mitotic chromatin in vivo.	Jun 2002
hOGG1 Ser(326)Cys polymorphism and modification by environmental factors of stomach cancer risk in Chinese.	Jun 2002
Association of the hOGG1 Ser326Cys polymorphism with lung cancer risk.	Apr 2002
Reciprocal "flipping" underlies substrate recognition and catalytic activation by the human 8-oxo-guanine DNA glycosylase.	Mar 2002
Expression of 8-oxoguanine DNA glycosylase is reduced and associated with neurofibrillary tangles in Alzheimer's disease brain.	Jan 2002
Structure and chromosome location of human OGG1.	Month 1999
Expression and differential intracellular localization of two major forms of human 8-oxoguanine DNA glycosylase encoded by alternatively spliced OGG1 mRNAs.	May 1999
Genetic polymorphisms and alternative splicing of the hOGG1 gene, that is involved in the repair of 8-hydroxyguanine in damaged DNA.	Jun 1998
Augmented expression of a human gene for 8-oxoguanine DNA glycosylase (MutM) in B lymphocytes of the dark zone in lymph node germinal centers.	Nov 1997
Opposite base-dependent reactions of a human base excision repair enzyme on DNA containing 7,8-dihydro-8-oxoguanine and abasic sites.	Oct 1997
Molecular cloning and functional expression of a human cDNA encoding the antimutator enzyme 8-hydroxyguanine-DNA glycosylase.	Jul 1997
Cloning and characterization of hOGG1, a human homolog of the OGG1 gene of Saccharomyces cerevisiae.	Jul 1997

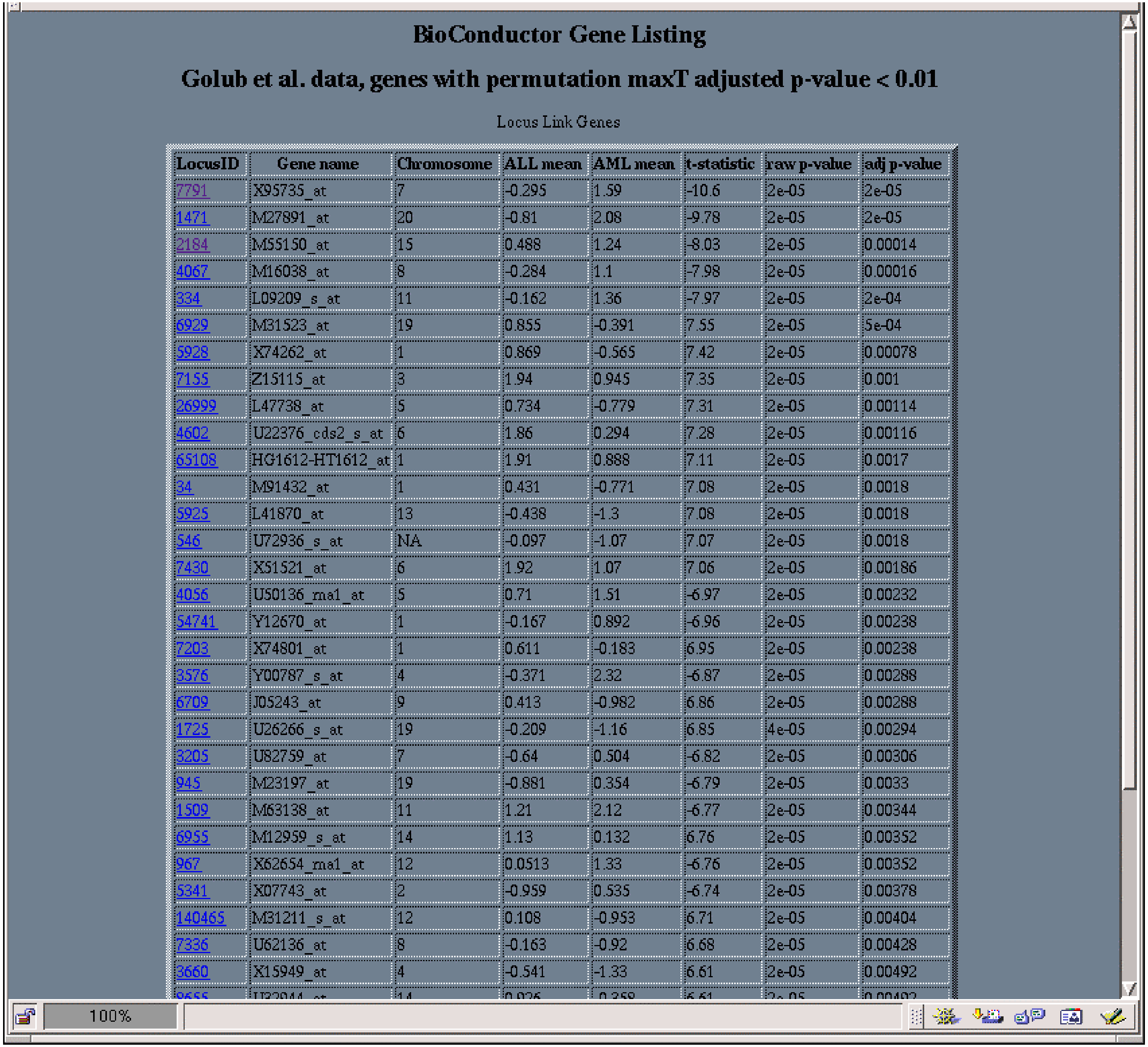
Document: Done

pmAbst2html
function from
annotate package

pm.html

`annotate`: analysis reports

- A simple interface, `ll.htmlpage`, can be used to generate an HTML report of analysis results.
- The page consists of a table with one row per gene, with links to LocusLink.
- Entries can include various gene identifiers and statistics.



BioConductor Gene Listing

Golub et al. data, genes with permutation maxT adjusted p-value < 0.01

Locus Link Genes

LocusID	Gene name	Chromosome	ALL mean	AML mean	t-statistic	raw p-value	adj p-value
7791	X95735_at	7	-0.295	1.59	-10.6	2e-05	2e-05
1471	M27891_at	20	-0.81	2.08	-9.78	2e-05	2e-05
2184	M55150_at	15	0.488	1.24	-8.03	2e-05	0.00014
4067	M16038_at	8	-0.284	1.1	-7.98	2e-05	0.00016
334	L09209_s_at	11	-0.162	1.36	-7.97	2e-05	2e-04
6929	M31523_at	19	0.855	-0.391	7.55	2e-05	5e-04
5928	X74262_at	1	0.869	-0.565	7.42	2e-05	0.00078
7155	Z15115_at	3	1.94	0.945	7.35	2e-05	0.001
26999	L47738_at	5	0.734	-0.779	7.31	2e-05	0.00114
4602	U22376_cds2_s_at	6	1.86	0.294	7.28	2e-05	0.00116
65108	HG1612-HT1612_at	1	1.91	0.888	7.11	2e-05	0.0017
34	M91432_at	1	0.431	-0.771	7.08	2e-05	0.0018
5925	L41870_at	13	-0.438	-1.3	7.08	2e-05	0.0018
546	U72936_s_at	NA	-0.097	-1.07	7.07	2e-05	0.0018
7430	X51521_at	6	1.92	1.07	7.06	2e-05	0.00186
4056	U50136_rna1_at	5	0.71	1.51	-6.97	2e-05	0.00232
54741	Y12670_at	1	-0.167	0.892	-6.96	2e-05	0.00238
7203	X74801_at	1	0.611	-0.183	6.95	2e-05	0.00238
3576	Y00787_s_at	4	-0.371	2.32	-6.87	2e-05	0.00288
6709	J05243_at	9	0.413	-0.982	6.86	2e-05	0.00288
1725	U26266_s_at	19	-0.209	-1.16	6.85	4e-05	0.00294
3205	U82759_at	7	-0.64	0.504	-6.82	2e-05	0.00306
945	M23197_at	19	-0.881	0.354	-6.79	2e-05	0.0033
1509	M63138_at	11	1.21	2.12	-6.77	2e-05	0.00344
6955	M12959_s_at	14	1.13	0.132	6.76	2e-05	0.00352
967	X62654_rna1_at	12	0.0513	1.33	-6.76	2e-05	0.00352
5341	X07743_at	2	-0.959	0.535	-6.74	2e-05	0.00378
140465	M31211_s_at	12	0.108	-0.953	6.71	2e-05	0.00404
7336	U62136_at	8	-0.163	-0.92	6.68	2e-05	0.00428
3660	X15949_at	4	-0.541	-1.33	6.61	2e-05	0.00492
9655	U32944_at	14	0.026	0.358	6.61	2e-05	0.00492

ll.htmlpage function from **annotate** package

genelist.html

`annotate`: `chromLoc` class

Location information for one gene

- **`chrom`**: chromosome name.
- **`position`**: starting position of the gene in bp.
- **`strand`**: chromosome strand +/-.

`annotate`: `chromLocation` class

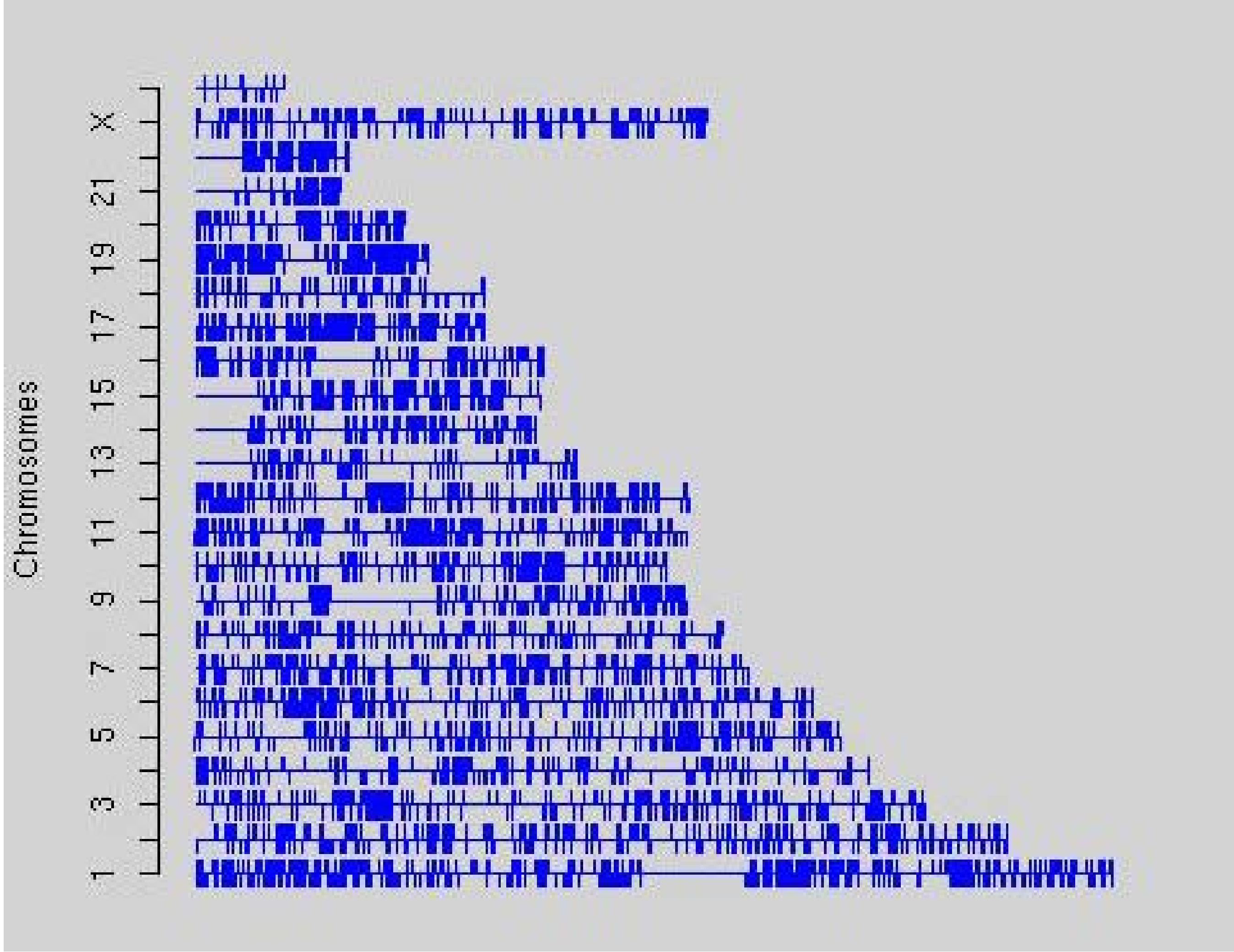
Location information for a set of genes

- **`species:`** species that the genes correspond to.
- **`datSource:`** source of the gene location data.
- **`nChrom:`** number of chromosomes for the species.
- **`chromNames:`** chromosome names.
- **`chromLocs:`** starting position of the genes in bp.
- **`chromLengths:`** length of each chromosome in bp.
- **`geneToChrom:`** hash table translating gene IDs to location.

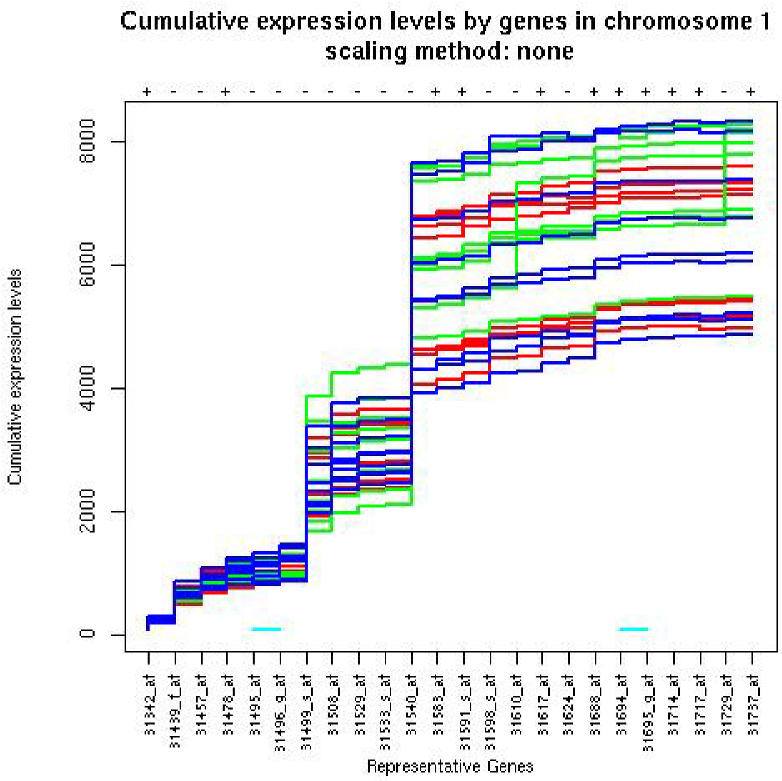
Function **`buildChromClass`**.

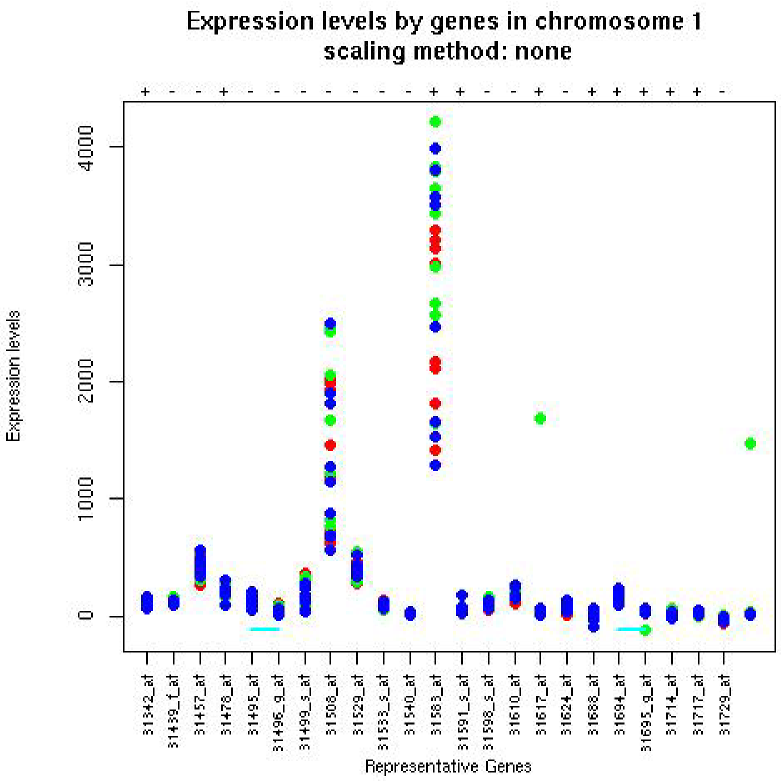
Visualization

geneplotter: cPlot

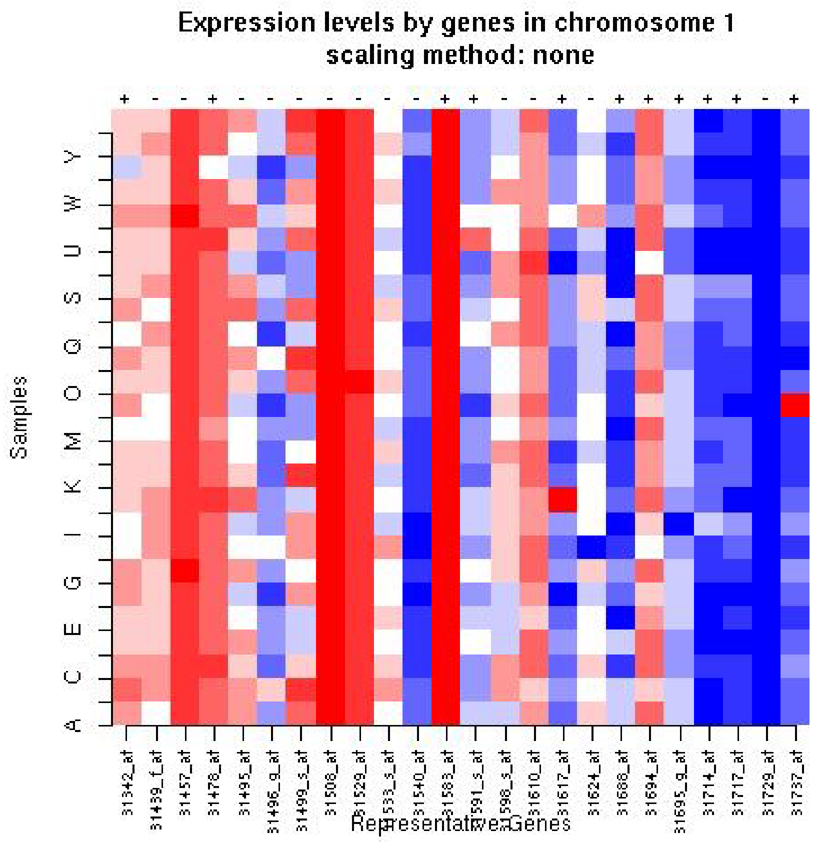


geneplotter: alongChrom

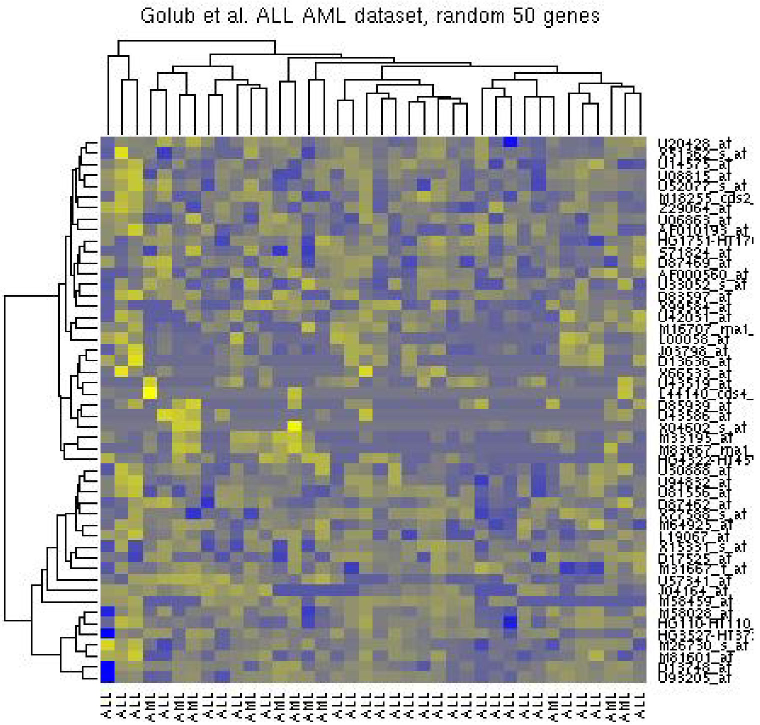


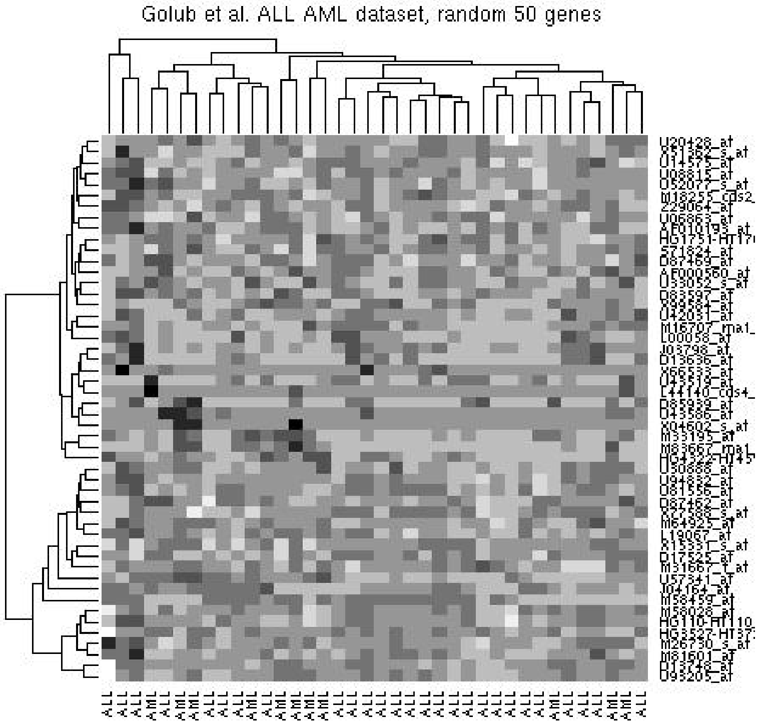


geneplotter: alongChrom



mva: heatmap


Golub et al. ALL AML dataset, random 50 genes
U20428_at
X51362_s_at
U14575_at
U08815_at
U52077_s_at
M18255_cds2_
Z29064_at
U06863_at
AF010193_at
HG1751-HT17
S71824_at
D87469_at
AF000560_at
U33052_s_at
D83597_at
X99584_at
U42031_at
M16707_rna1_
L00058_at
J03798_at
D13636_at
X66533_at
U43519_at
L44140_cds4_
D85939_at
D43586_at
X04602_s_at
M33195_at
M83667_rna1_
HG4322-HT45
U30888_at
U94832_at
U81556_at
D87462_at
X77588_s_at
M64925_at
L19067_at
X15331_s_at
D17525_at
M31667_f_at
U57341_at
J04164_at
M58459_at
M58028_at
HG110-HT110
HG3527-HT37
M26730_s_at
M81601_at
D13748_at
D93205_at
ALL
ALL
ALL
AML
ALL
AML
AML
ALL
ALL
ALL
AML
ALL
AML
AML
AML
AML
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
AML
ALL
ALL
ALL
AML
AML
ALL


Golub et al. ALL AML dataset, random 50 genes
U20428_at
X51362_s_at
U14575_at
U08815_at
U52077_s_at
M18255_cds2_
Z29064_at
U06863_at
AF010193_at
HG1751-HT17
S71824_at
D87469_at
AF000560_at
U33052_s_at
D83597_at
X99584_at
U42031_at
M16707_rna1_
L00058_at
J03798_at
D13636_at
X66533_at
U43519_at
L44140_cds4_
D85939_at
D43586_at
X04602_s_at
M33195_at
M83667_rna1_
HG4322-HT45
U30888_at
U94832_at
U81556_at
D87462_at
X77588_s_at
M64925_at
L19067_at
X15331_s_at
D17525_at
M31667_f_at
U57341_at
J04164_at
M58459_at
M58028_at
HG110-HT110
HG3527-HT37
M26730_s_at
M81601_at
D13748_at
D93205_at
ALL
ALL
ALL
AML
ALL
AML
AML
ALL
ALL
ALL
AML
ALL
AML
AML
AML
AML
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
ALL
AML
ALL
ALL
ALL
AML
AML
ALL