Lecture 8

An introduction to Microarray

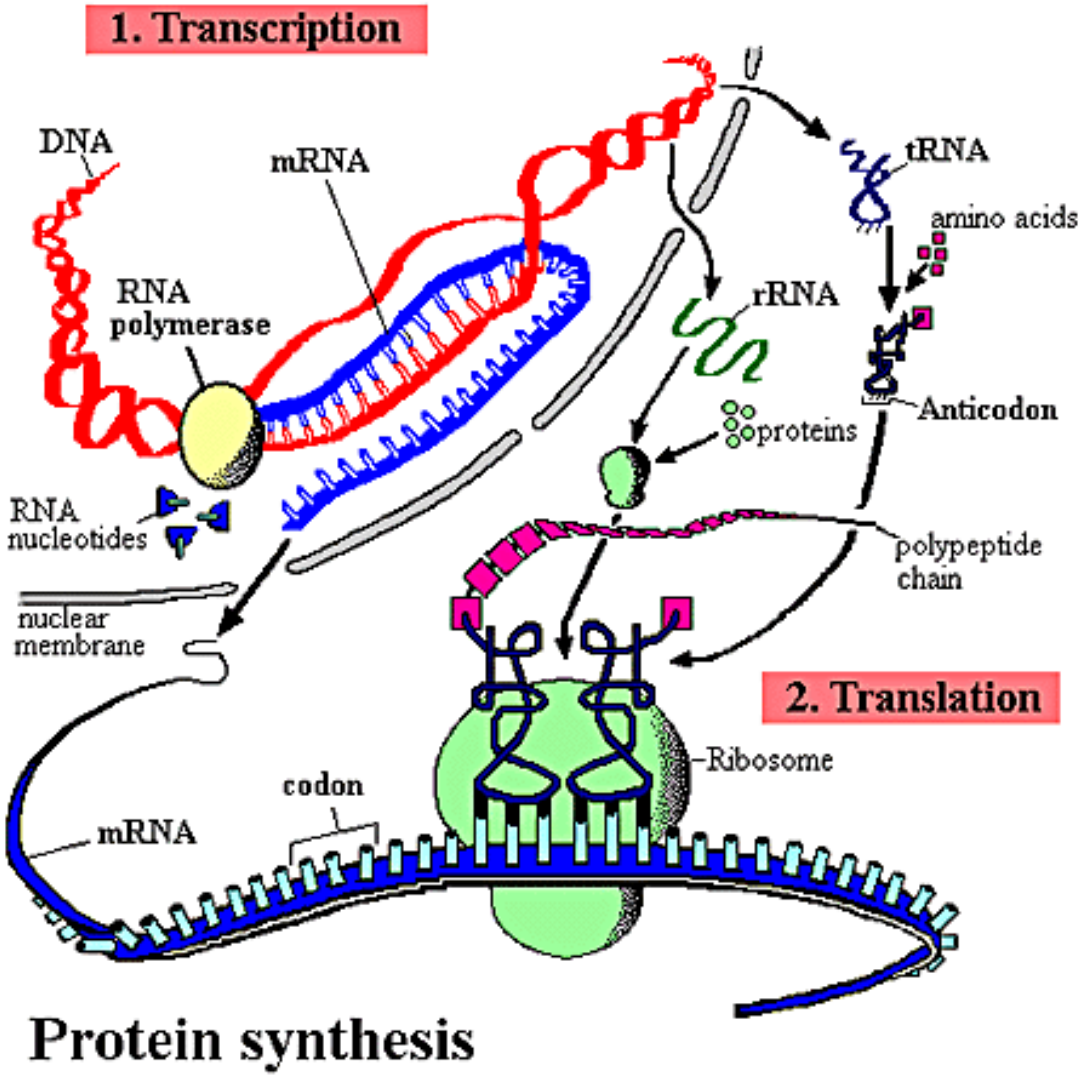
MCB 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

Outline

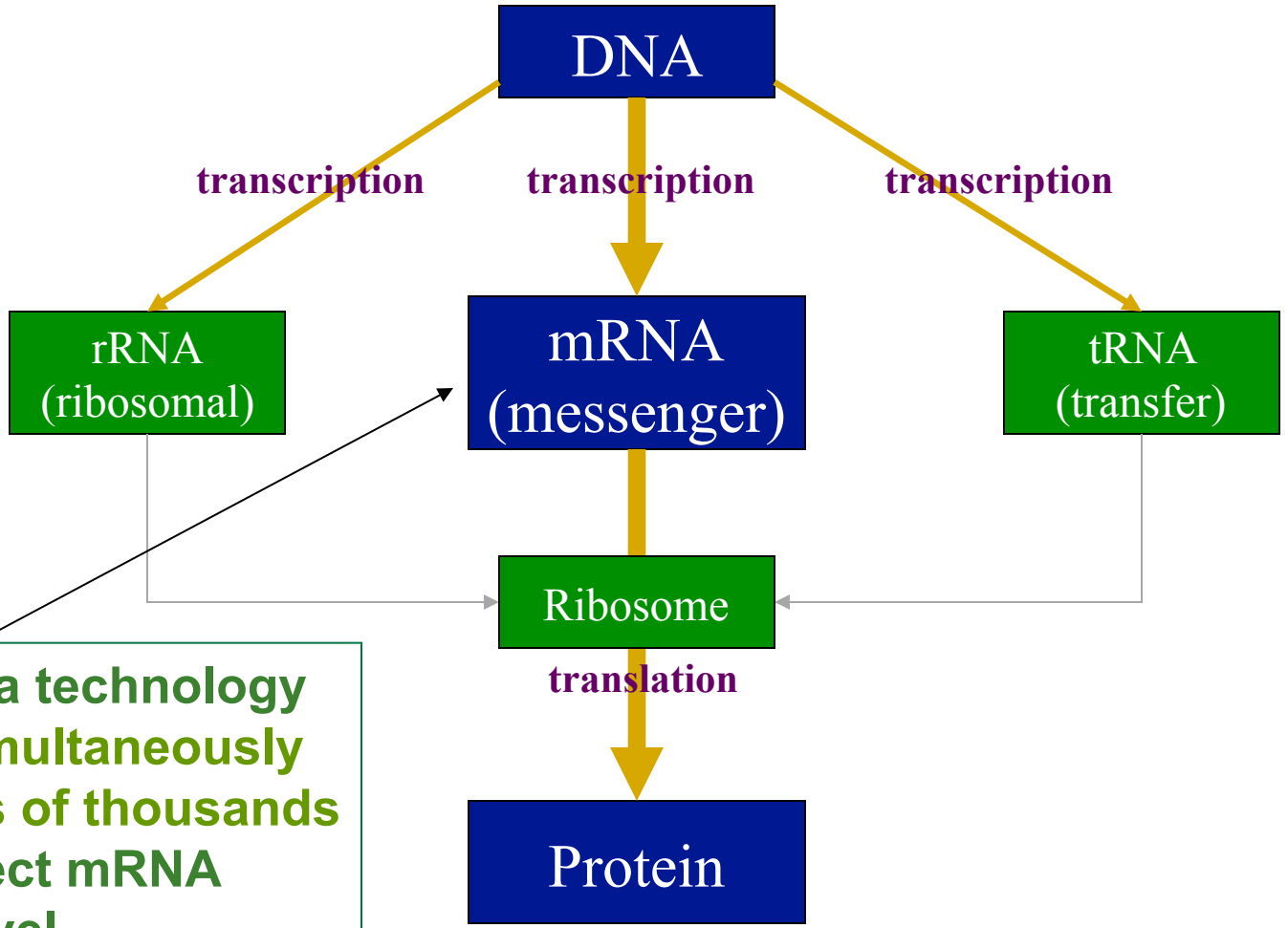
- Why microarray?
- Affymetrix microarray experiment
- cDNA microarray experiment
- Other platforms and comparison
- Microarray experiment

Review



Protein synthesis

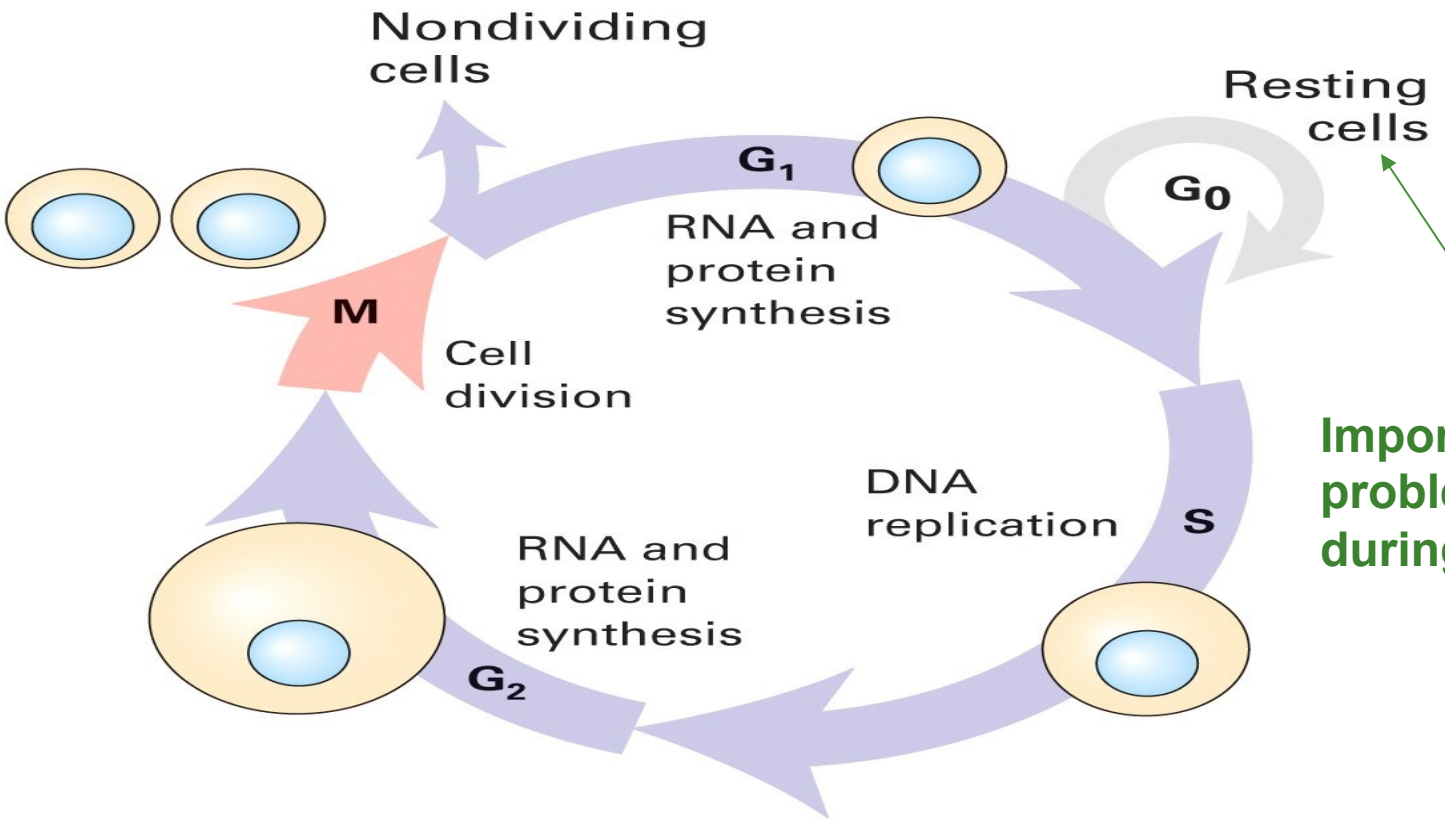
The central dogma of molecular biology:



Microarray is a technology to globaly (simultaneously detecting tens of thousands of genes) detect mRNA expression level.

Why detect expression level of mRNA?

Cell cycle



Important to repair problematic mutations during cell division.

Cancer cells are malignant cells who don't die but reproduce rapidly instead.

Why detect expression level of protein or mRNA?

Early diagnosis of a disease:

If mechanism known, detecting expression level can help identifying cancer patients (e.g. unusual p53 or Kras expression activity).

Exploratory:

In general, microarray can help identify candidate genes that contribute to tumor progression and propose hypothesis of the underlying genetic network.

What is microarray?

- multiplex technology used in molecular biology and in medicine.
- consist of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10^{-12} moles) of a specific DNA sequence, known as *probes*
- *Probes* are used to hybridize to cDNA or cRNA samples (called *targets*)

Old techniques and their limitations

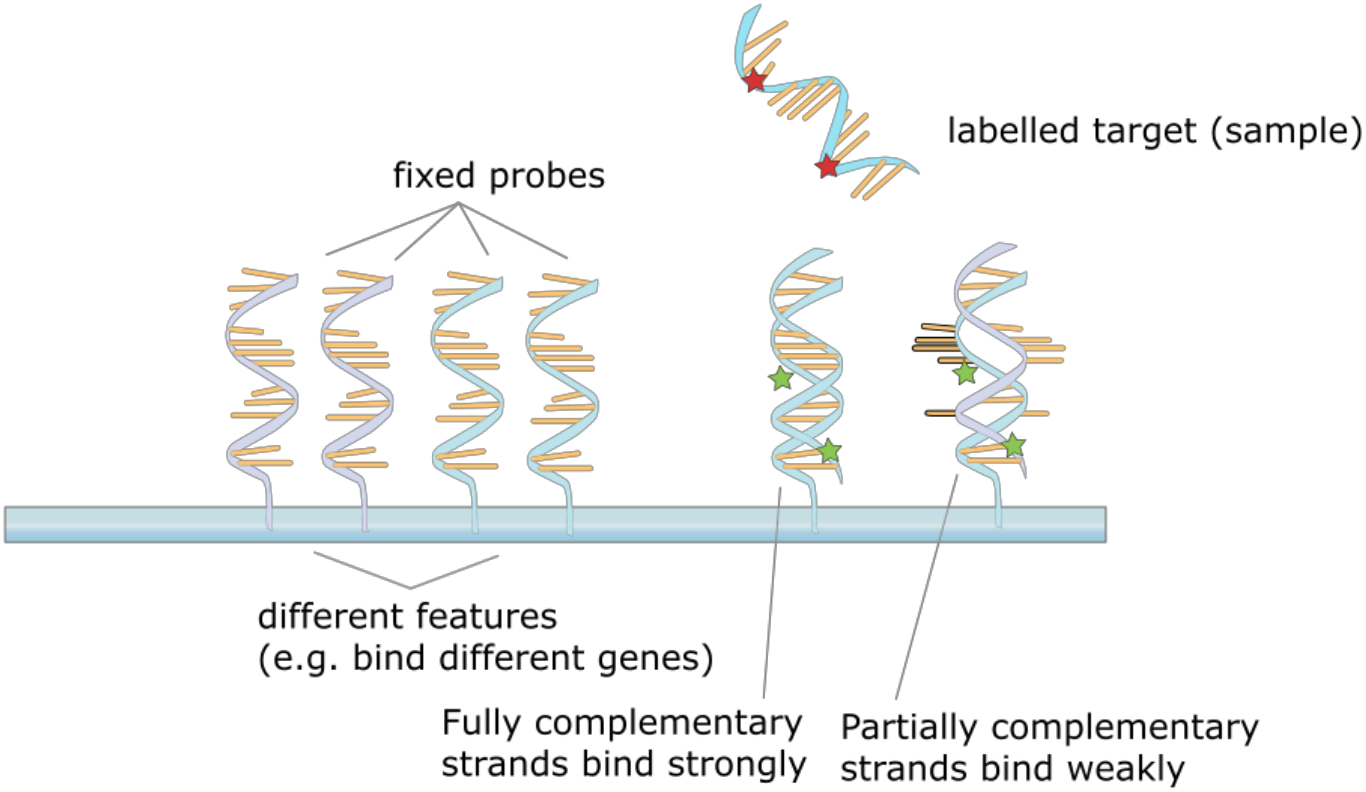
1. Northern blot, RT-PCR, …

2. Can only detect up to dozens of genes.
   (gene-by-gene analysis)

3. Labor intensive

4. Need to know the target sequences. For RT-PCR, at least need to know the primer to start the PCR.

Hybridization

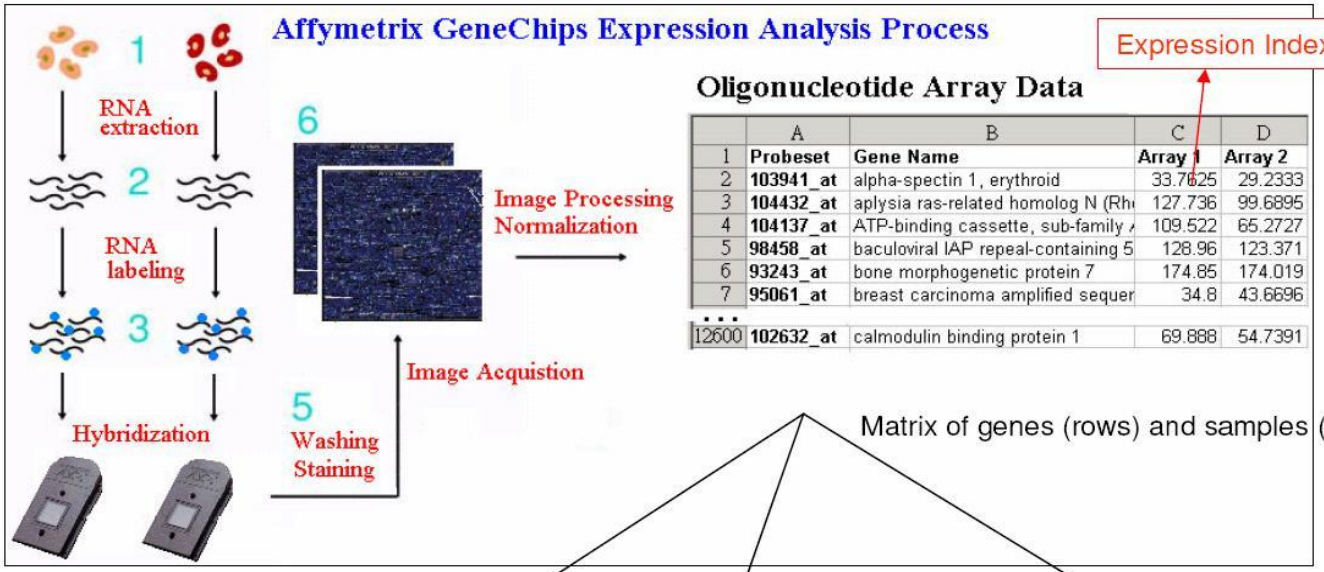


The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs.

Affymetrix GeneChip

Overview of the Affymetrix GeneChip experiment

Affymetrix array: from experiments to analysis



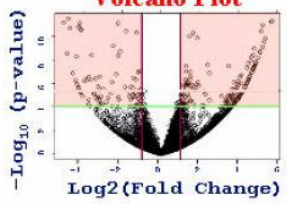
	A	B	C	D
1	Probeset	Gene Name	Array 1	Array 2
2	103941_at	alpha-spectin 1, erythroid	33.7625	29.2333
3	104432_at	aplysia ras-related homolog N (Rh	127.736	99.6895
4	104137_at	ATP-binding cassette, sub-family	109.522	65.2727
5	98458_at	baculoviral IAP repeal-containing 5	128.96	123.371
6	93243_at	bone morphogenetic protein 7	174.85	174.019
7	95061_at	breast carcinoma amplified sequer	34.8	43.6696
...				
12600	102632_at	calmodulin binding protein 1	69.888	54.7391

Discovery of differentially expressed genes

Parametric : t-test

Non-parametric : Wilcoxon, Mann-Whitney test

Volcano Plot

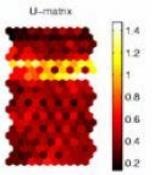


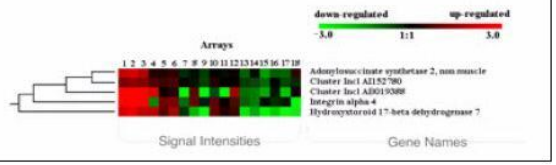
Unsupervised: clustering

Hierarchical clustering

K-means clustering

Self-organizing maps

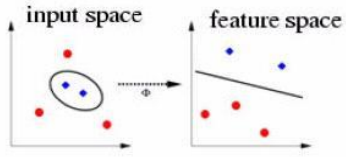




Supervised: classification

- **Linear discriminants**
- **Decision trees**
- **Support vector machines**

Support Vector Classifiers



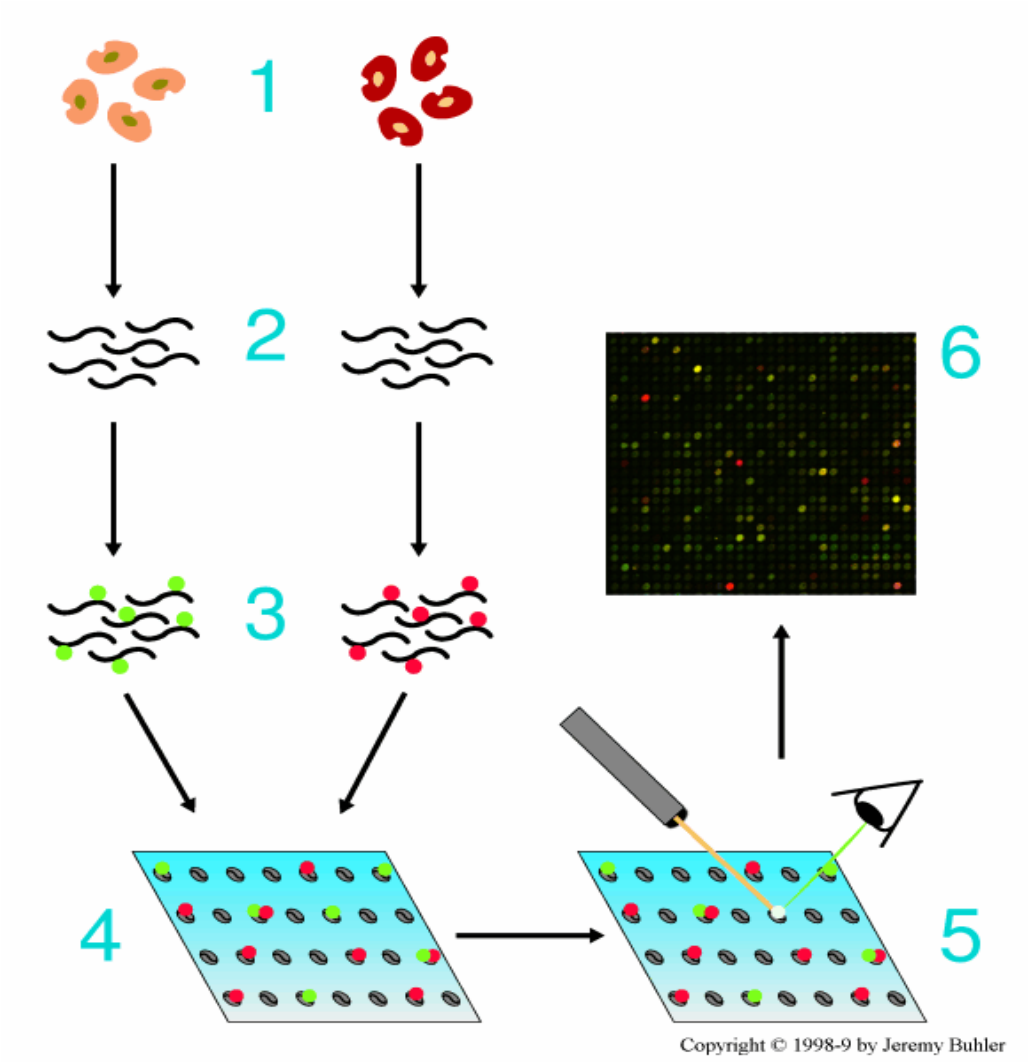
Boser, Guyon, and Vapnik (1992)

Chips

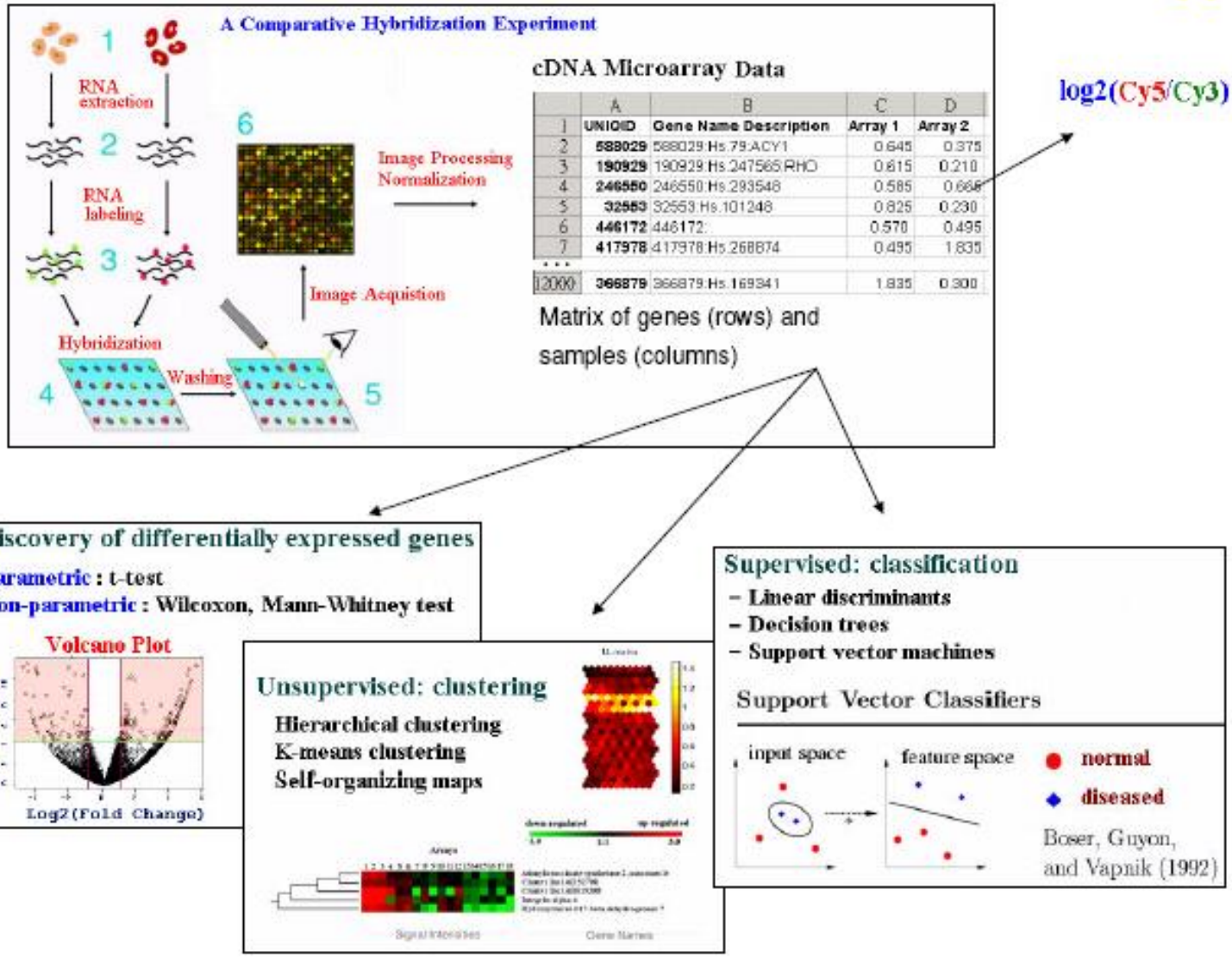
	HG-U95	HG-U133 Set	HG-U133 Plus 2.0 Array	Human Exon 1.0 ST Array
sequence source	Build 95 UniGene database	Build 133 UniGene database	Build 133 UniGene database	UCSC hg16
# of probes	~16	11	11	Average four probes per exon and 40 probes per gene
# of arrays	5	2	**1**	**1**
# of transcripts	~54000 genes	~33,000 genes	~38500 genes	>1 million exons

Spotted cDNA microarray

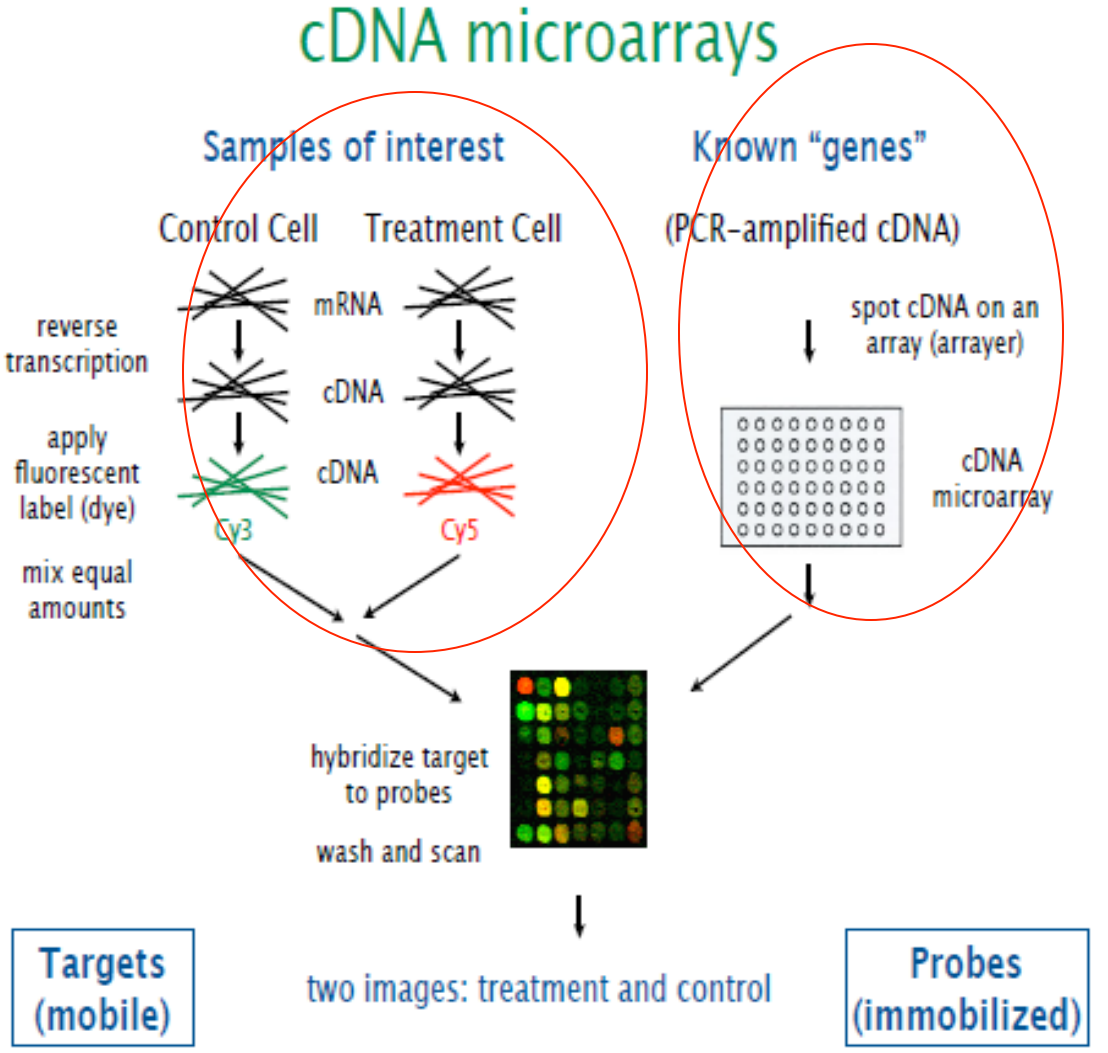
Overview of cDNA array experiment



cDNA array: from experiments to analysis



cDNA array: details



Adapted from "The Science Creative Quarterly", http://www.scq.ubc.ca/

An image example

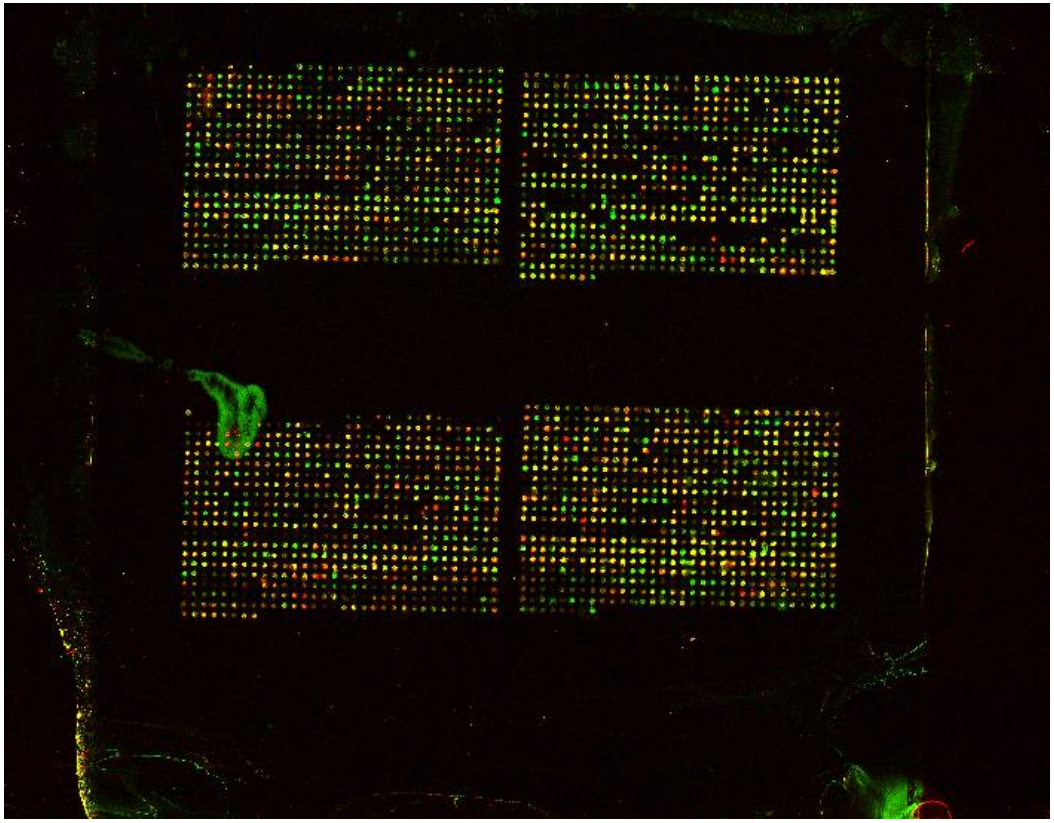


Image analysis is more difficult than Affy array. The probes are spotted by robot instead of synthesized and the exact physical location is not known.

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

Comparison of cDNA array and GeneChip

	cDNA	GeneChip
Probe preparation	Probes are cDNA fragments, usually amplified by PCR and spotted by robot.	Probes are short oligos synthesized using a photolithographic approach.
colors	Two-color (measures relative intensity)	One-color (measures absolute intensity)
Gene representation	One probe per gene	11-16 probe pairs per gene
Probe length	Long, varying lengths (hundreds to 1K bp)	25-mers
Density	Maximum of ~15000 probes.	38500 genes * 11 probes = 423500 probes

More details can be found at:
http://grf.lshtm.ac.uk/microarrayoverview.htm

Advantage and disadvantage of cDNA array and GeneChip

cDNA microarray	Affymetrix GeneChip
The data can be noisy and with variable quality	Specific and sensitive. Result very reproducible.
Cross(non-specific) hybridization can often happen.	Hybridization more specific.
May need a RNA amplification procedure.	Can use small amount of RNA.
More difficulty in image analysis.	Image analysis and intensity extraction is easier.
Need to search the database for gene annotation.	More widely used. Better quality of gene annotation.
Cheap. (both initial cost and per slide cost)	Expensive (~$400 per array+labeling and hybridization)
Can be custom made for special species.	Only popular species are available
Do not need to know the exact DNA sequence.	Need the DNA sequence for probe selection.

Comparison with more …

	cDNA	GeneChip	Codelink	Agilent
Probe preparation	Probes are cDNA fragments, usually amplified by PCR and spotted by robot.	Probes are short oligos synthesized using a photolithographic approach.	3-D aqueous gel matrix	Probes are printed by Inkjet technology from HP
colors	Two-color (measures relative intensity)	One-color (measures absolute intensity)	One-color	One- or two-color
Gene representation	One probe per gene	11-16 probe pairs per gene	One probe per gene	One probe per gene
Probe length	Long, varying lengths (hundreds to 1K bp)	25-mers	30-mers	60-mers
Density	Maximum of ~15000 probes.	38500 genes * 11 probes = 423500	~57000	~22000 probes
Manufacturer	Stanford and many labs.	Affymetrix company	GE company	Agilent company

Mechanisms in microarray

Important mechanisms that make microarray work:

1. Reverse transcription: mRNA => cDNA. This is usually also the step to label dyes.

   (Protein can not be reverse translated to mRNA or to another form. So difficult to label dyes.)

2. Double strand binding of complimentary DNA sequences.

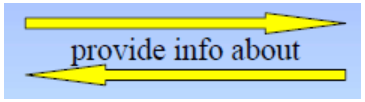
   (Protein does not enjoy such a good property; there are 20 amino acids without complementary binding)

Questions and answers: a point of view

Biological questions first, then statistical methods (design, analysis) and thinking, leading to tentative answers, together with an assessment of the uncertainty in those answers

Expression microarrays: - Underlying assumption and concepts

- Measuring relative changes in levels of specific mRNAs provide information about what's going on in the cells from which the mRNA came.

Samples 

Genes

- A gene expression profile is a molecular phenotype of a cell in a specific state.

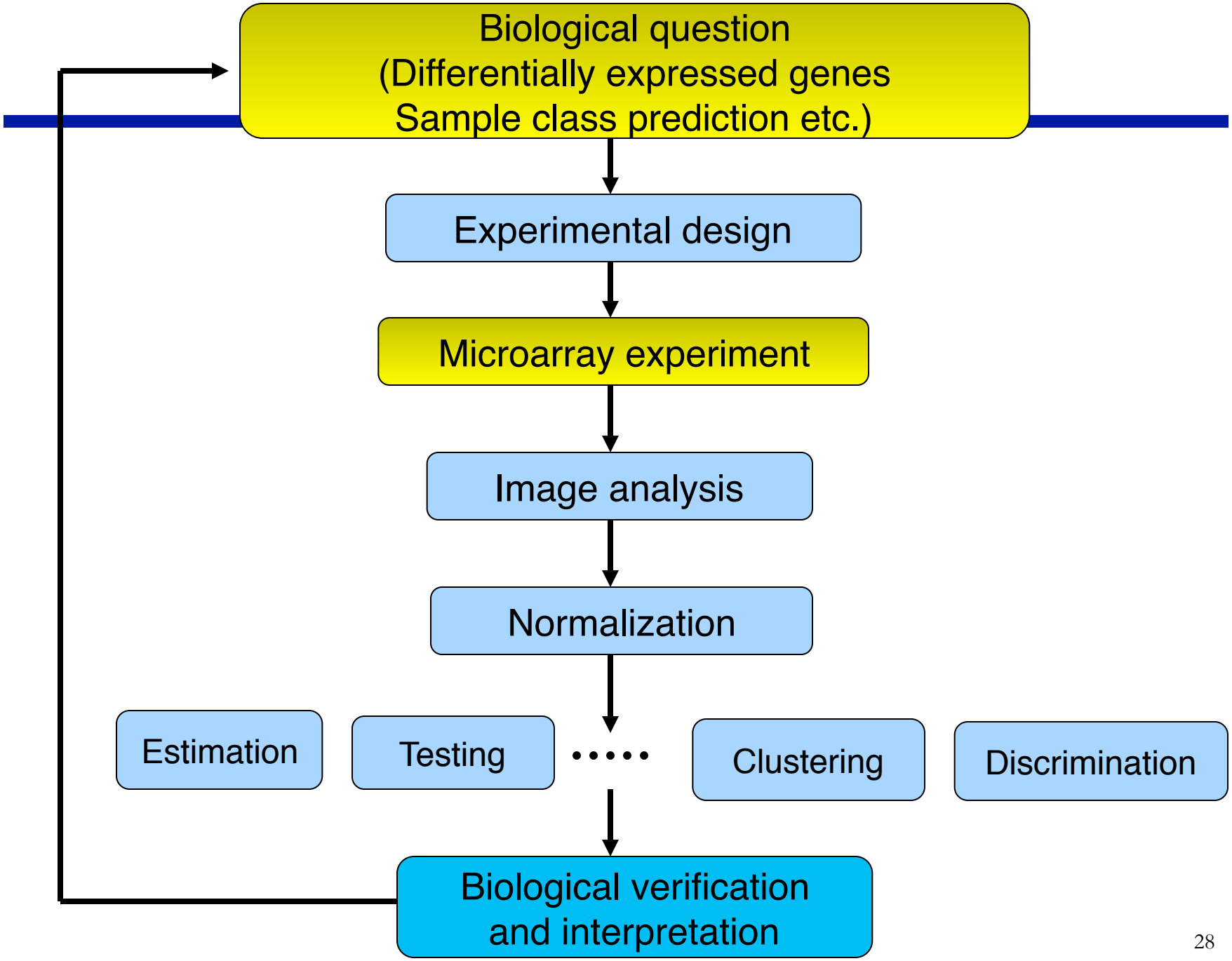
Experimental design: Most important question

- Why are you doing this experiment?
  (Be as specific as possible.)

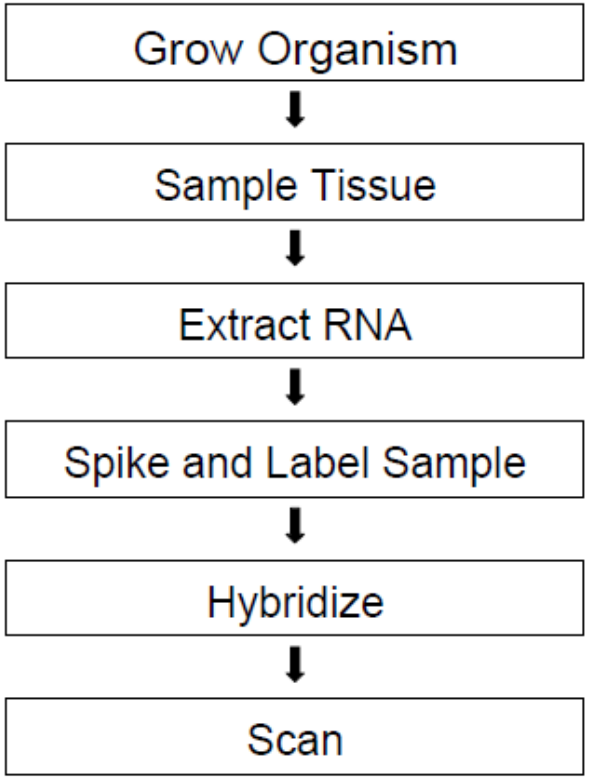
- "To learn something interesting about my cells" is usually not the best answer.

Common partial experimental objectives

- **Comparison:** identify differentially expressed genes
  - Cell types (brain vs. liver)
  - Developmental (fetal vs. adult)
  - Response to stimulus (rich vs poor media)
  - Gene activity (wild type vs. mutant)
  - Disease states (healthy vs. diseased)
- **Discovery:** identify clusters of genes or samples
- **Prediction:** use a gene expression profile to label a cell sample


Biological question
(Differentially expressed genes
Sample class prediction etc.)
Experimental design
Microarray experiment
Image analysis
Normalization
Estimation
Testing
•••••
Clustering
Discrimination
Biological verification
and interpretation

From Organism to Data



From Organism to Data