3. Gene Expression & the Transcriptome

Yow-Ling Shiue
Institute of Biomedical Science
National Sun Yat-sen University

Beyond the Human Genome

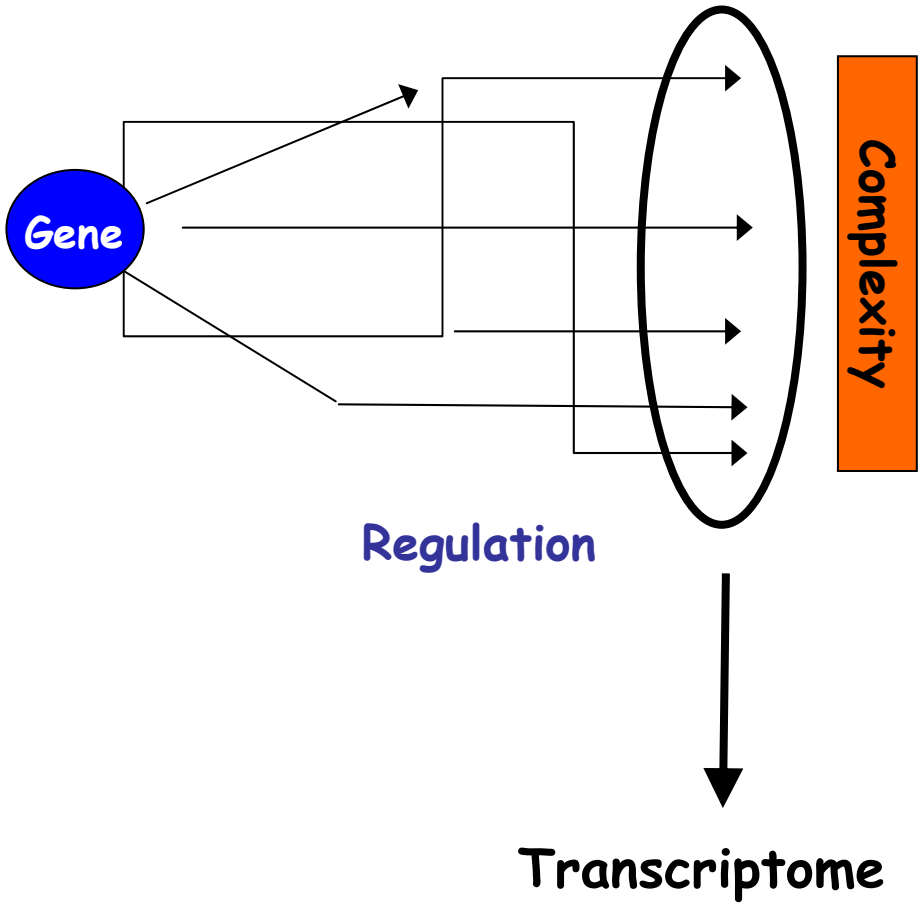
- 1995
  - Human genome sequencing begins in earnest " Mapping the Book of Life"
- 1999
  - Approx 140,000 genes
- 2000
  - First draft: 30,000 to 40,000? ?
- 2003
  - Essential completion: 24,195 genes! ! ! ? ? ?

•Commemorative stained glass window for FC Crick, designed by M. McClafferty

•Gonville & Gaius College, Cambrideg, UK

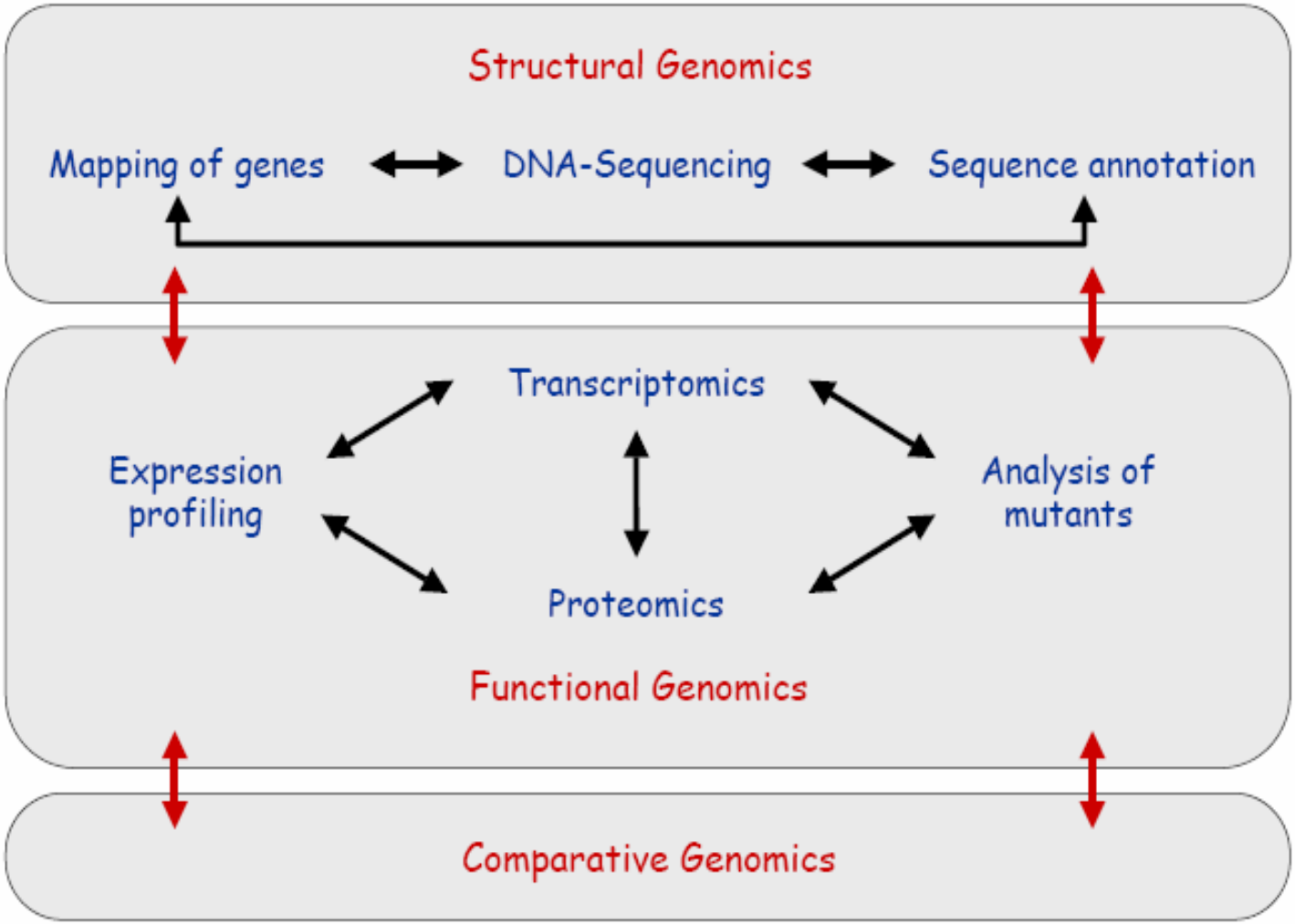
Beyond the Human Genome

Gene Number & Complexity



- Commemorative stained glass window for FC Crick, designed by M. McClafferty
- Gonville & Gaius College, Cambrideg, UK

Three levels of genome research



Transcriptome

- The **second major branch** of genome science
- Documenting gene expression on **a genome-wide scale**

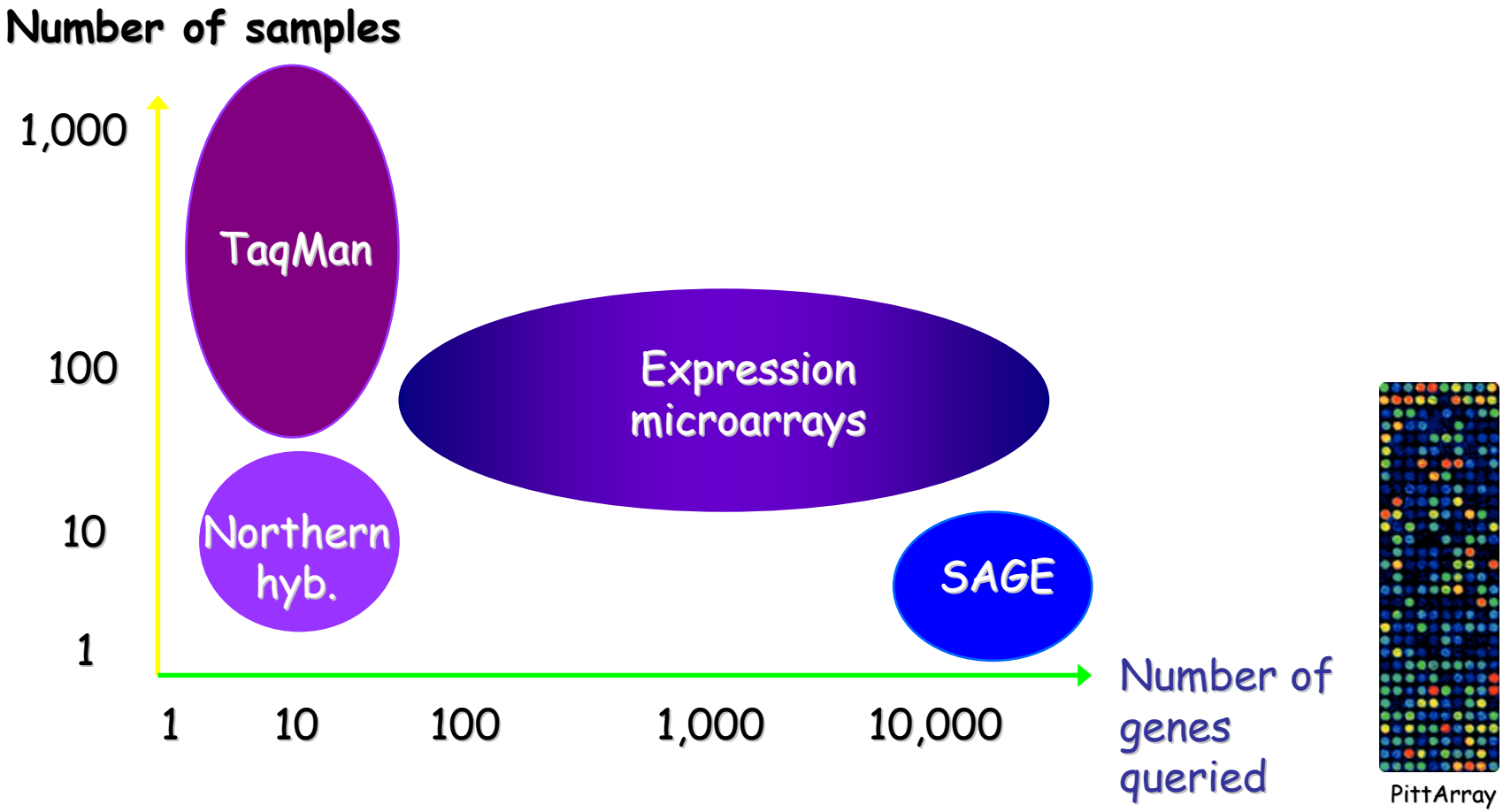
- The complete set of transcripts and their relative levels of expression in **a particular cell or tissue type** under defined conditions

- Transcription is **only one level** of gene regulation
- Transcript levels do not necessarily translate into protein expression or activity (vs. Proteome: the structure & expression of the proteins encoded in the genome)

Technologies Adopted

- Parallel Analysis of Gene Expression: Microarray (relative expression analysis)
  - cDNA microarray technology
  - Oligonucleotide microarray technology
  - Microarray data mining
- Absolute Expression Analysis
  - Series analysis of gene expression (SAGE)
  - Microbeads
  - Differential Display
- Single Gene Analysis
  - Northern blots
  - RNase protection
  - Quantitative PCR

Technology: Different Methods for Obtaining Gene Expression Data



Taken from Dr. Karoly Mirnics talk Bioinformatics@Pitt

Transcriptome & Transcriptomics

- The transcriptome consists in **the complete set** of transcripts expressed in **a particular cell or tissue type** under a defined condition

- " Transcriptome, the mRNAs expressed by a genome at any given time i-"Abbott 1999

- The massively parallel analysis of the relative expression level of a given cell or tissue's transcriptome in different conditions is termed " **Transcriptomics**"

Transcriptome: An Evolving Definition (1)

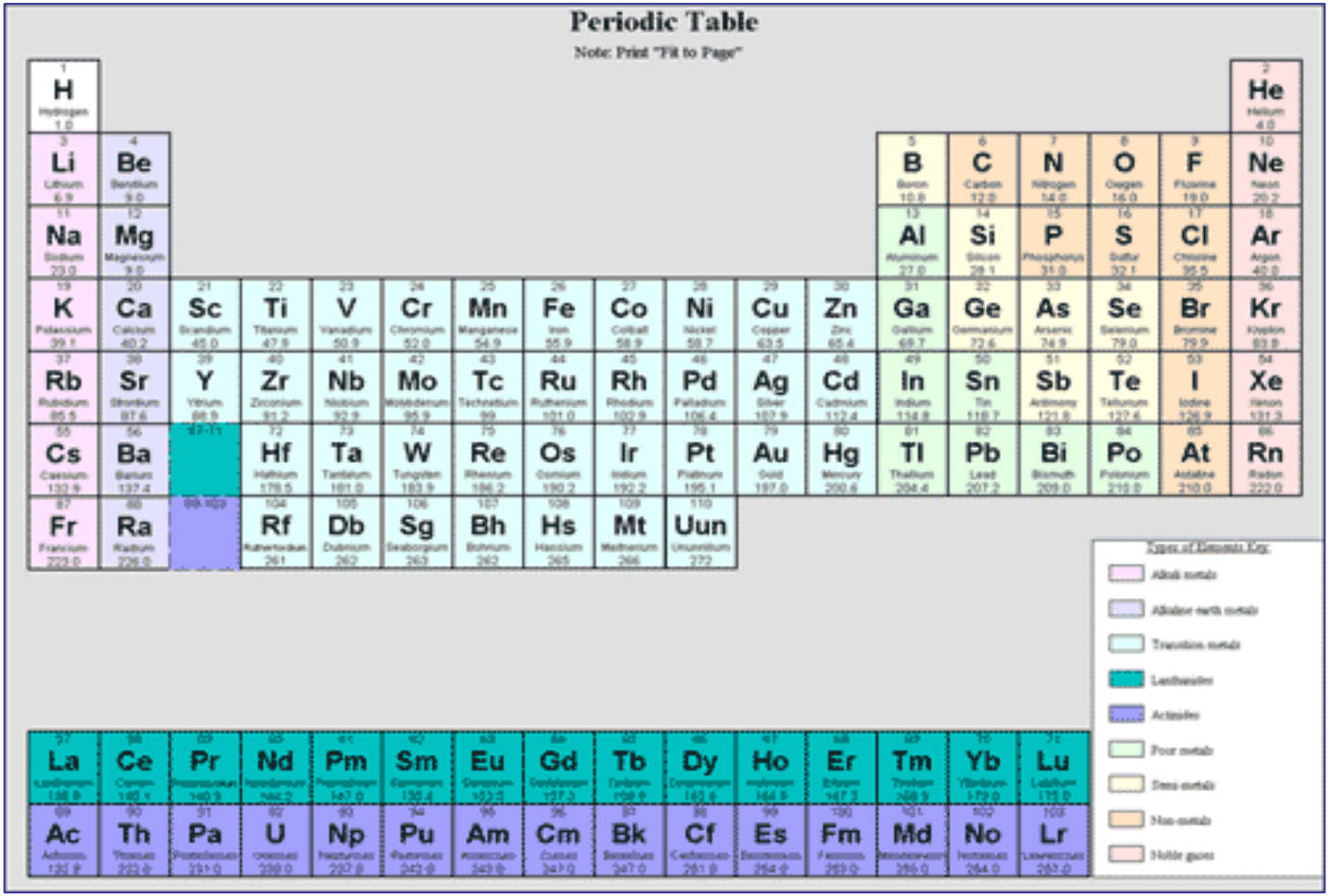
- The population of mRNAs expressed by a genome **at any given time**
  - Abbott 1999

- The complete collection of **transcribed elements** of the genome
  - Affymetrix 2004

Transcriptome: An Evolving Definition (2)

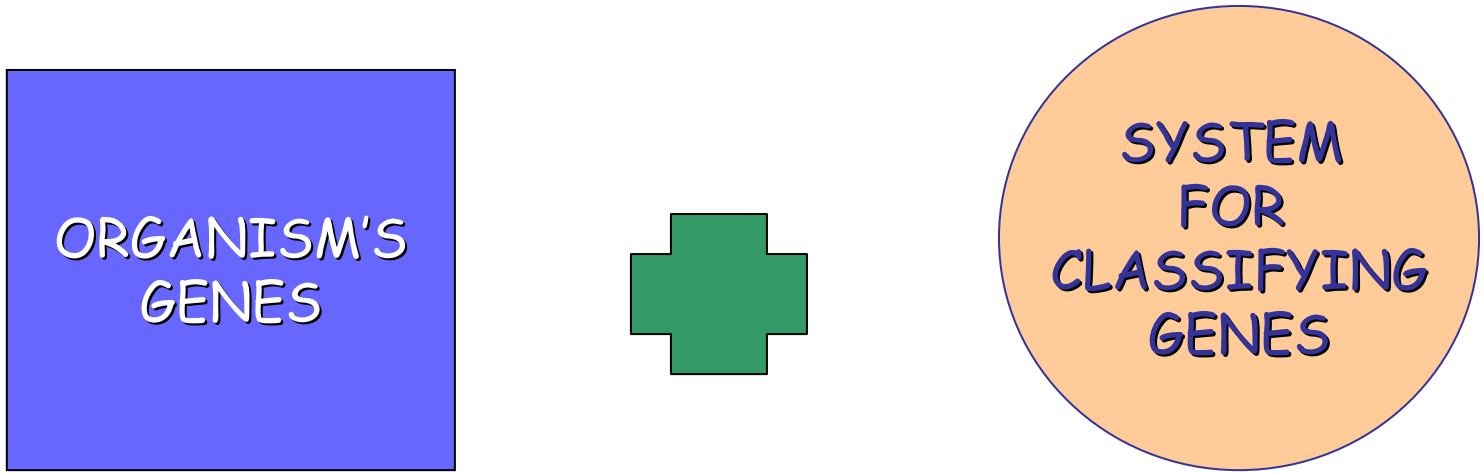
- **mRNAs: 35,913 transcripts**
- **Non-coding RNAs**
  - tRNAs: 497 genes
  - rRNAs: 243 genes
  - snmRNAs (small non-messenger RNAs)
- **MicroRNAs and siRNAs (small interfering RNAs)**
  - snoRNAs (small nucleolar RNAs)
  - snRNAs (small nuclear RNAs)
  - Pseudogenes: about 2,000

The Periodic Table: Functional Grouping of Chemical Elements



Biologist' s Periodic Table (1)

GENOMICS



Gopalakrishnan V. 2002

Biologist' s Periodic Table (2)

- Will not be two-dimensional

- Will reflect **similarities** at **diverse levels**
- Primary DNA sequence in **coding** and **regulatory** regions
- **Polymorphic variation** within a species or subgroup
- Time & place of expression of RNAs during development, physiological response & disease
- Subcellular localization & intermolecular interaction of protein products (protein-protein interaction)

Array of Hope?

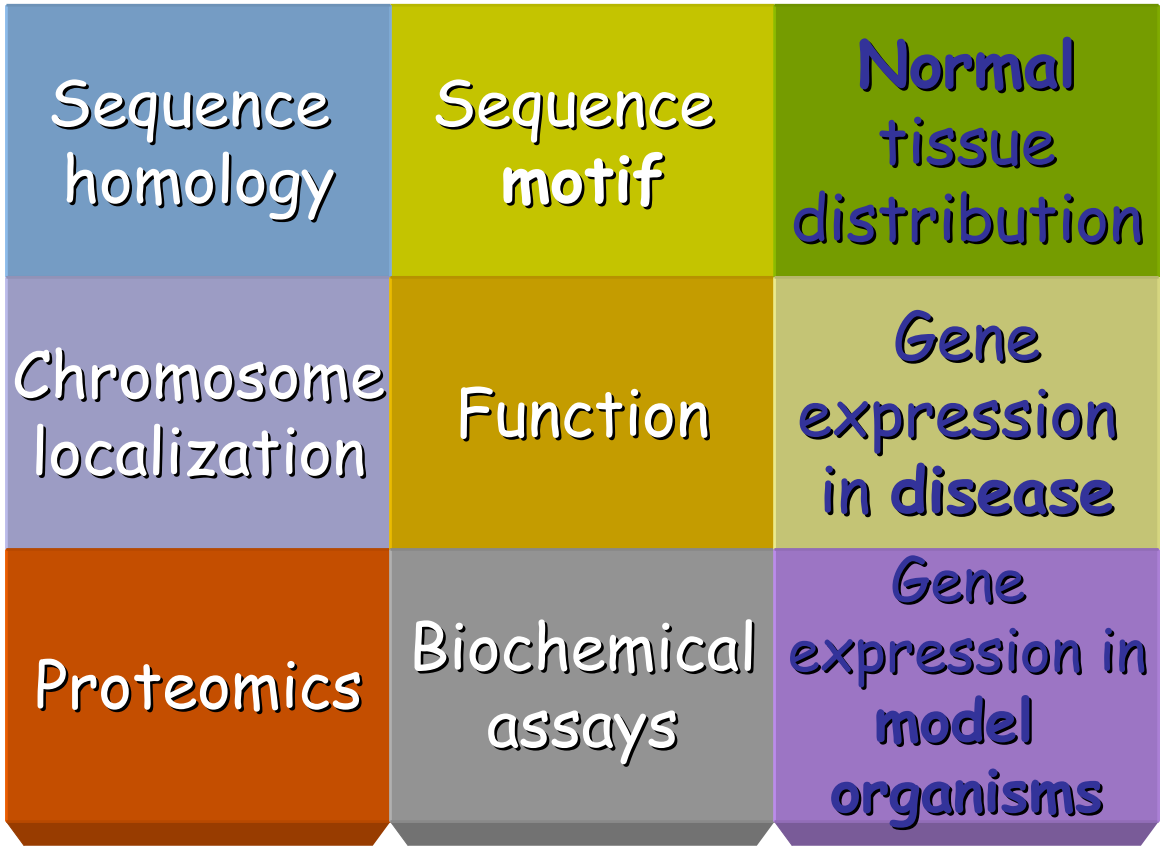
- Eric S. Lander, Whitehead Institute, MIT
  - Jan. 1999 Nature Genetics supplement, volume 21

- **Arrays** offer hope for **global views** of biological processes
  - Systematic way to study DNA & RNA variation
  - Standard tool for molecular biology research & clinical diagnostics

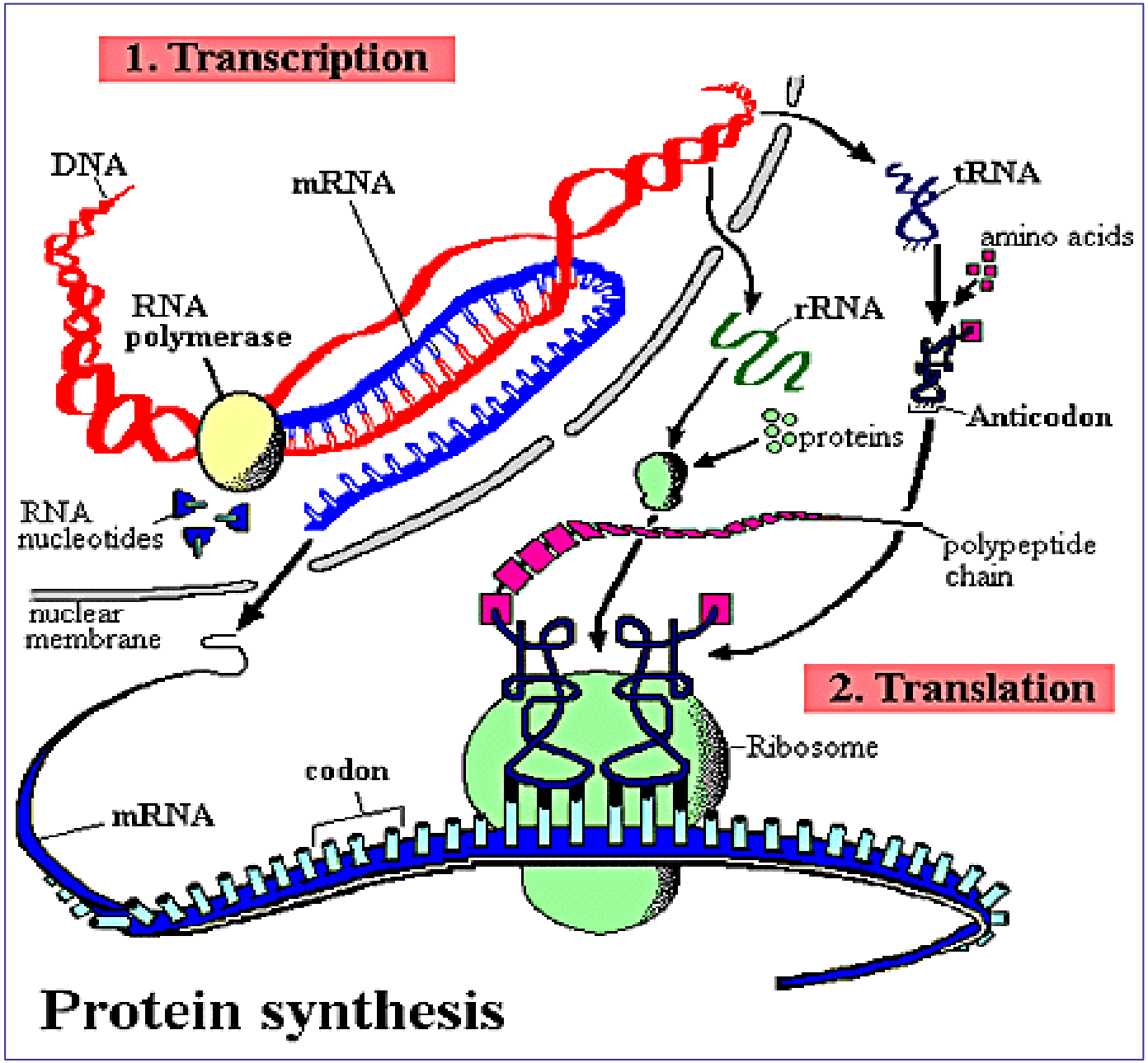
- Ed **Southern's** key insight: labeled nucleic acid molecules could be used to interrogate nucleic acid molecules attached to solid support
  - Probe vs. target

Determining Gene Function

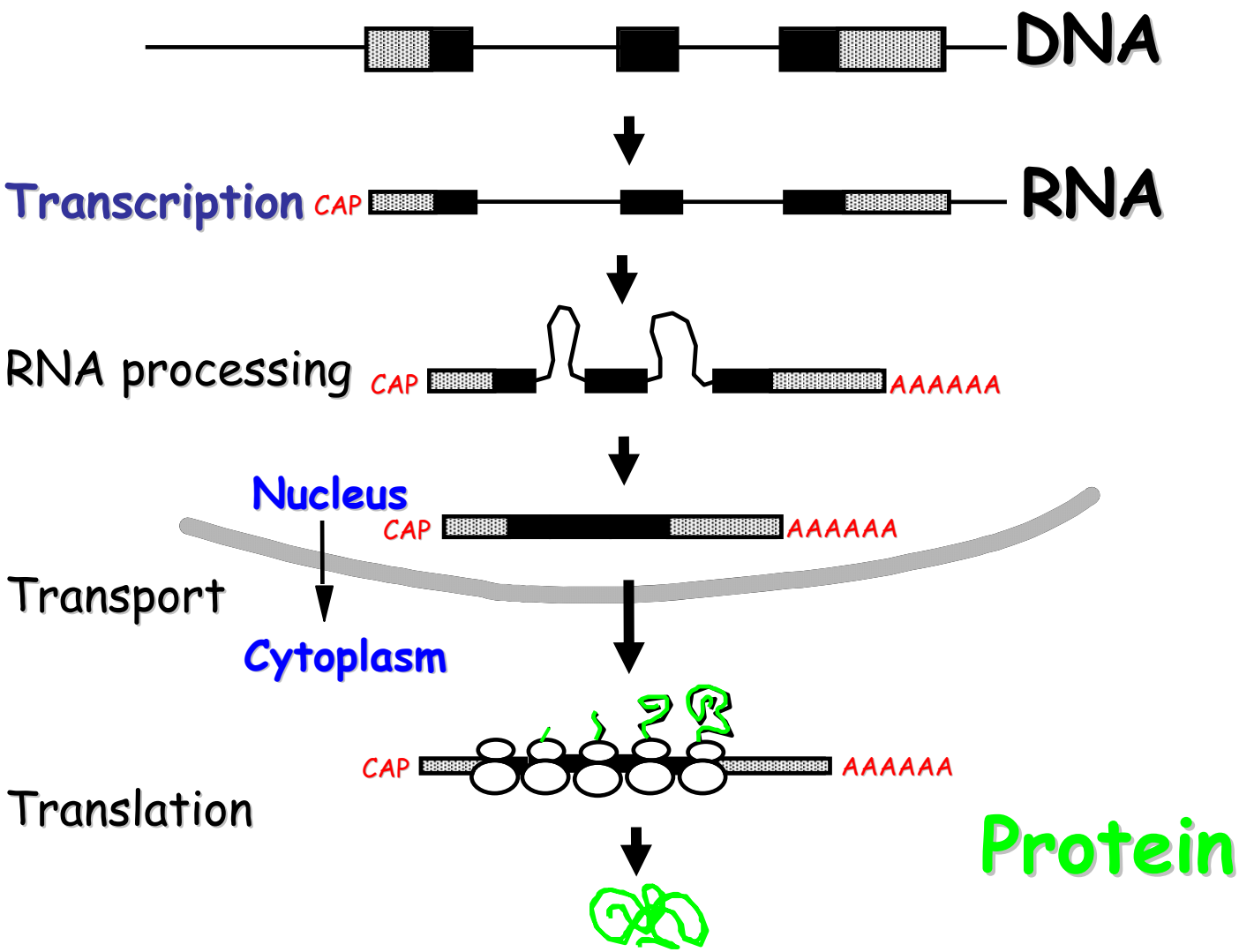
- C. Debouck and P.N. Goodfellow (1999)DNA Microarrays in drug discovery and development. Nature supplement Genetics, vol. 21.



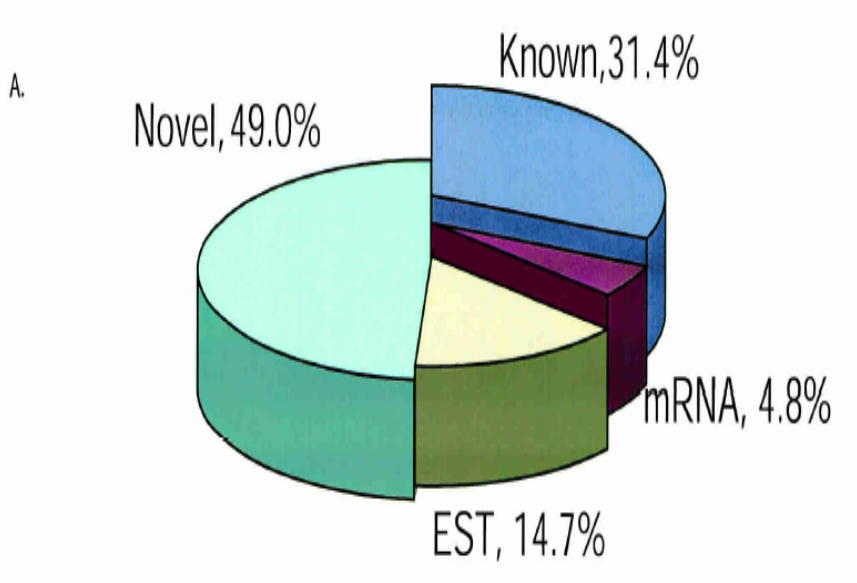
- Central dogma of molecular biology



Basic Mechanisms of Gene Expression



The Human Transcriptome



- High density oligonucleotide arrays across 11 different cell lines
- ~70% of transcripts non-coding
- ~78-88% have multiple transcripts

- Kapranov et al. 2002
- ~90% of transcribed nucleotides outside annotated exons

- The dimension of the unique transcriptome? ?
- Current 40,000 estimate

Kampa et al., Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Research. 2004

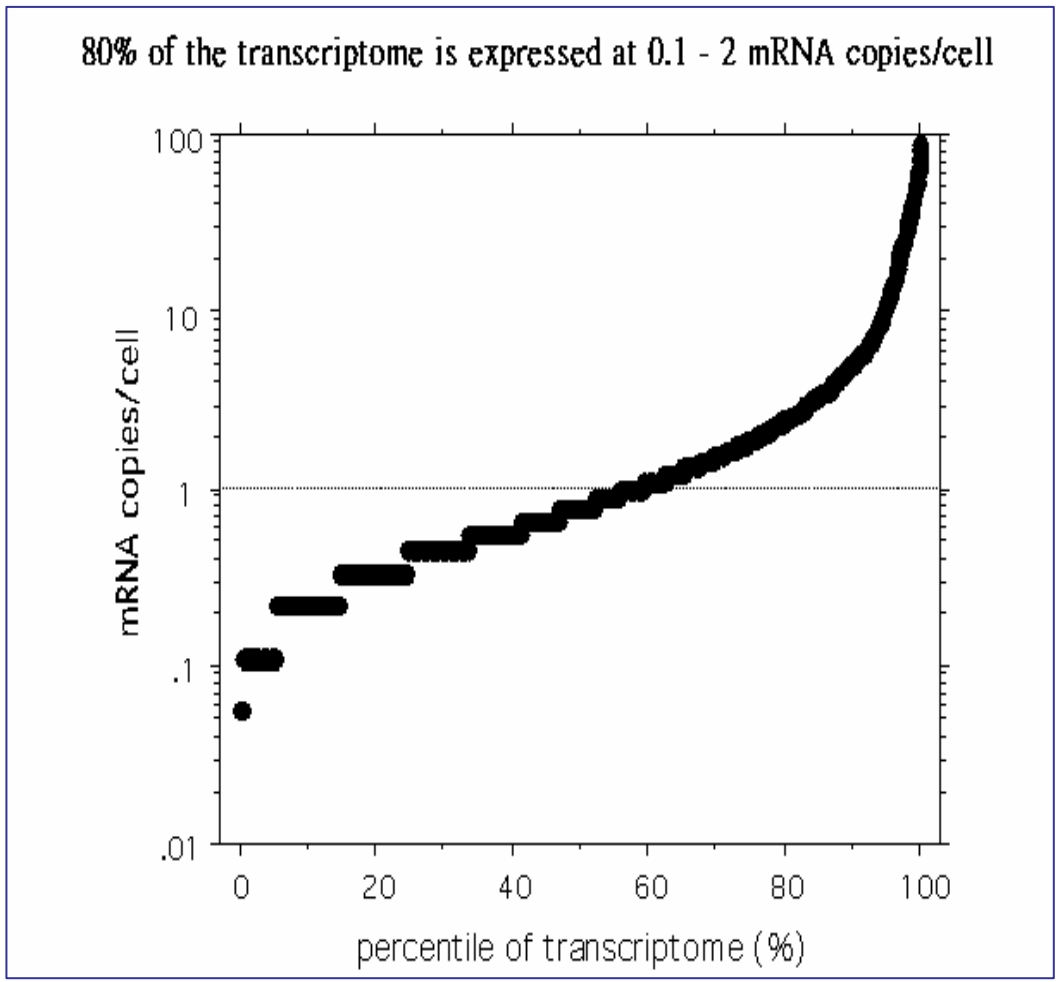
Applications of Transcriptome (1)

- Discovery of new **proteins**
  - That are present in specific tissues
  - That have specific cell locations
  - That respond to specific cell states
- Discovery of new **variants**
  - Of important genes
  - The work to increase/decrease the activity of the "native" protein

Applications of Transcriptome (2)

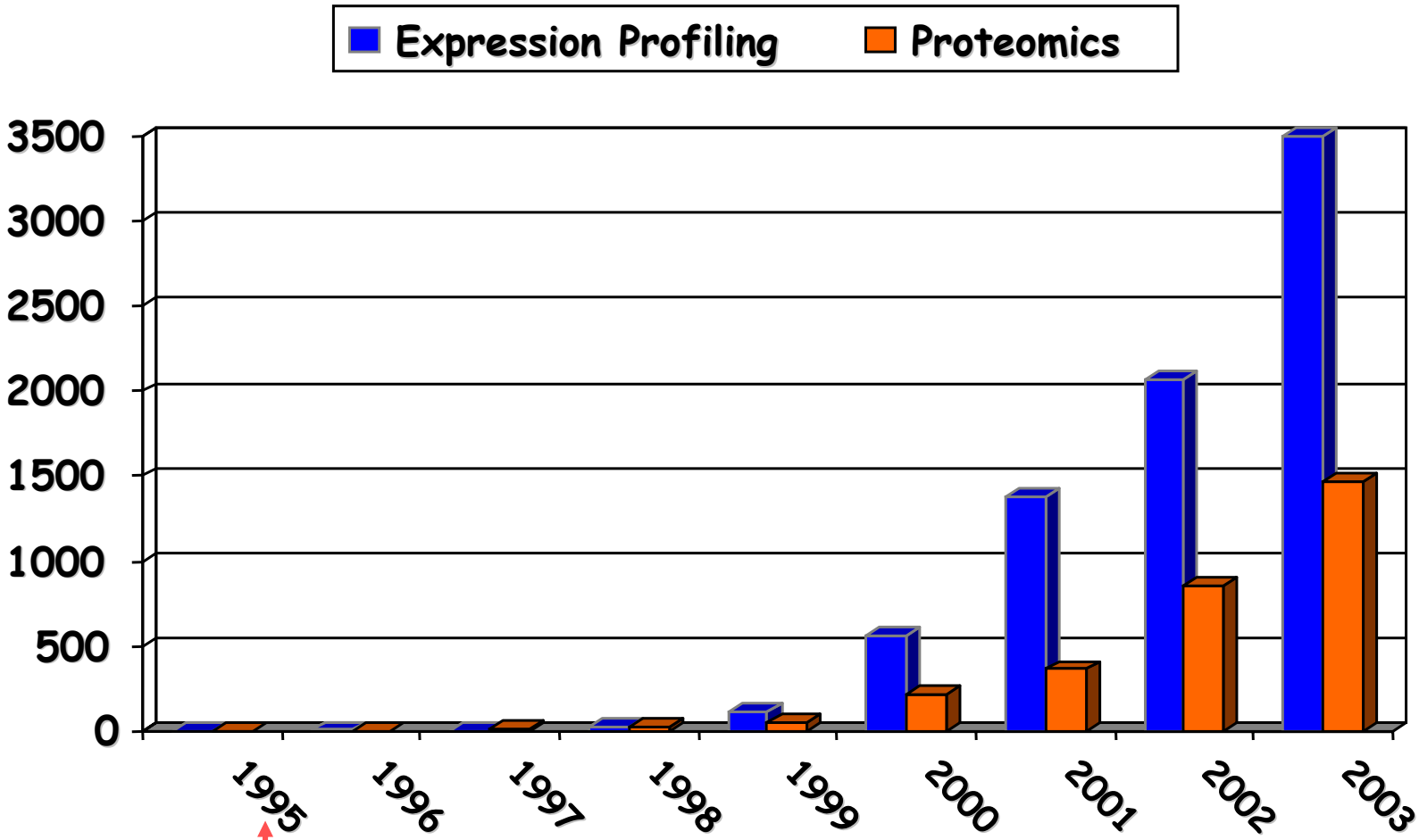
- The transcriptome reflects tissue source (cell type, organ) and also tissue activity and state such as the stage of development, growth and death, cell cycle, disease or healthy, response to therapy or stress

Yeast Transcriptome Statistics



- 5,460 transcripts
- 15,000 poly A-RNA's per cell
- Average level: **2.8 copies/cell**
- Median level: 0.79 copies/cell

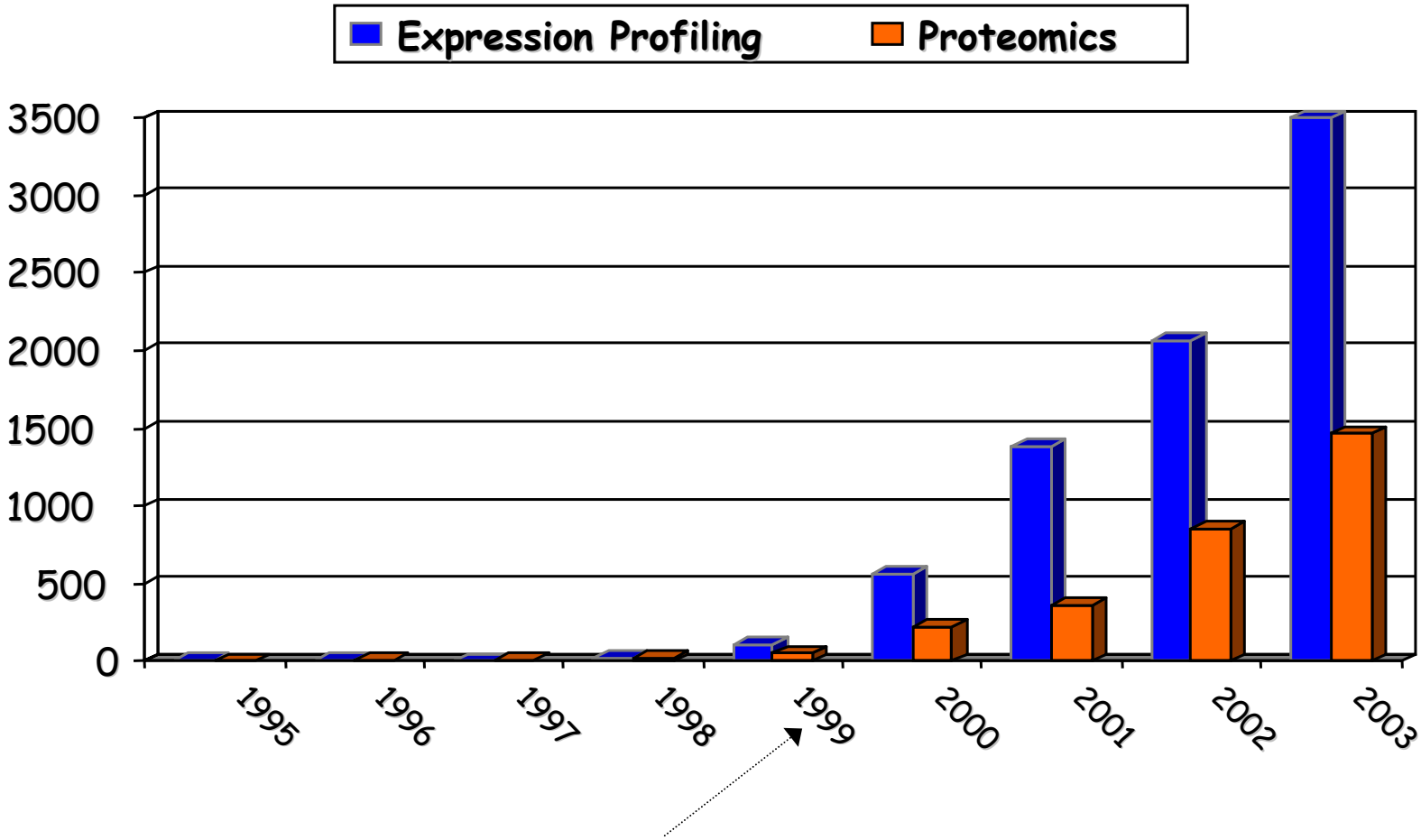
Publications: Expression Profiling vs. Proteomics



Quantitative monitoring of gene expression patterns with a complementary DNA microarray.
Schena M, Shalon D, Davis RW, Brown PO.
Stanford University Medical Center, CA.

Data from PubMed

Publications: Expression Profiling vs. Proteomics



"The challenge is no longer in the expression arrays themselves, but in **developing experimental designs** to exploit the full power of a global perspective."

Eric Lander

Data from PubMed

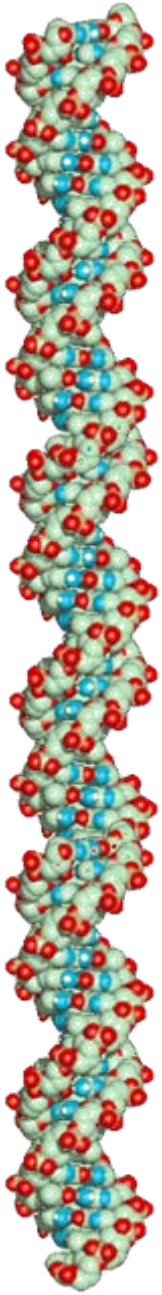
Human Transcriptome Map

- http://bioinfo.amc.uva.nl/HTMseq/controller
- UniGene database (NCBI)
  - dbEST, NCBI
- Cancer Genome Anatomy Project -CGAP
- SAGEmap, NCBI

dbEST & UniGene

- Database of expressed sequence tags; short, single pass read cDNA (mRNA) sequences
  - Also includes **cDNA sequences** from differential display experiments & RACE experiments
    - Subtractive libraries

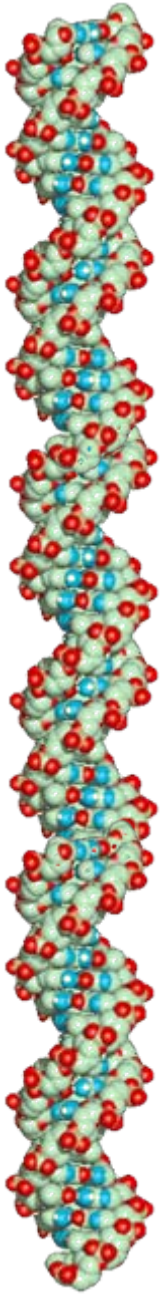
- **UniGene**
  - **ESTs** and **full-length mRNA sequences** organized into clusters that each represent a unique known or putative gene **within the organism** from which the sequences were obtained
  - UniGene clusters are annotated with mapping & expression information when possible (*e.g.*, for human), & include cross-references to other resources

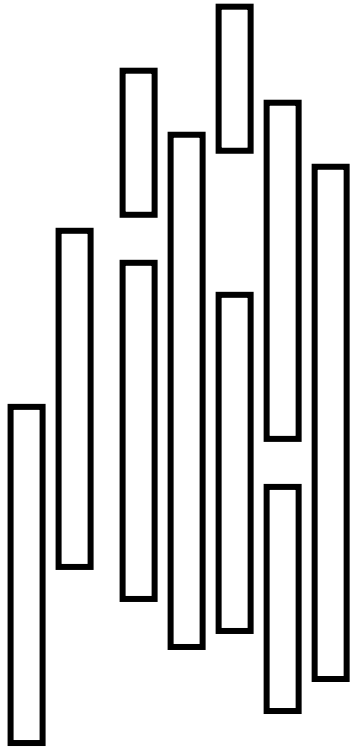


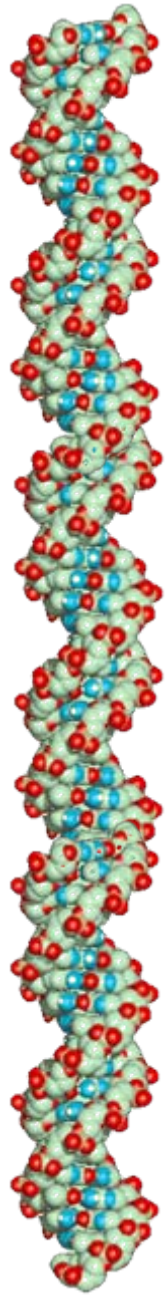
Cluster sizes in UniGene

This is a gene with 1 EST associated; the cluster size is 1

Cluster Sizes in UniGene







This is a gene with
10 ESTs associated;
the cluster size is 10

UniGene: Unique Genes via ESTs

- UniGene at NCBI
- UniGene clusters contain many ESTs
  - The best resource for *in silico* cloning
- UniGene data come from many cDNA libraries
  - When you look up a gene in UniGene
  - You get information on its abundance and its regional distribution (expressed tissues, or cell types)
- dbEST libraries for *Homo sapiens*
  - http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=9606

Homo sapiens (Hs)

- Unigene Build
  - Sequences included in Unigene
  - Build method
  - Final number of clusters (sets)
  - Histogram of cluster sizes for Unigene Hs build
  - DDD

UniGene - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=188401
移至 連結
Google se protection AND glossary 搜尋 PageRank 20 已攔截 檢查 選項 RNase protection glossary

complete cds

BC007320.2 Homo sapiens annexin A10, mRNA (cDNA clone MGC:1303 IMAGE:2988009), complete cds **PA**

EST Sequences (10 of 76) [Show all sequences]

BG611770.1	Clone IMAGE:4738455	prostate	5' read **PM**
BG500743.1	Clone IMAGE:4669621	prostate	5' read **PM**
BG489720.1	Clone IMAGE:4636865	lung	5' read **PM**
BG402780.1	Clone IMAGE:4525405	bladder	5' read **PM**
BG437158.1	Clone IMAGE:4622570	lung	5' read **PM**
BF208214.1	Clone IMAGE:4092524	bladder	5' read **PM**
BE789806.1	Clone IMAGE:3884315	lung	5' read **PM**
BE304485.1	Clone IMAGE:2988009	colon	5' read **PM**
BU540688.1	Clone IMAGE:6572183	lung	5' read **PM**
BU183283.1	Clone IMAGE:6106778	pancreas	5' read **PM**

human
ANXA10

Key to Symbols

P Has similarity to known **P**roteins (after translation)

網際網路

UniGene - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=unigene 移至 連結
Google se protection AND glossary 搜尋 PageRank 20 已攔截 檢查 選項 RNase protection glossary

NCBI

UniGene
ORGANIZED VIEW OF THE TRANSCRIPTOME

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search UniGene for ANXA10 Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort by Send to

All: 2 | Fungi: 0 | Insects: 0 | Mammals: 2 | Plants: 0

Items 1 - 2 of 2 One page.

NCBI

UniGene
Homepage
Query Tips
FAQs
Finding cDNAs
Library Browser
DDD
Download UniGene

Related Databases
Gene
HomoloGene
dbEST
Trace Archive

NIH cDNA Projects
MGC
ZGC
XGC
CGAP

1: Mm.42179 Links
Anxa10: Annexin A10
Mus musculus, 39 sequence(s)

2: Hs.188401 NIH cDNA clone, Links
ANXA10: Annexin A10
Homo sapiens, 81 sequence(s)

完成 網際網路

Cancer Genome Anatomy Project -CGAP (1)

- Interdisciplinary program to identify the human genes expressed **in different cancerous states**, based on cDNA (EST) libraries, to determine
  - The molecular profiles of normal, precancerous & malignant cells

- Collaboration among **the National Cancer Institute** (NCI), the NCBI, and numerous research laboratories
  - 1937 established
  - The Federal Government's principal agency for cancer research and training

Cancer Genome Anatomy Project -CGAP (2)

- **Goal**
  - To generate the **information** and **technological** tools needed to decipher the molecular anatomy of the cancer cell

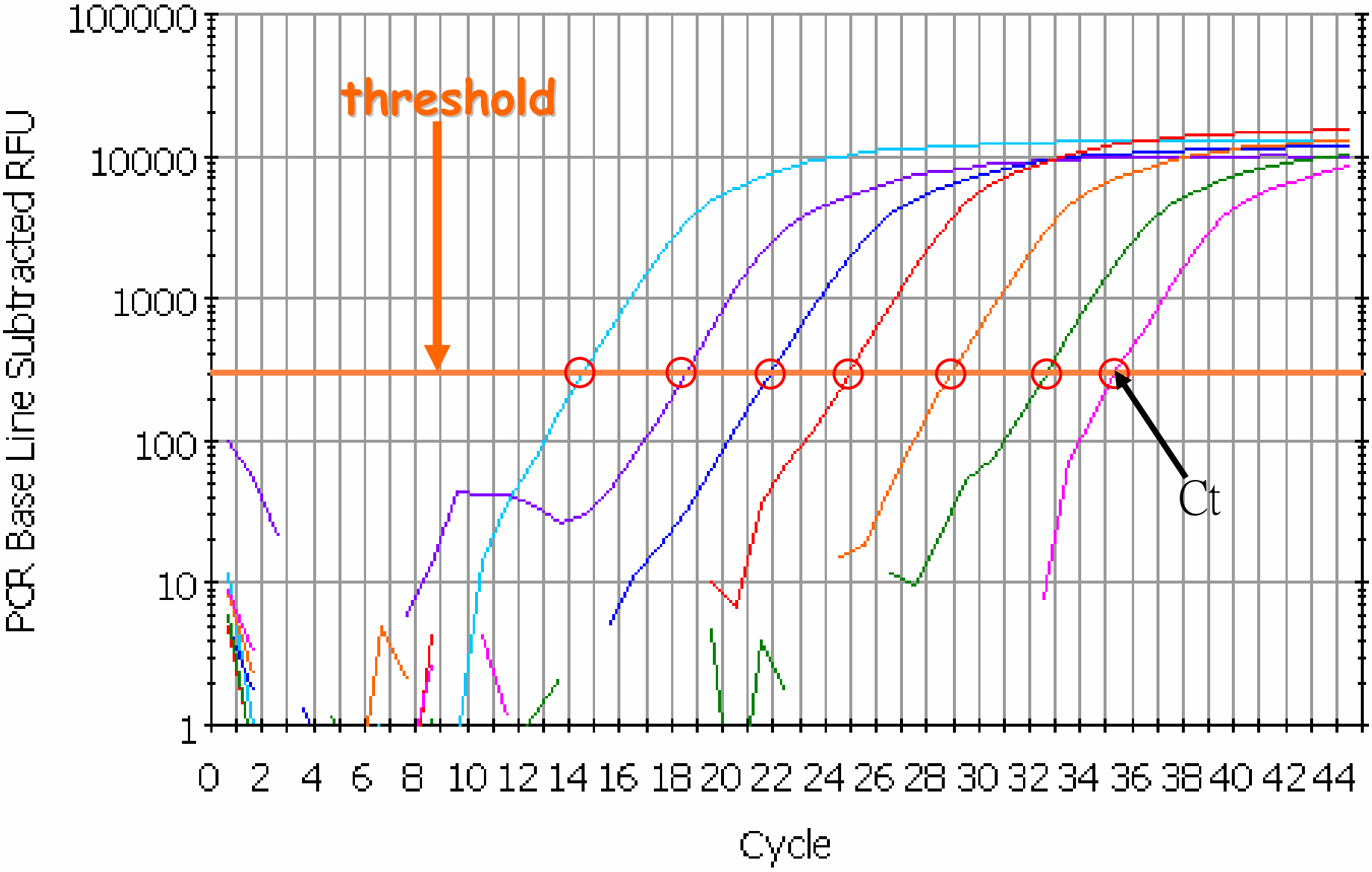
- http://cgap.nci.nih.gov/

Single Gene Analysis

- After identify candidate genes
  - A small subset of genes
- To use **a different experimental procedure** to confirm that the gene really is differential expressed
- **Three major methods**
  - Northern blots (μg)
  - RNAse protection (μg)
  - **Quantitative RT-PCR (ng)**

Quantitative PCR (1)

- **Quantitative reverse-transcription PCR**
  - Q-PCR or Q-RT-PCR
- **General PCR**
  - **Not quantitative**: the end product is observed after **the bulk** of the product has been synthesized, at a point where **the rate** of synthesis of new molecules has reached **a plateau** ⇨ **small differences** in the amount of target at the start of the reaction are masked
- **Quantitative measures** of nucleic acid polymerization must be made **during the linear phase of the reaction**



Series of 10-fold dilutions

DNA Microarray

- High-throughput **Southern Hybridizations (DNA/DNA)**

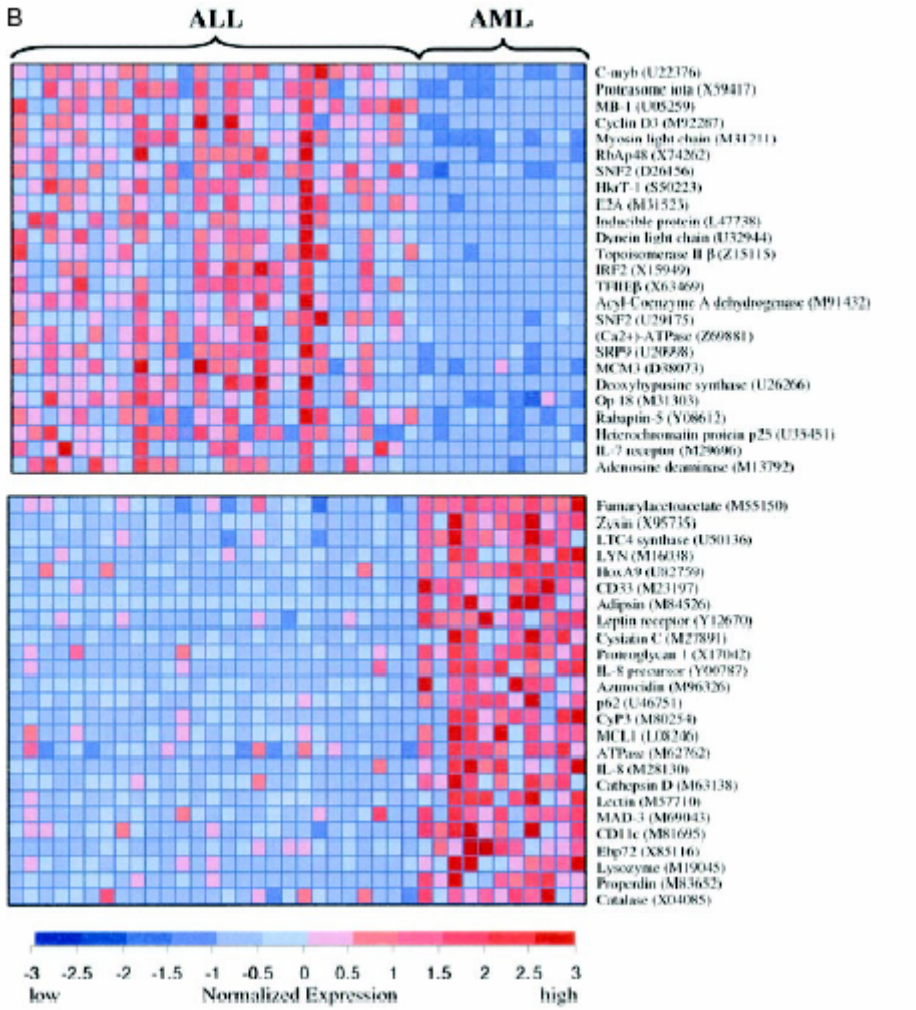
- **Three types**
  - cDNA microarray (generic, multiple manufactures)
  - cDNA membranes (radioative detection)
  - Oligonucleotide arrays (GeneChips®)

- More information
  - http://www.genetics.pitt.edu/services/labpage.html?whichlab=pittarray

Classic Human Transcriptome Profiling Studies: Trancriptome Reflects Biology

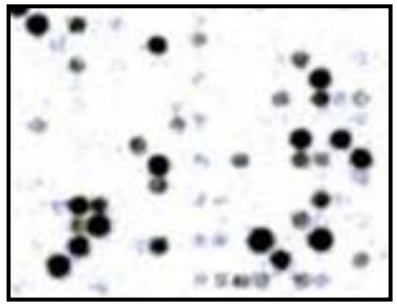
- Golub et al.
- Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science 1999**.
- ALL – acute lymphoblastic leukemia
- AML – acute myeloid leukemia

- Scherf et al.
- A gene expression database for the molecular pharmacology of cancer
- **Nature Genetics 2000**
- 60 human cancer cell lines
  - http://genome-www.stanford.edu/nci60/

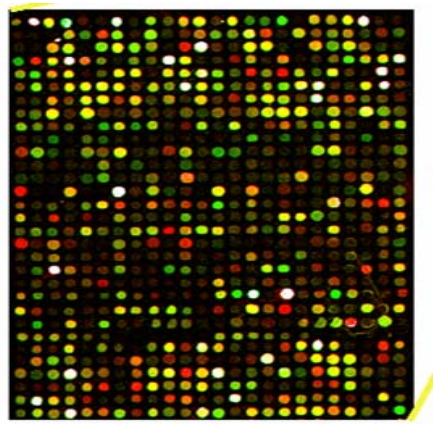


Platforms and Formats (1)

- **Isotope**
  - Nylon – cDNA (300-900 nt)

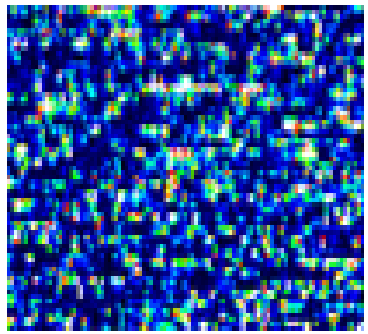


- **Two-colour**
  - Glass
  - cDNA or Oligo (80 nt)
  - 500 – 11,000 elements

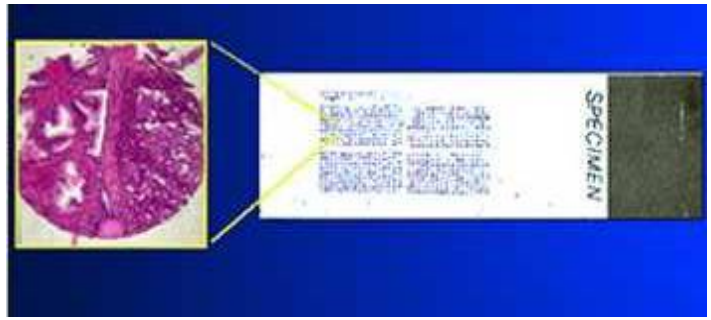


Platforms and Formats (2)

- Affymetrix®
  - Silicone – oligo (25 nt)
  - 22,000 elements



- Tissue Arrays
  - Glass
  - Tissue Discs (20-150)



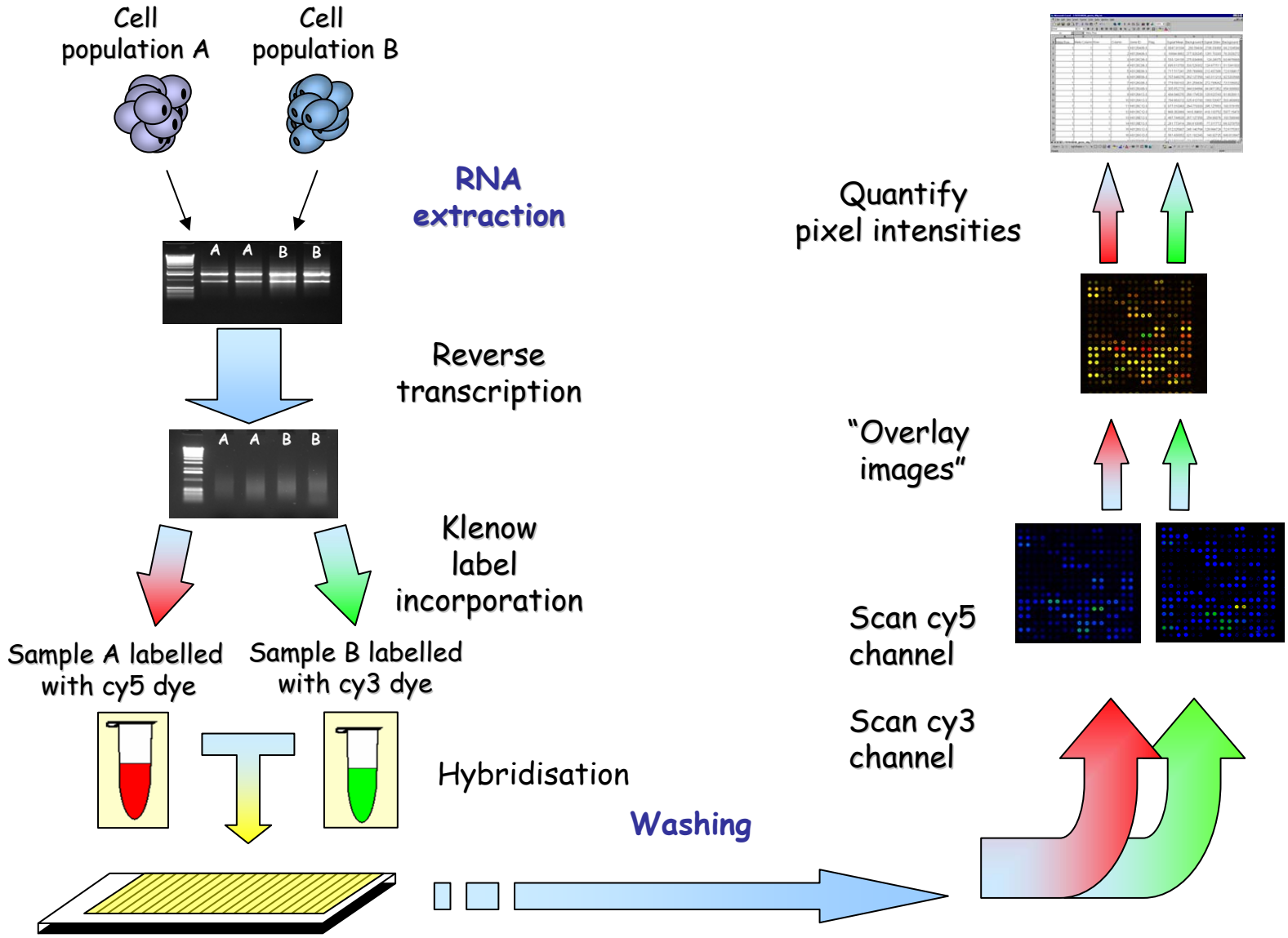
Bioinformatics Challenges (1)

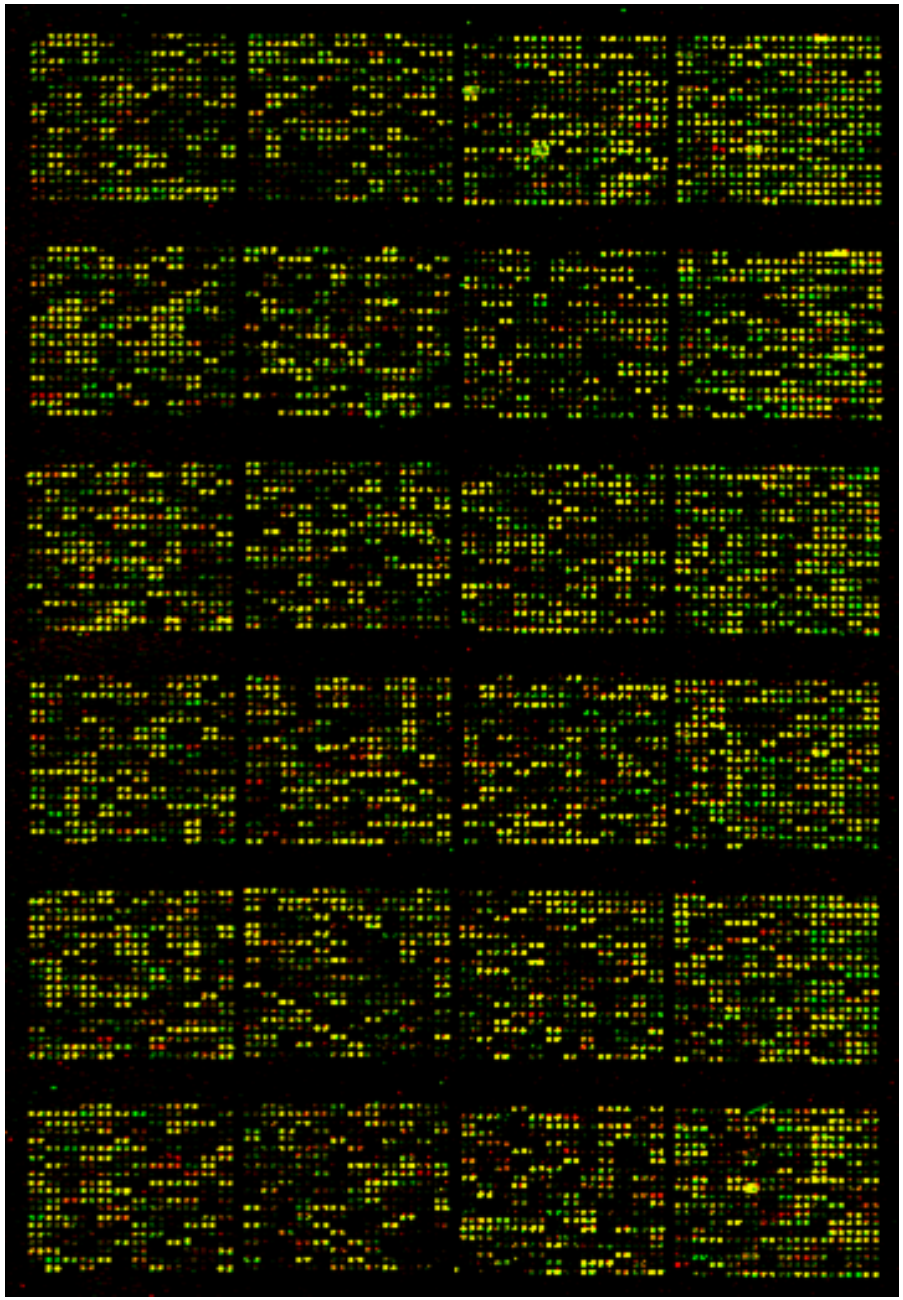
- **High dimensional data**
  - **Tens of thousands** of genes in **a single tissue sample** that are studied **simultaneously**
  - Many of the genes are **irrelevant** **(noisy data)**
- Problem is incompletely characterized
  - Not sure what "**difference in expression**" really means

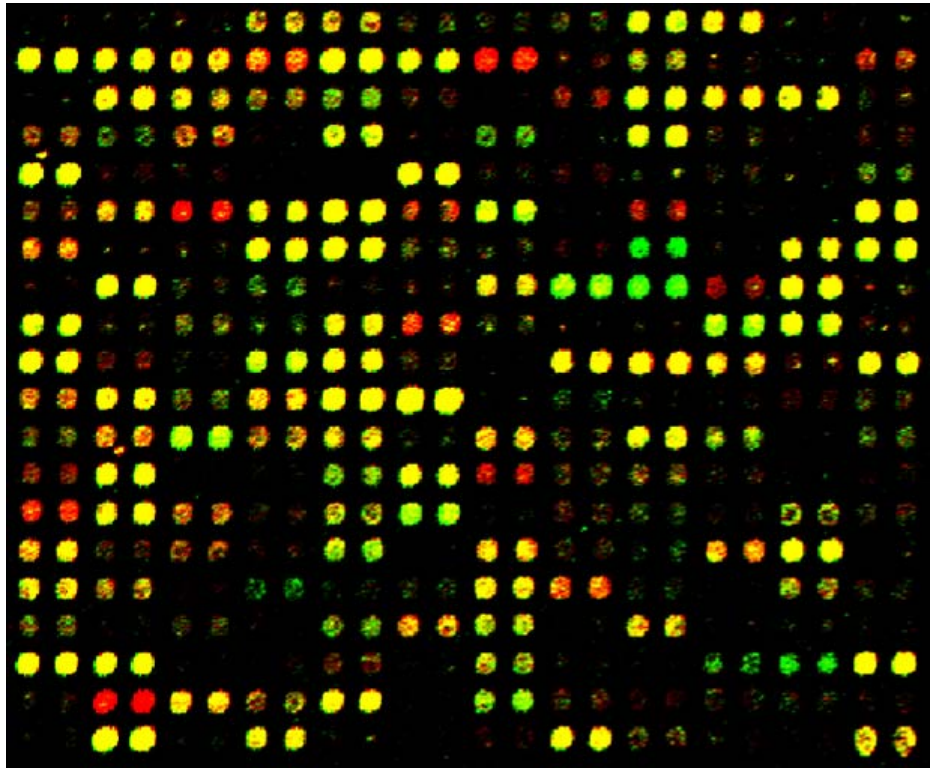
Bioinformatics Challenges (2)

- **No standards** for data collected from experiments
  - **Different techniques** are used for studies
  - Need to understand **what type of data** is **available** and **where and under what assumptions** it is produced

Experimental Overview







Red: increase of Cy5 sample transcripts

Green: increase of Cy3 sample transcripts

Yellow: equal abundance

Limits: 1:30,000 transcripts

~20 transcripts/cell

Choice & Amplification of ESTs (1)

- cDNA microarray have a current limit of **30,000 spots** (normally ~5,000 spots)
- **Ideally**, **each EST** should represent **a unique gene** or **alternative splice variant** (*i.e.*, a **unique set**)
- As an alternative, for **ill-defined genomes** or **particular tissues**, **entire cDNA libraries** can be used to build up microarrays
  - Highly expressed genes will be however over-represented (redundancy)

Choice & Amplification of ESTs (2)

- **Multigene families** comprising several members will be represented by **many different clones**
  - Unless ESTs corresponding to **non-conserved regions** of the genes are used for spotting, the risk of **cross-hybridization** between transcripts from different genes exists

- Completion of whole genome projects leads to the assembly of **new Unigene sets**, that include genomic sequences representing **predicted genes** for which the corresponding EST has not yet been identified

Printing (1)

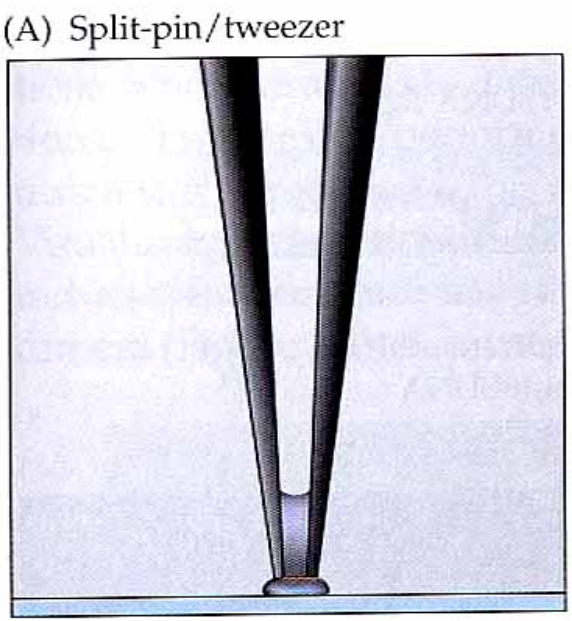
- **Supports**
  - Membrane (nylon, nitrocellulose) ⇨ radioactive labeling of probes
- **Robots**
  - DNA samples from 384-well plates are deposited onto a field of more than 100 slides
  - Spot **diameter** is normally about **50 micron**, distances between sport centers are **150-250 microns** with a tolerance less than 20 microns
  - Spot volumes range from **pico- to microliters**
  - A single amplification round can easily give raise to printing of **more than 1,000 slides**

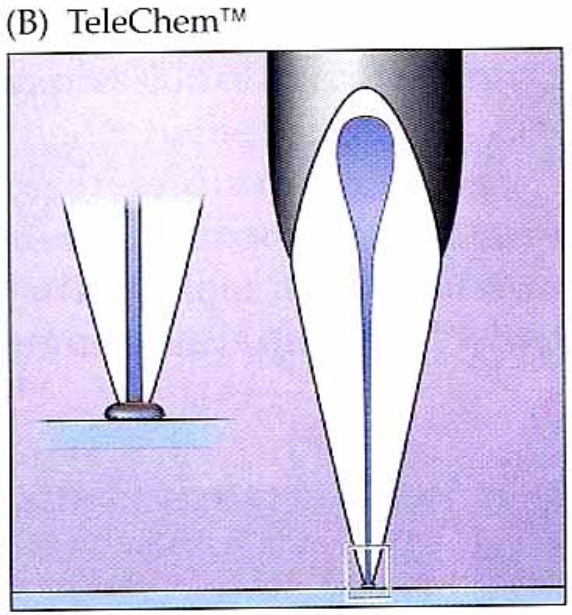
Printing (1)

- Printing
  - Capillary transfer
  - Pin-and-loop
  - Ink jet

Nanoliter

Capillary action

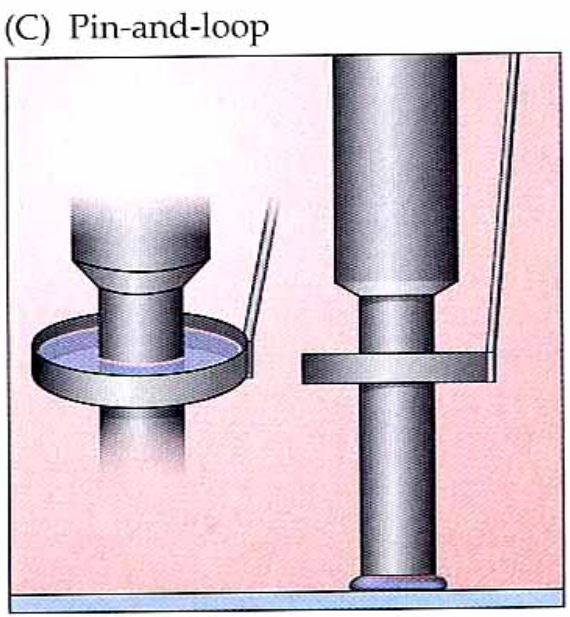


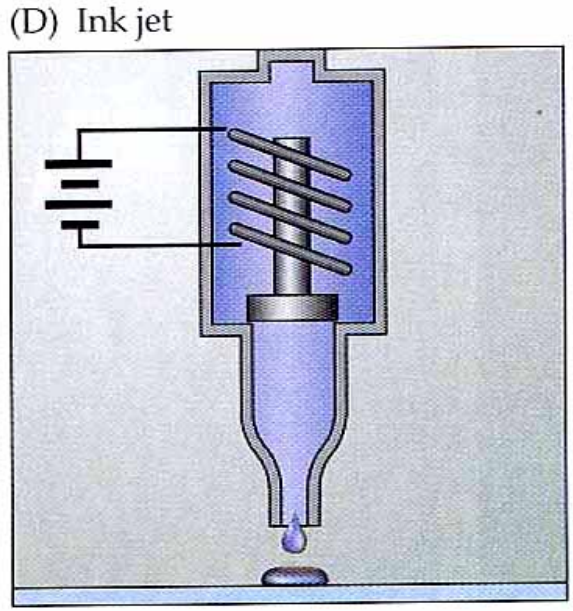


Small droplet

Contact

An uniform density



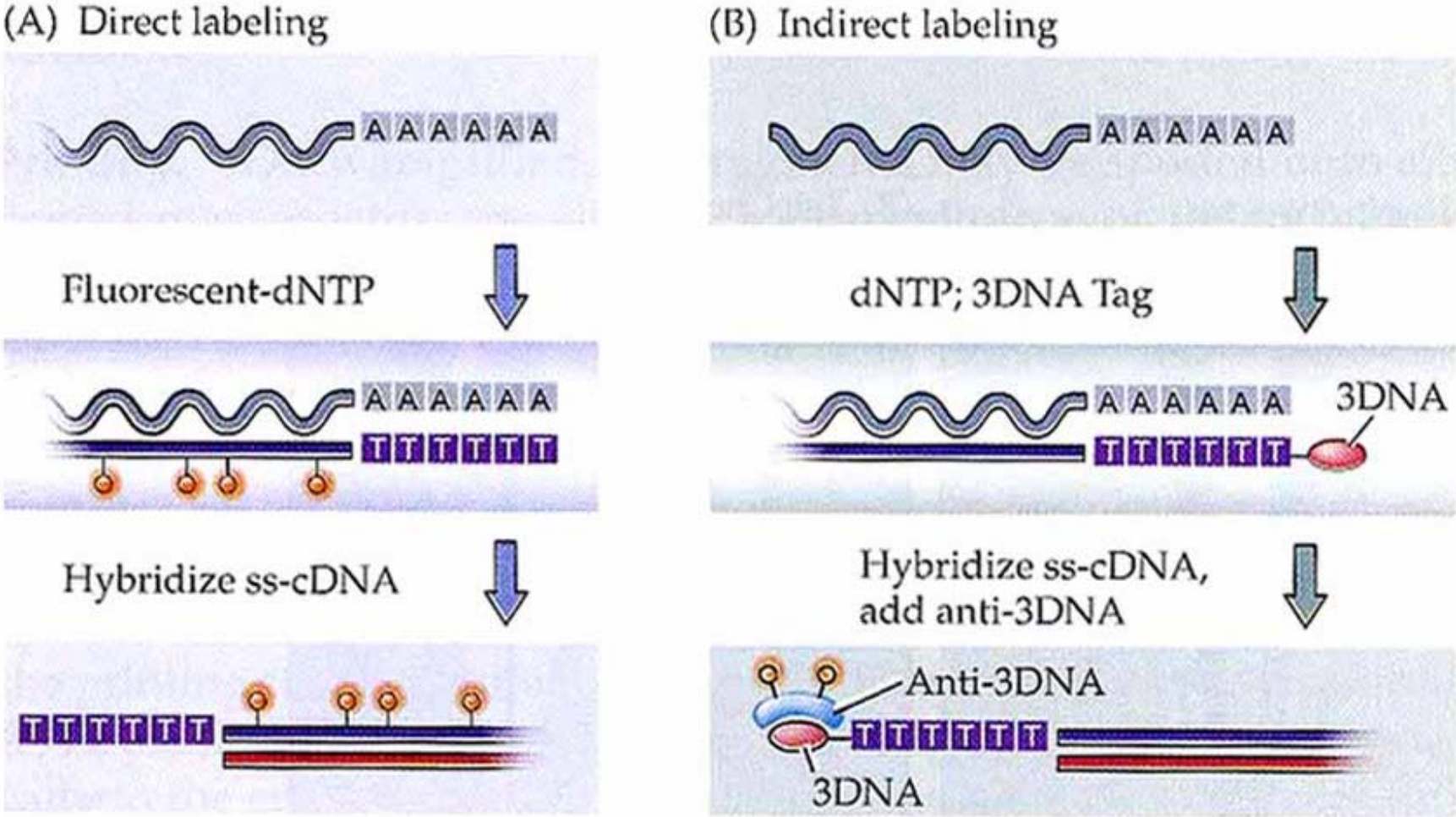


Picoliter

Under pressure

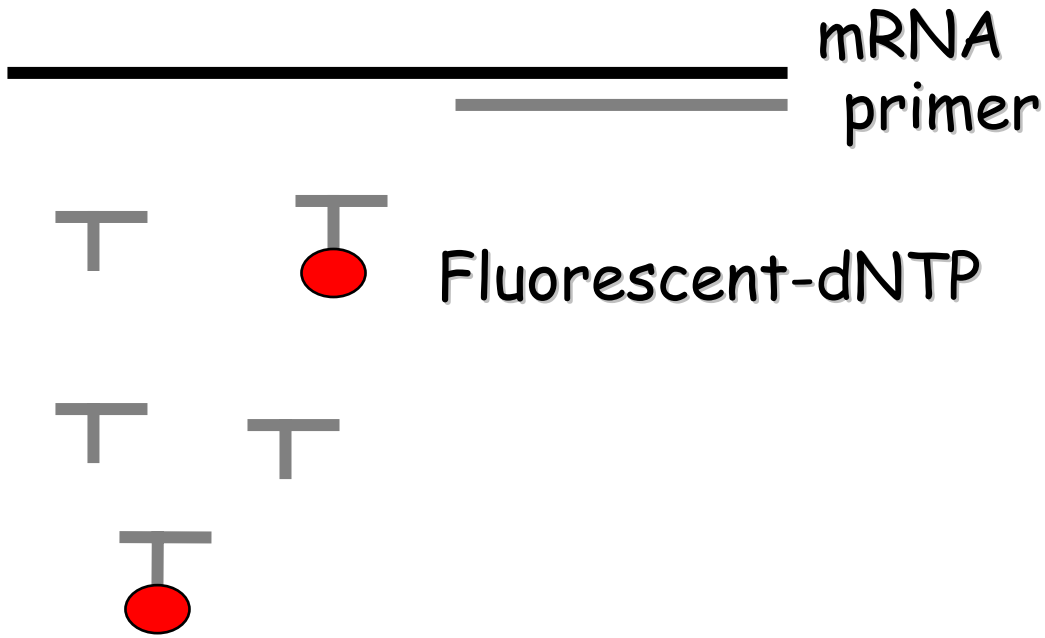
Labeling & Hybridization of cDNAs

- **Labeled cDNA** is prepared by reverse transcription of messenger RNA
  - Direct labeling (fluorescent dNTPs)
  - Indirect labeling (dNTP; 3DNA tag)
- Hybridization is performed in **humidified chambers**
  - High stringency conditions are used to minimize the risk of cross-hybridization
    - High Tm
    - Low salt
- Detection is performed by **laser-induced** fluorescence imaging
  - Confocal detector
  - CCD camera

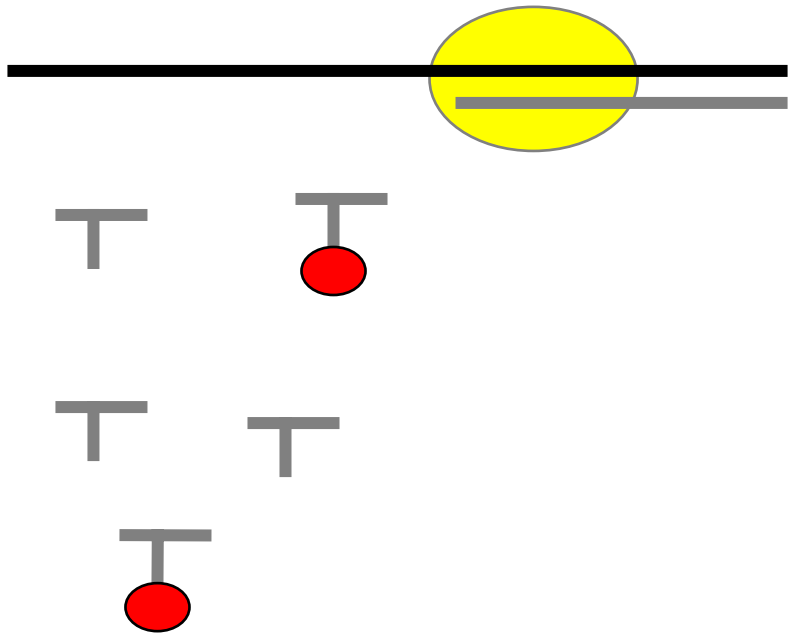


•Methods for labeling cDNA (A) Direct fluorescent dye incorporation during **reverse transcription** (B) Genisphere 3DNA submicro labeling, in which the 3D reagent is attached to an oligonucleotide that is complementary to **the 5'- end of the primer used in cDNA synthesis**, then hybridizes to it on the microarray (from Gibson & Muse 2001)

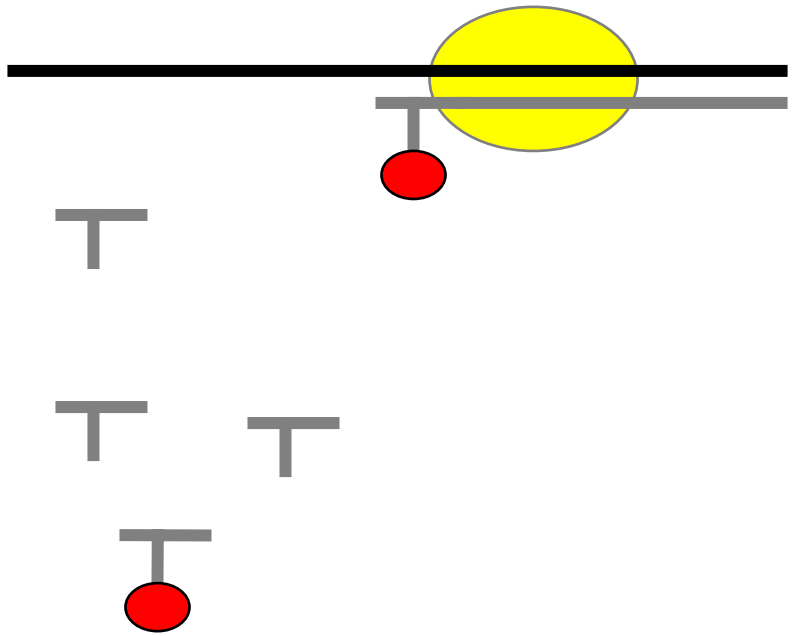
Direct Labeling (1)



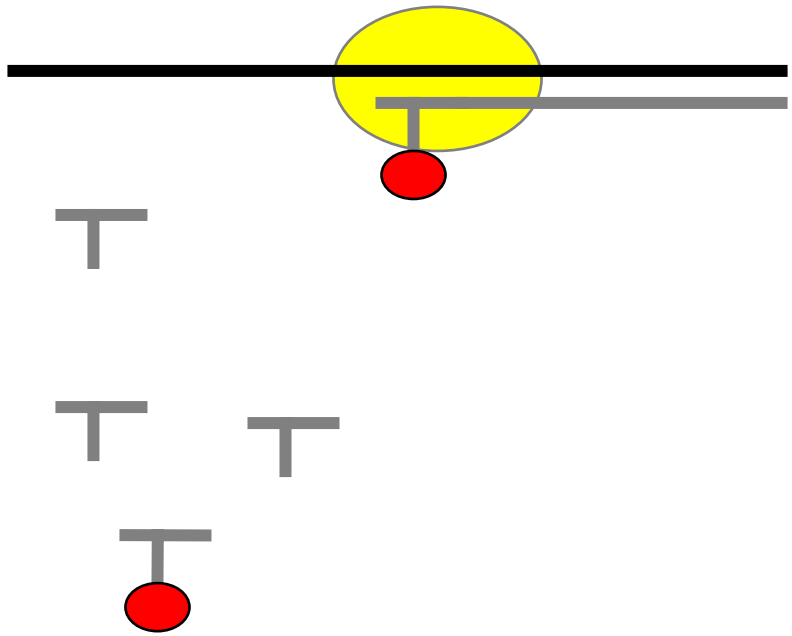
Direct Labeling (2)



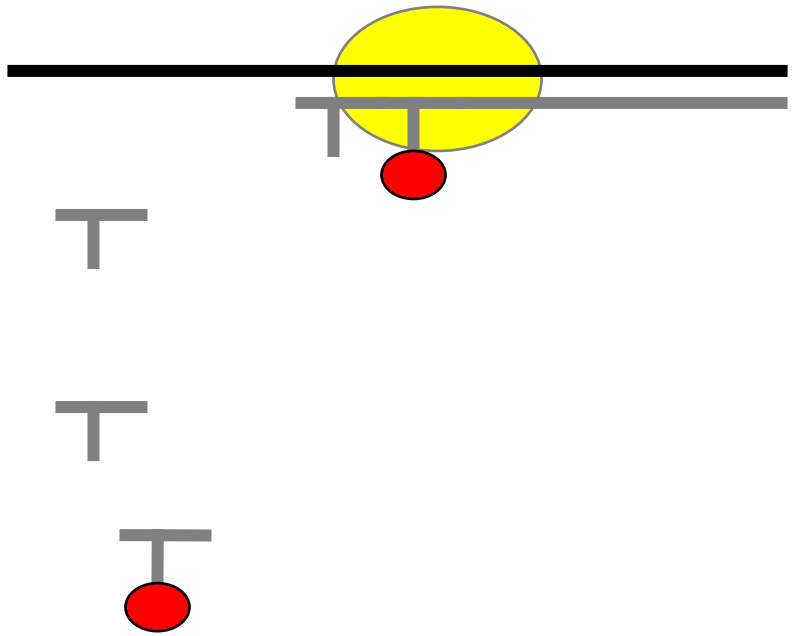
Direct Labeling (3)



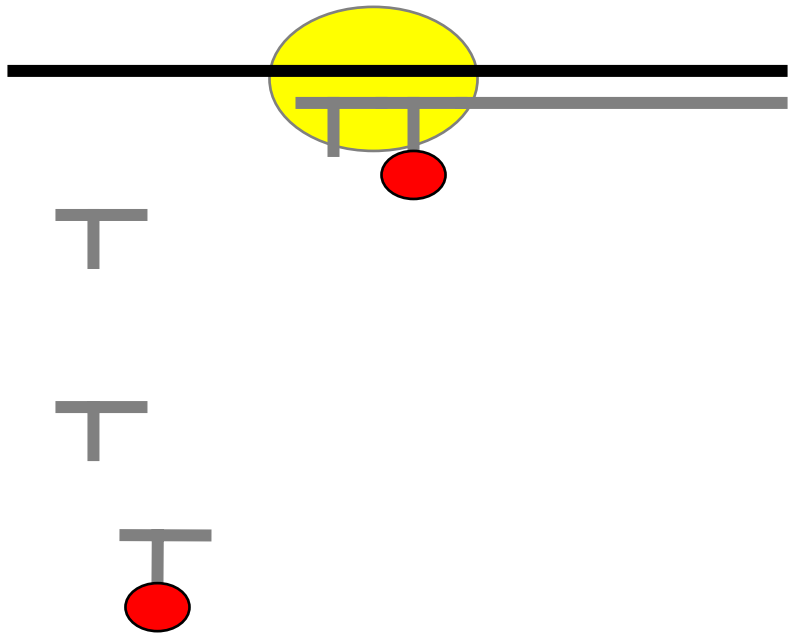
Direct Labeling (4)



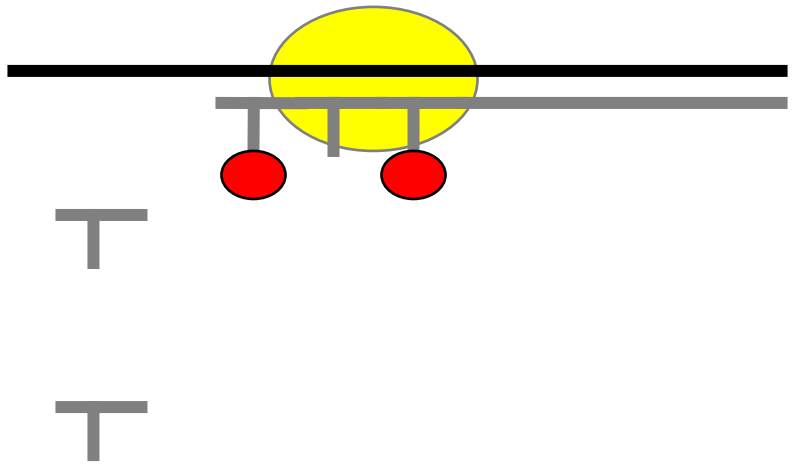
Direct Labeling (5)



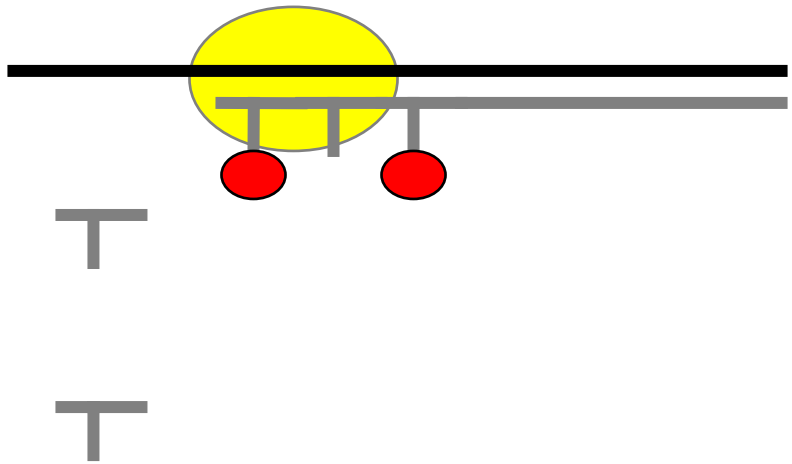
Direct Labeling (6)



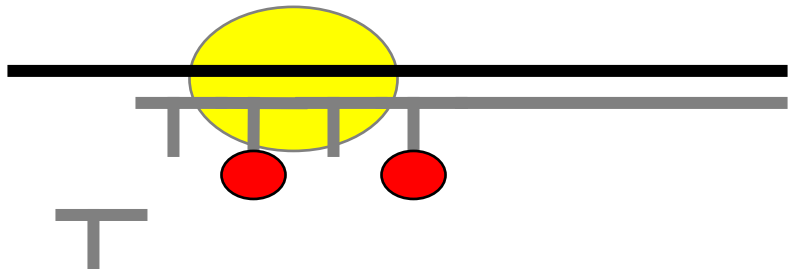
Direct Labeling (7)



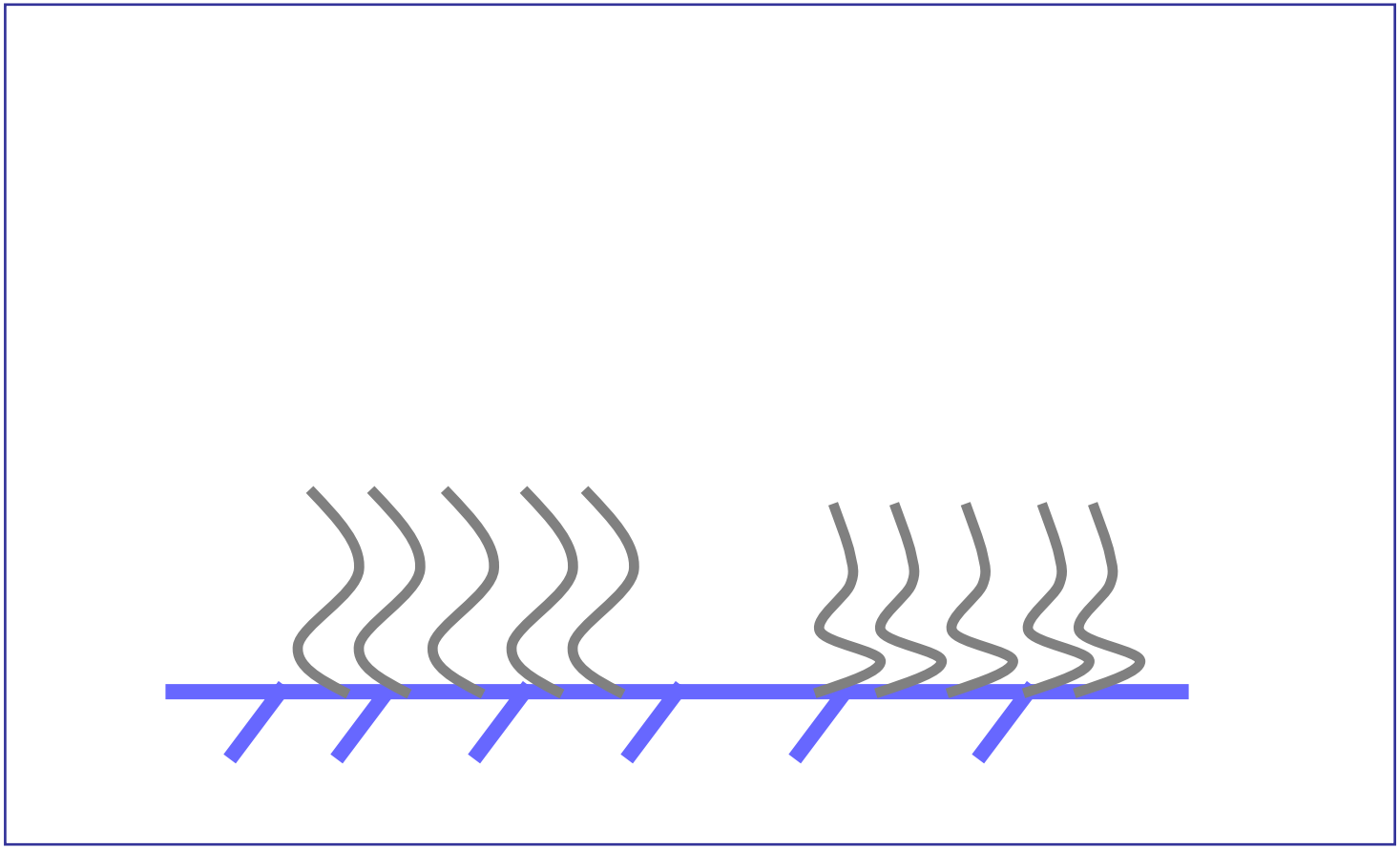
Direct Labeling (8)



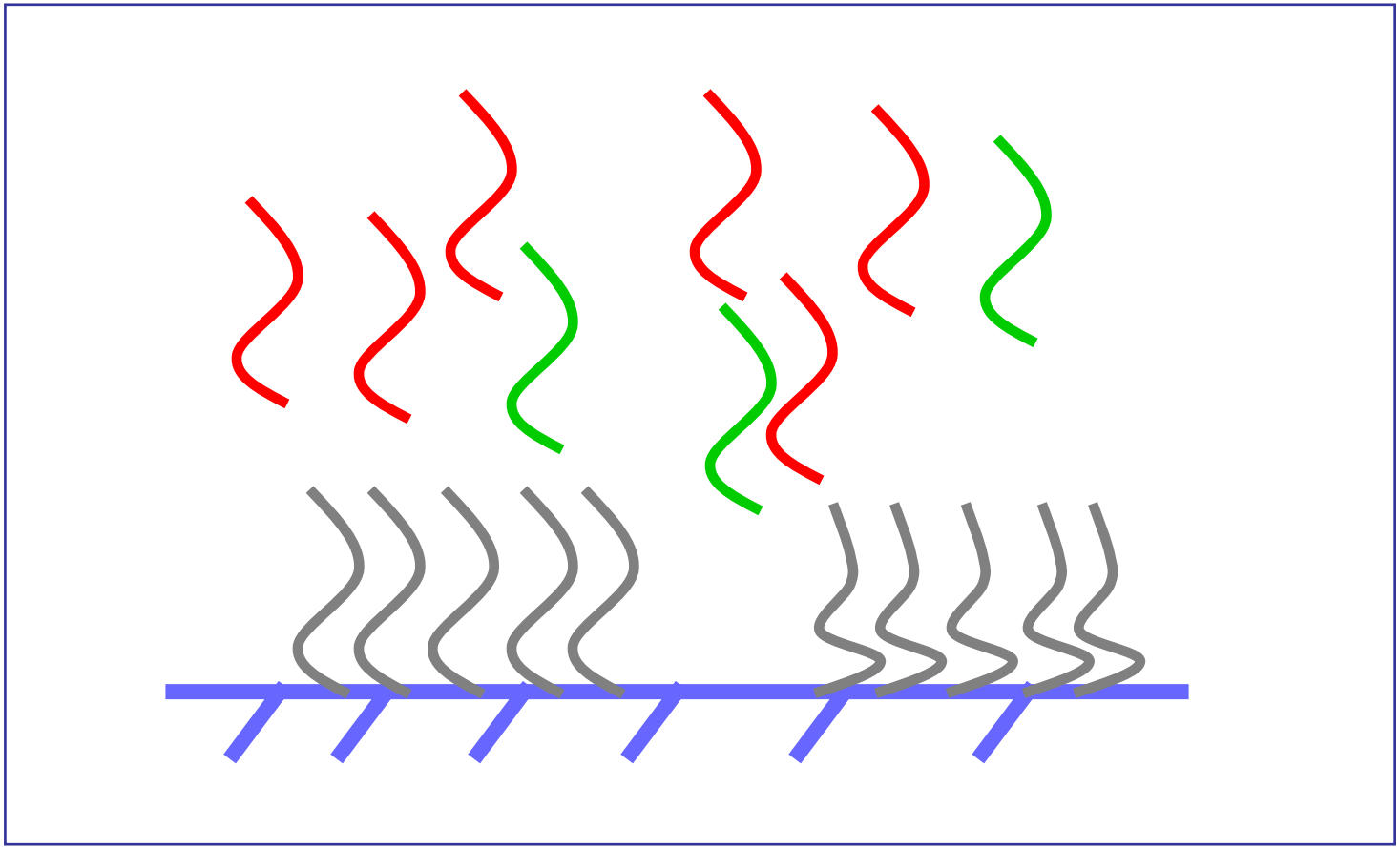
Direct Labeling (9)



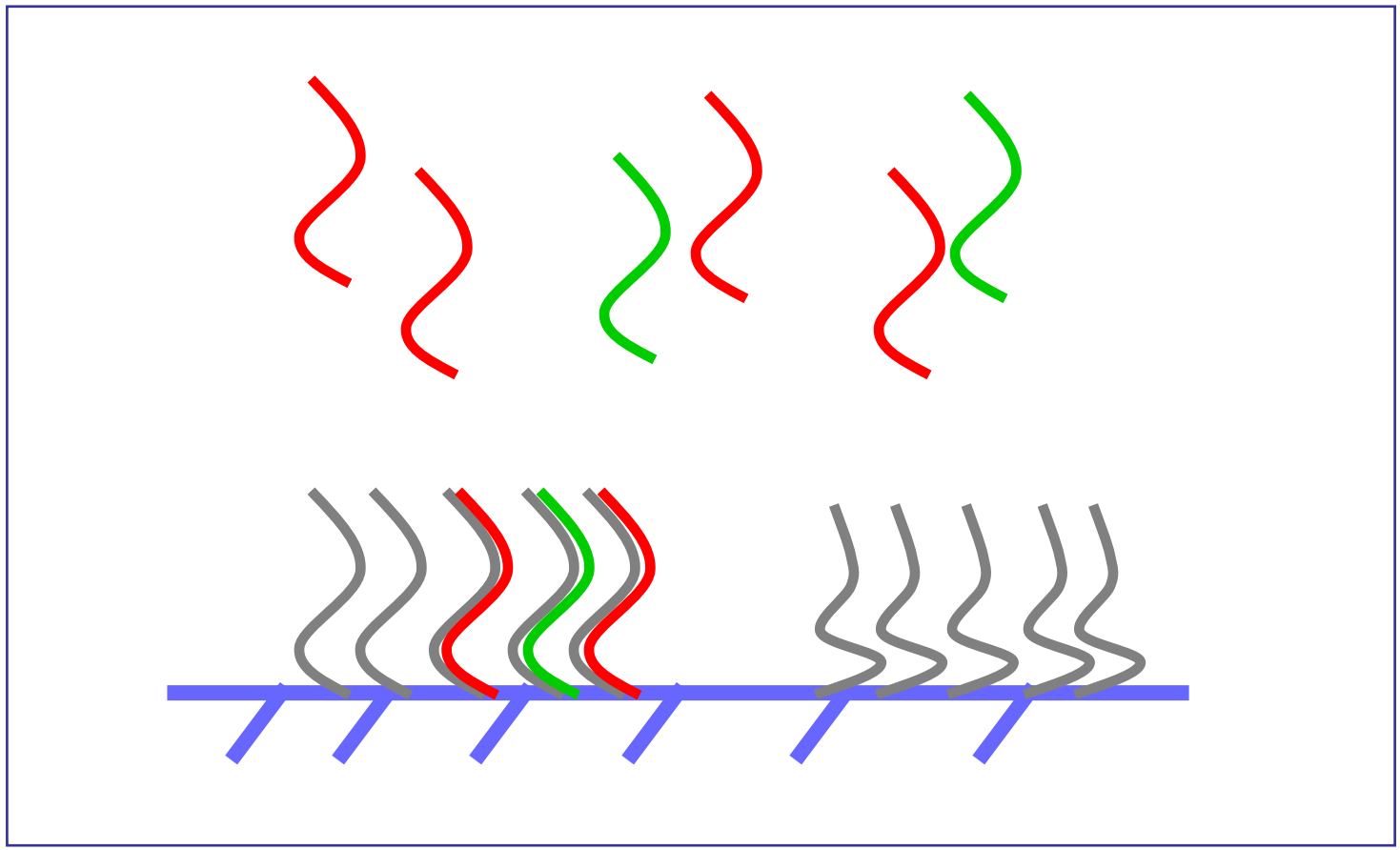
Hybridization (1)



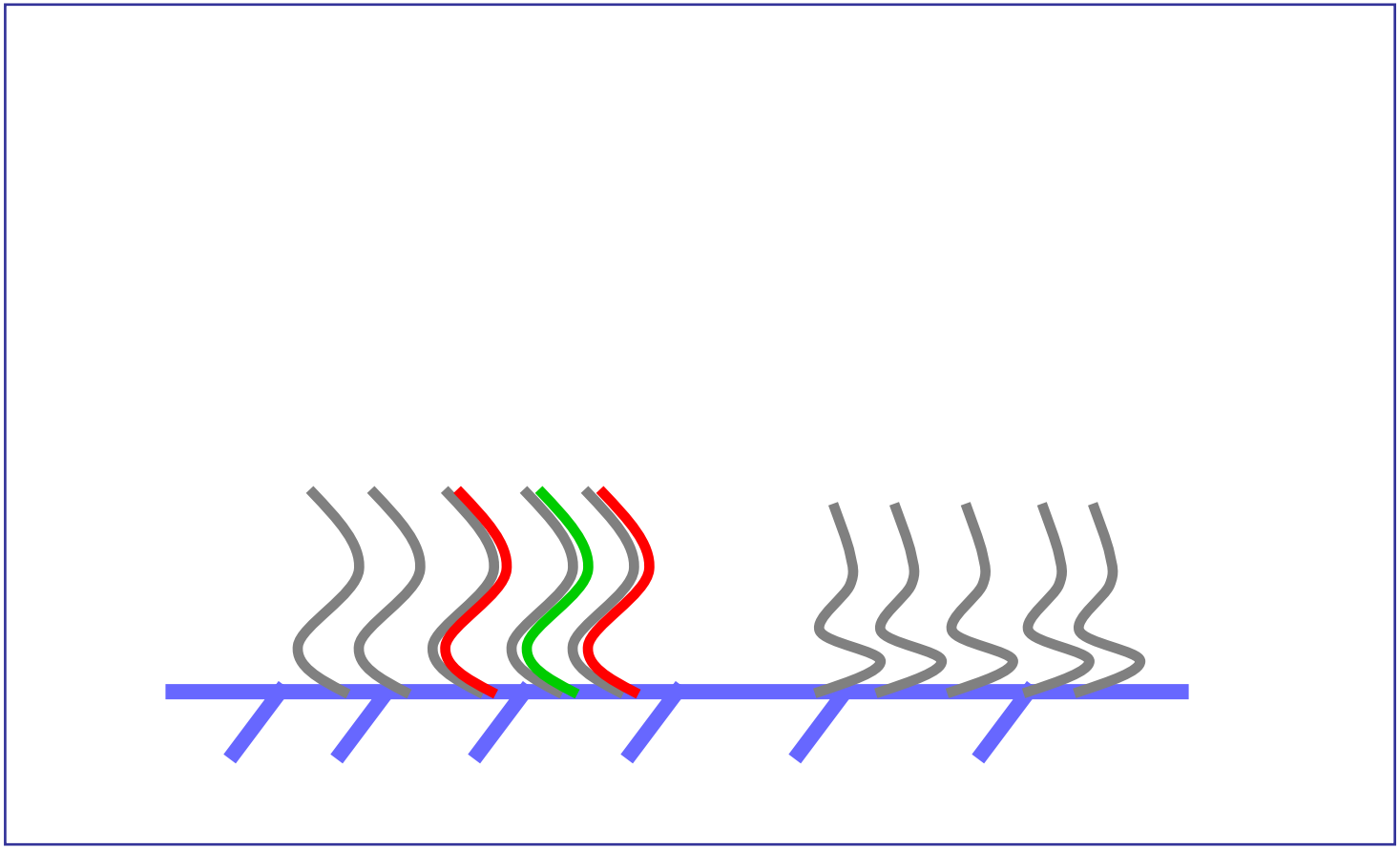
Hybridization (2)



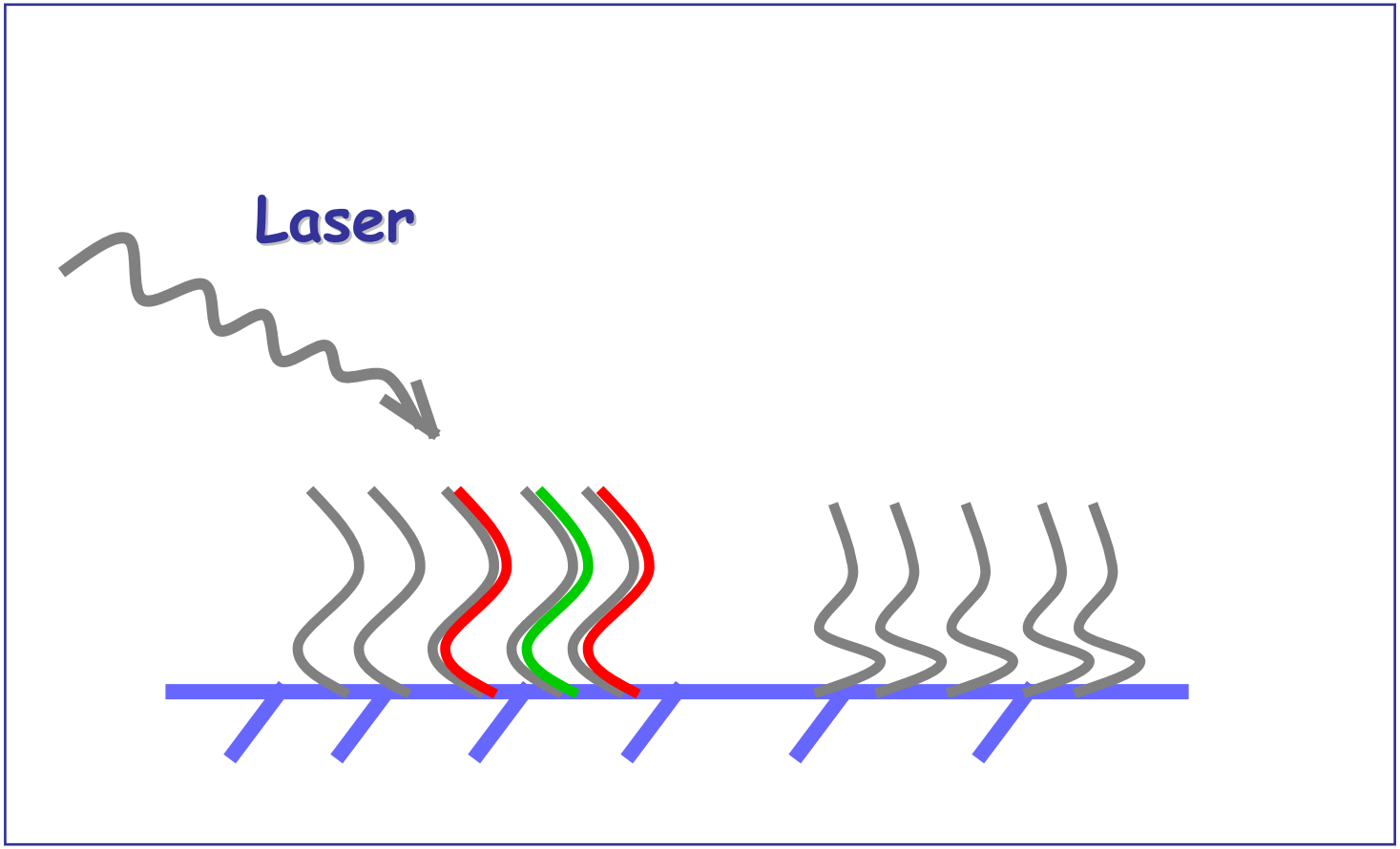
Hybridization (3)



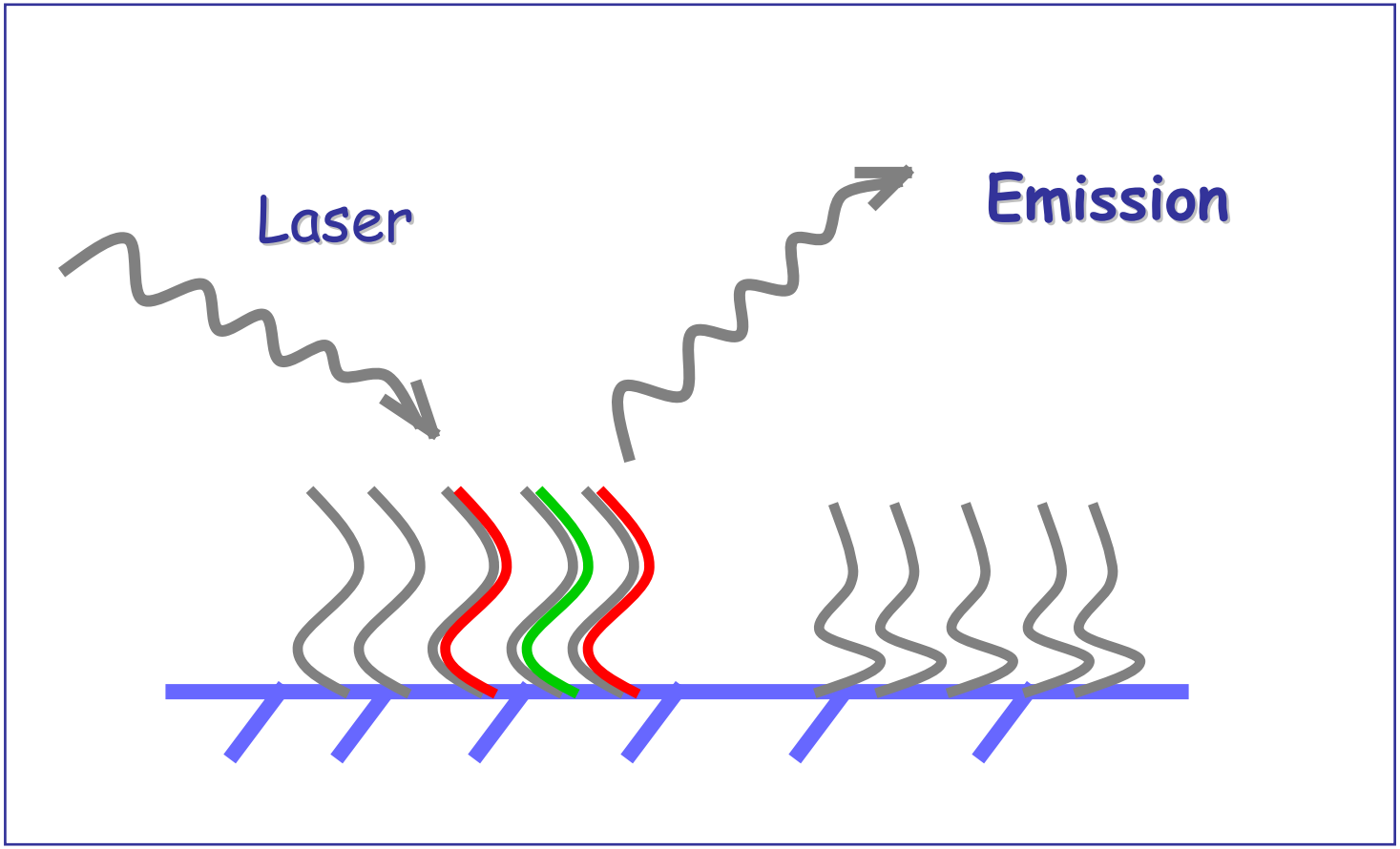
Wash



Laser Scanning (1)



Laser Scanning (2)



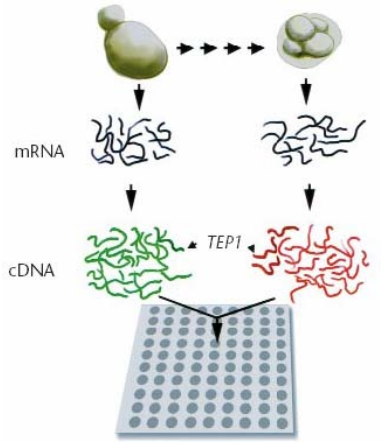
Microarray Data

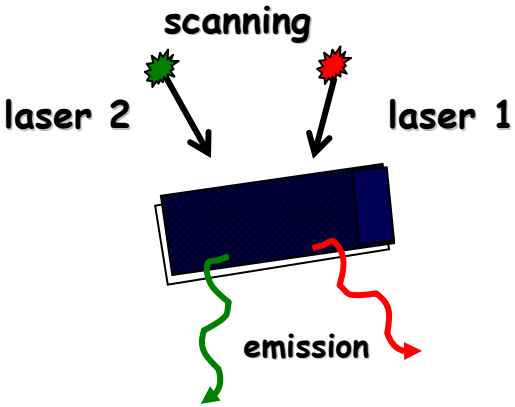
	ArrayExp1	Expt 2	Expt 3	…	Expt n
Gene1	$\log_2$(Cy5/Cy3)				
Gene2					
.					
.					
.					

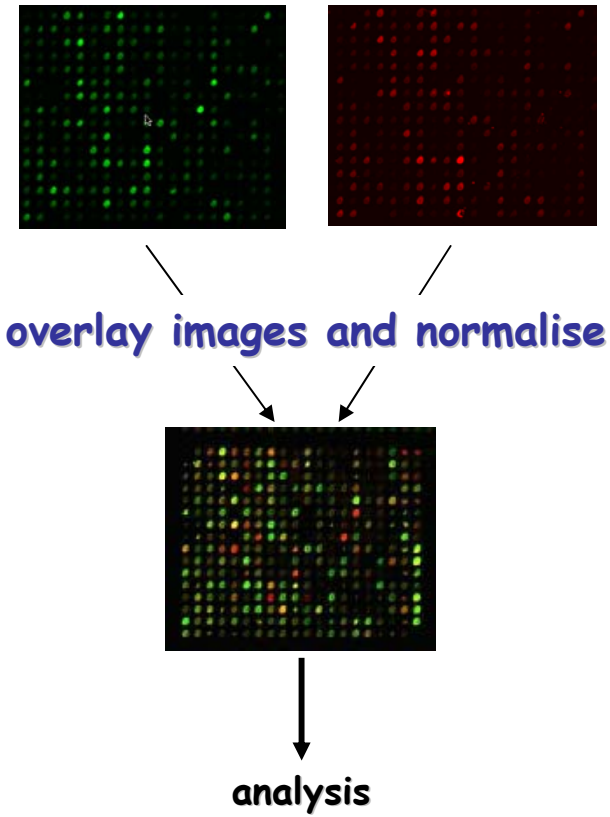
Intensity of control: green dye Cy3

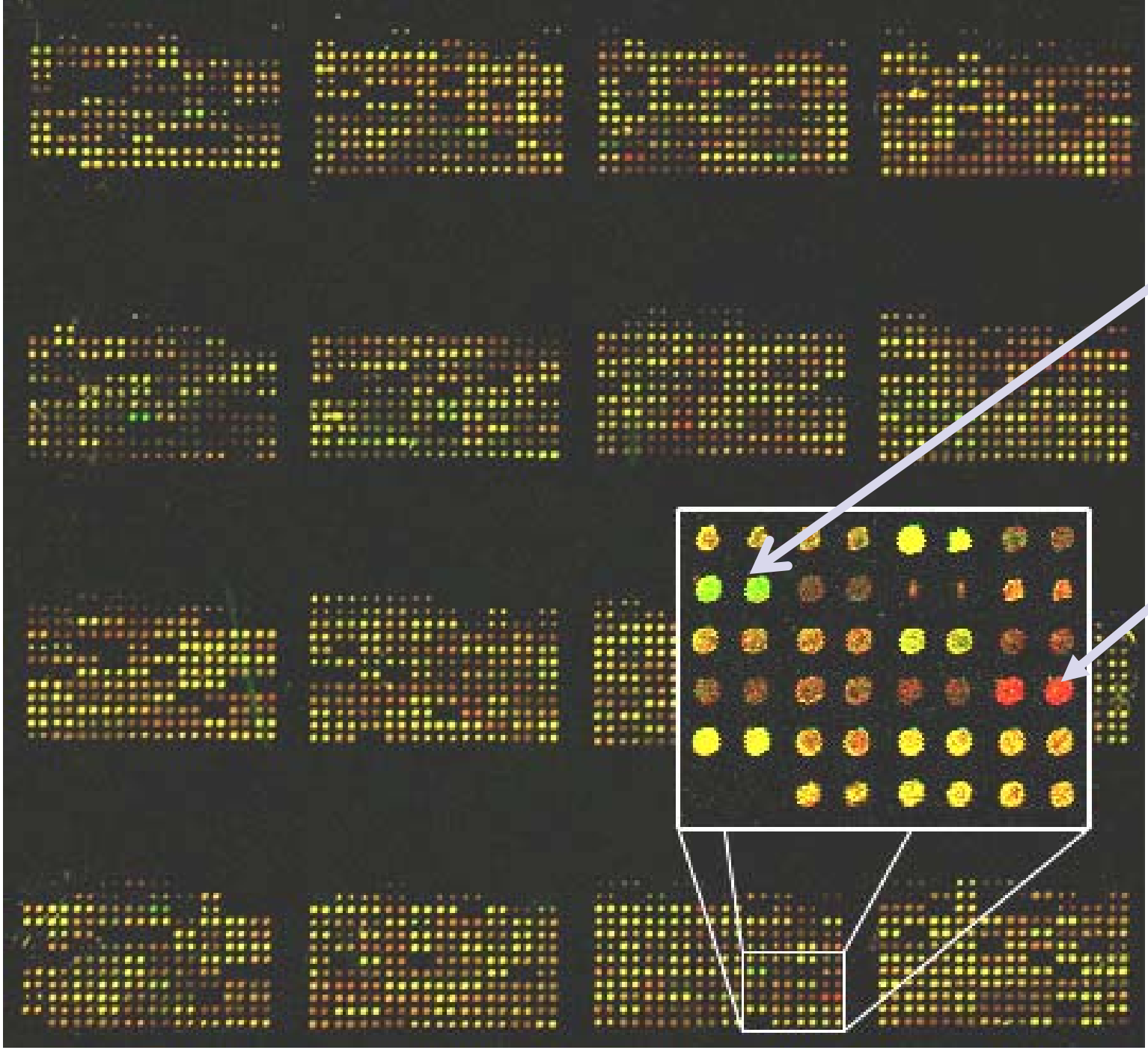
Intensity of experimental sample: red dye Cy5

The Output: the Image Raw Data









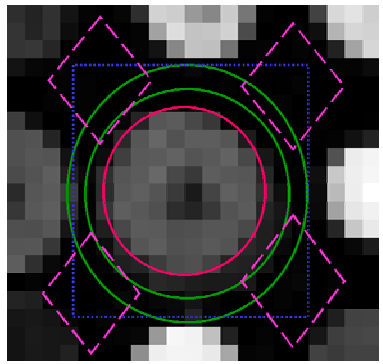
Gene D
Over-
expressed
in normal
tissue

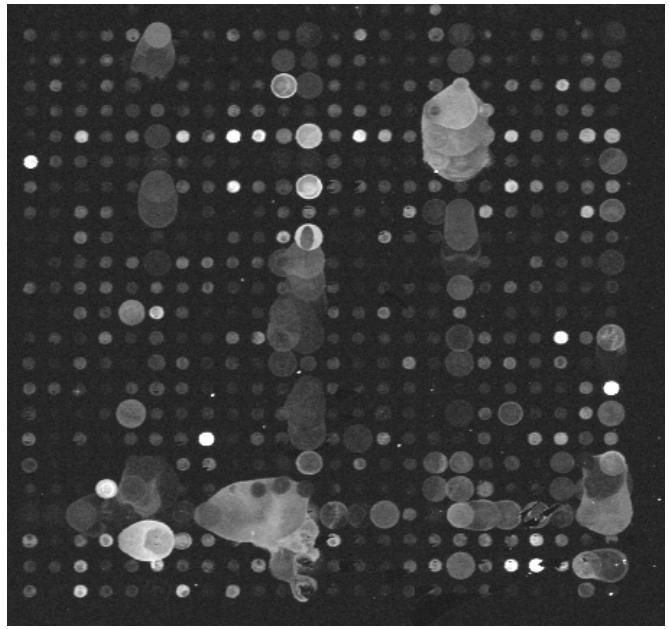
Gene E
Over-
expressed
in tumour

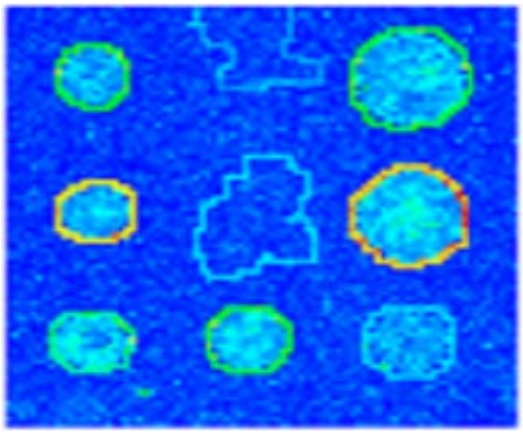
- Biomarkers of prognosis
- Genes affecting Treatment Response

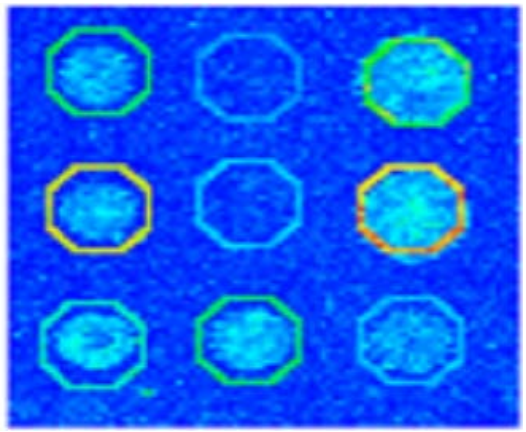
Problems in Image Analysis

- Computational methods
  - Noise
  - Spot detection and intensity
  - Alignment if overlay (grid align)



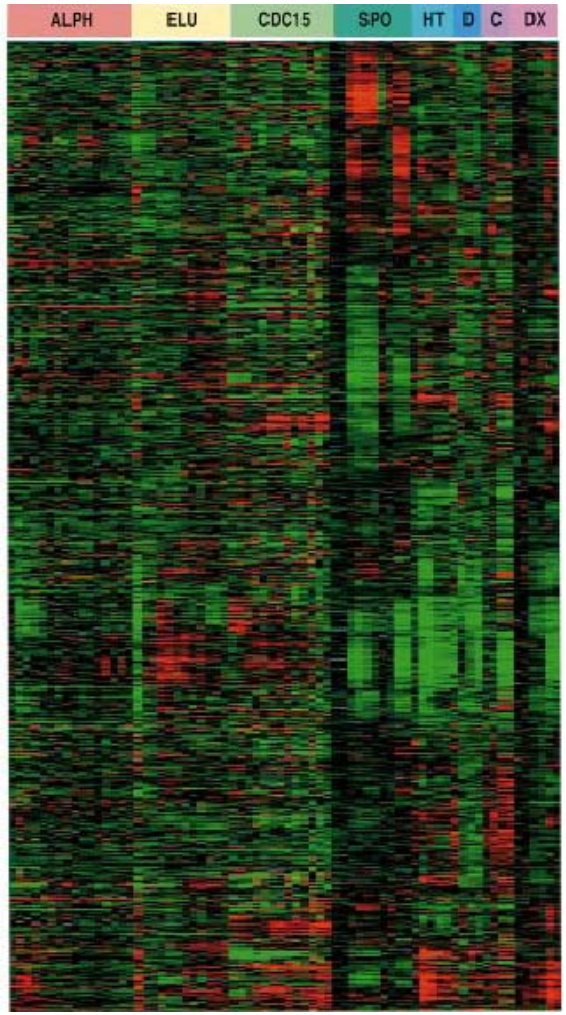




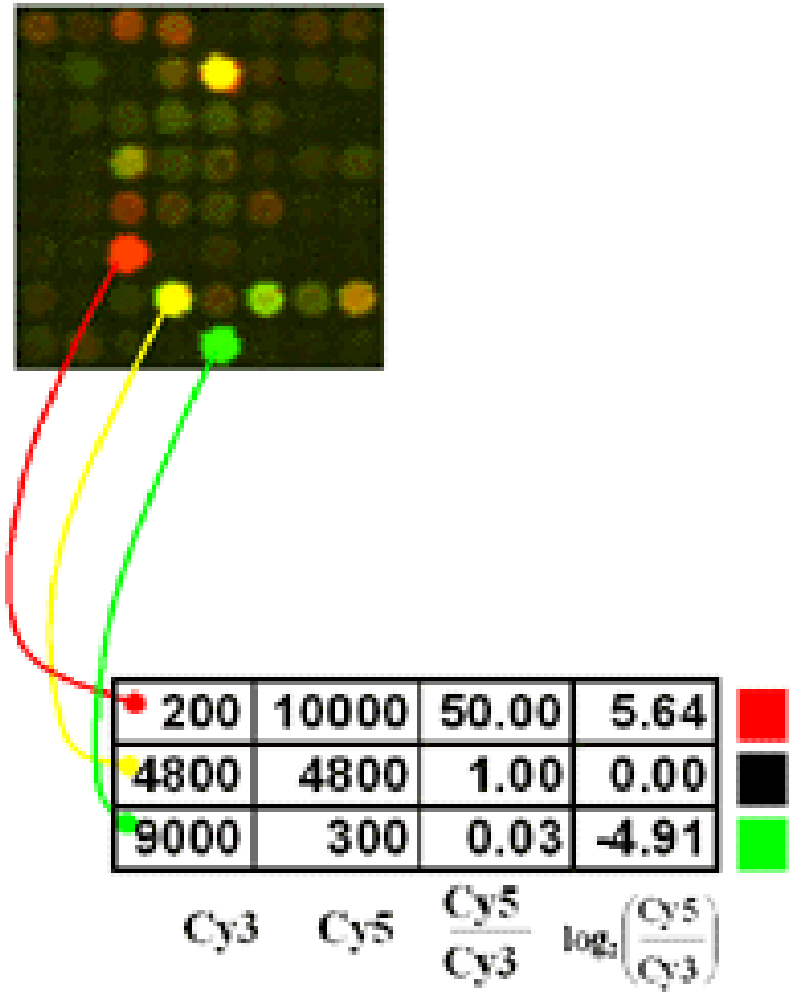


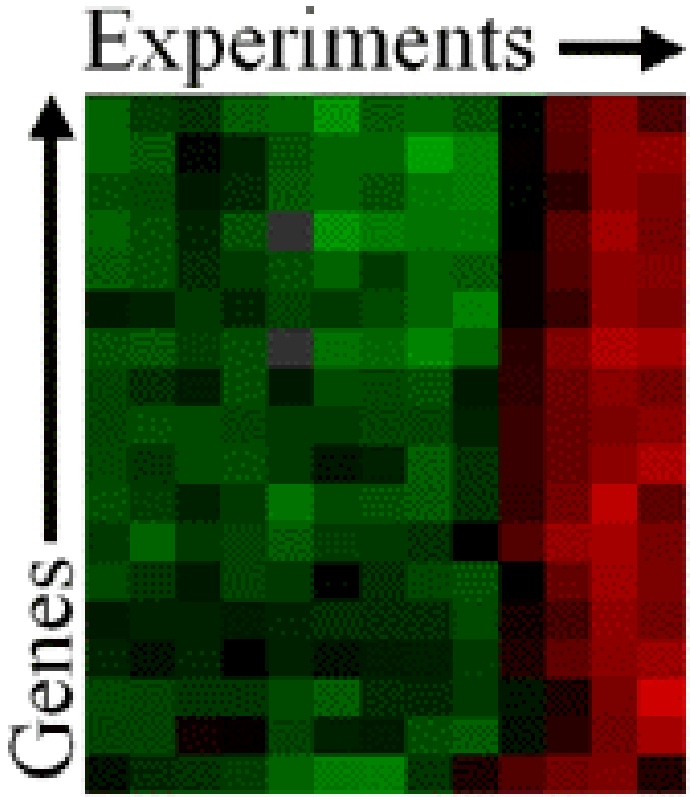
A Set of Experiments on Yeast...

- Each **row** represents **one gene**
  - Each **column** represents one **experiment**
  - The columns have been organized into **related sets of experiments**
- The **colors** indicate gene **activity** (from high to absent)
  - Numerical data **available**



Display of Datasets





From Gibson & Muse 2001

"Dye Swap"

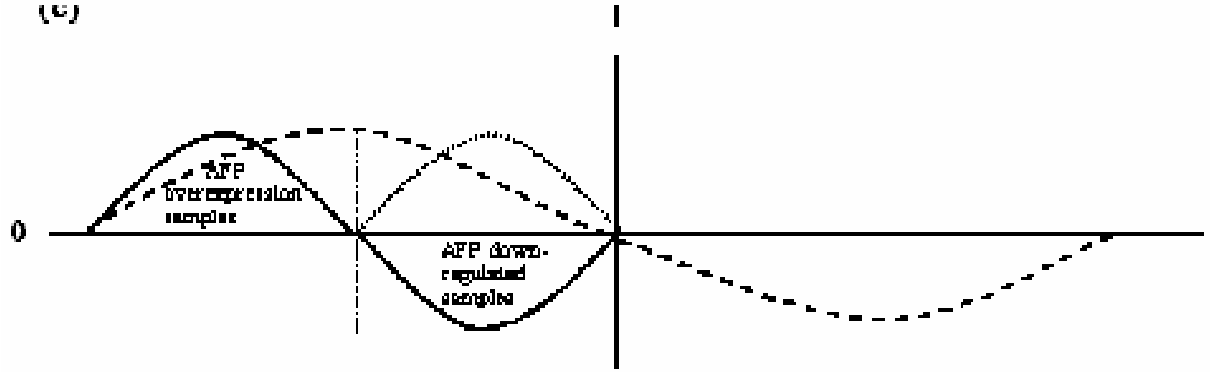
- One slide with **experimental** sample labeled with **Cy5**, and **reference sample** labeled with **Cy3**
  - **"Straight"**
- Replicated slide with **experiment** sample
  - Labeled **with Cy3**, and reference sample labeled with Cy5
  - **"Swap" or "switch"**
- ***A good microarray***
  - Self-hybridization test

Microarray Data Mining

- Place genes into **clusters** with **similar expression profiles** (Eisen *et al.* 1998)

- **Clustering** implies **co-regulation** ⇨ the genes are involved in **a similar biological process**
  - How individual genes respond to certain **treatments**
  - The level of **coordinate regulation** of gene expression **on the genome-wide scale**
  - Clustering process groups **unknown genes** with **annotated genes** ⇨ to **formulation of hypotheses** concerning **the possible function of the unknown genes**
    - *E.g.*, to identify genes expressed **at different stages** of the cell cycle, in a circadian manner

Data Mining on SMD Microarray Database



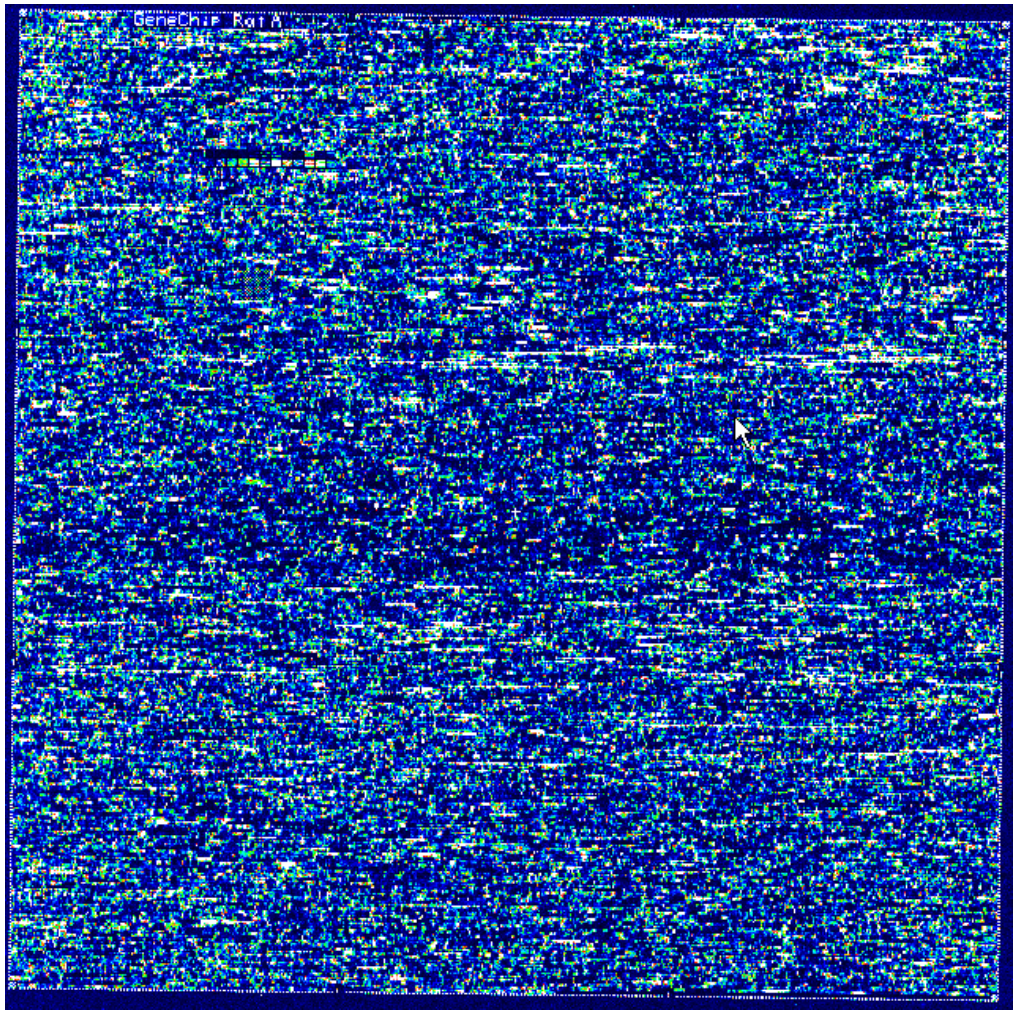
HCC samples (n=82)

Normal samples (n=74)

Gene/ESTs	Chr. localization	SAGE tags in liver tumor per 200,000	Subcellular localization Based on GO	Signal peptide (prediction*)
AFP	4q11-q13	3	Extracellular	+1~54
ATAD2	8q24.13	4	Unknown	-
CKS1B	1q21-22	68	Unknown	-
LOC54499	1q22-q25	12	Unknown	+ (1-72)
PCNA	20p12.3	14	Nucleus	-
PIGC	1q23-q25	7	Endoplasmic reticulum membrane	-
PYGB	20p11.2-p11.1	12	Unknown	-
PYGO2	1q22	4	Nucleus	-
SCAMP3	1q21-q22	21	Membrane	-
SNX27	1q21.3	-	Unknown	-
STMN1	1p35-p36	7	Cytoplasma	-

*Signal P 3.0 (http://www.cbs.dtu.dk/services/SignalP/)

Affymetrix GeneChip® (1)

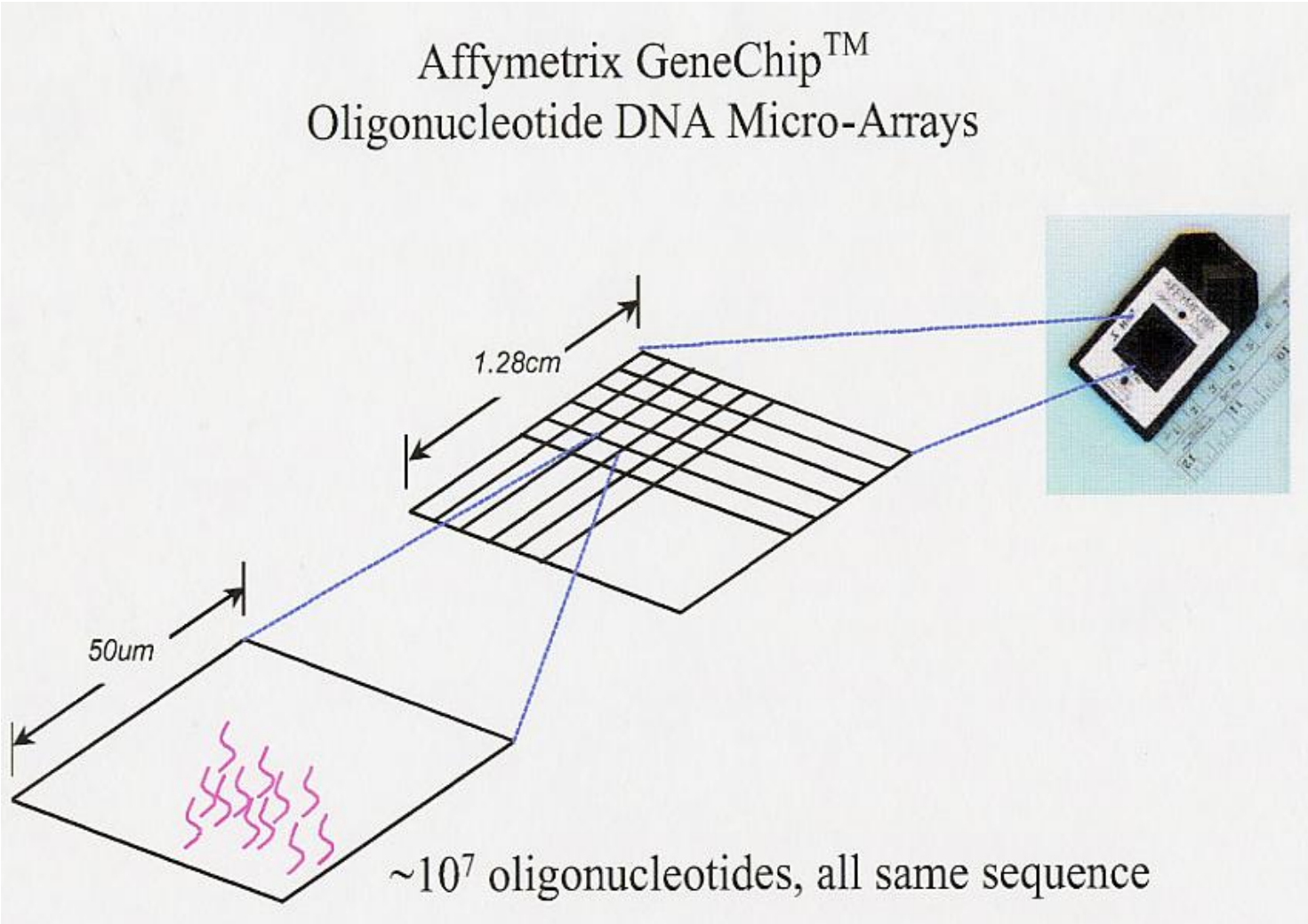


Limits: 1: 100,000 transcripts

~ 5 transcripts/cell

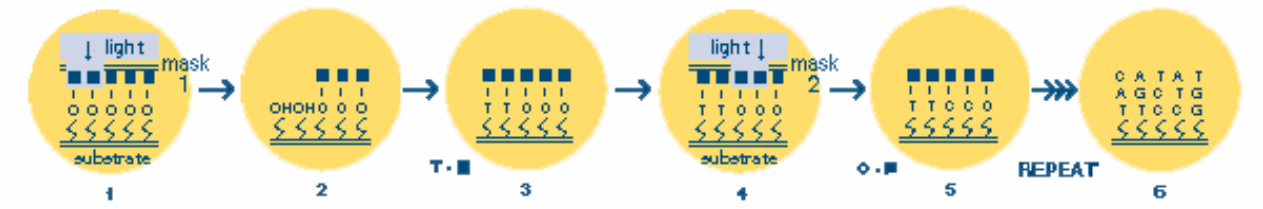
Affymetrix GeneChips® (2)

Properties

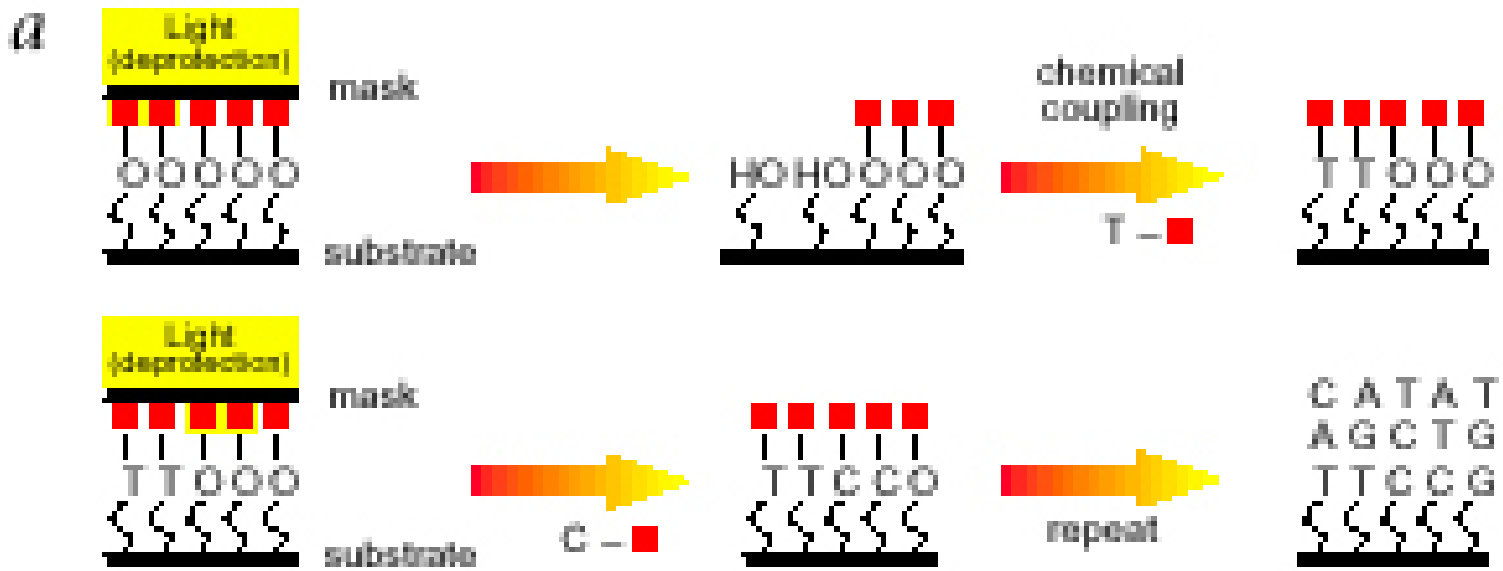


Affymetrix GeneChips® Production

Photolithgraphic method



Affymetrix GeneChip® (3)

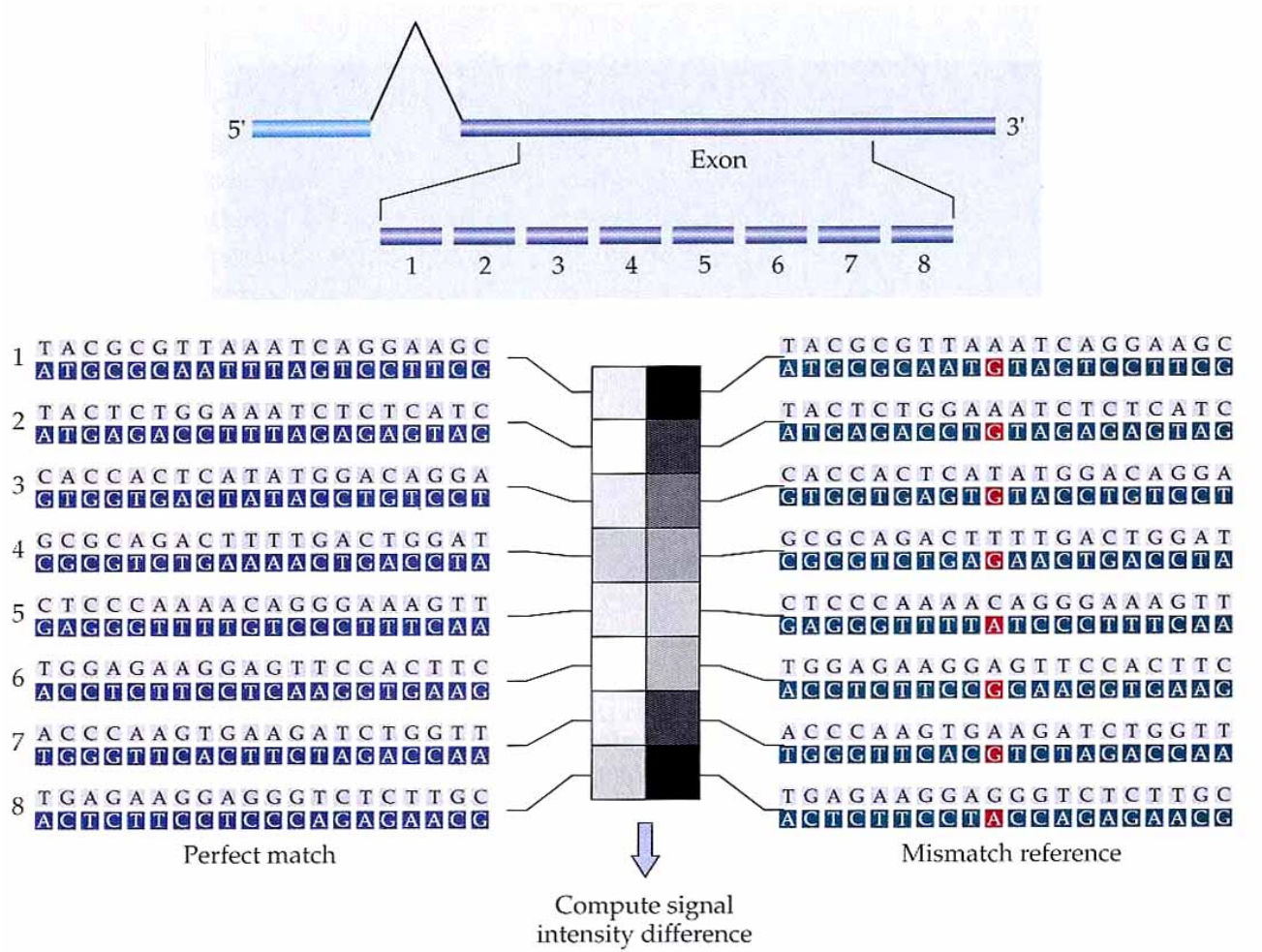


Nature Genetics 21: 2-24 1999

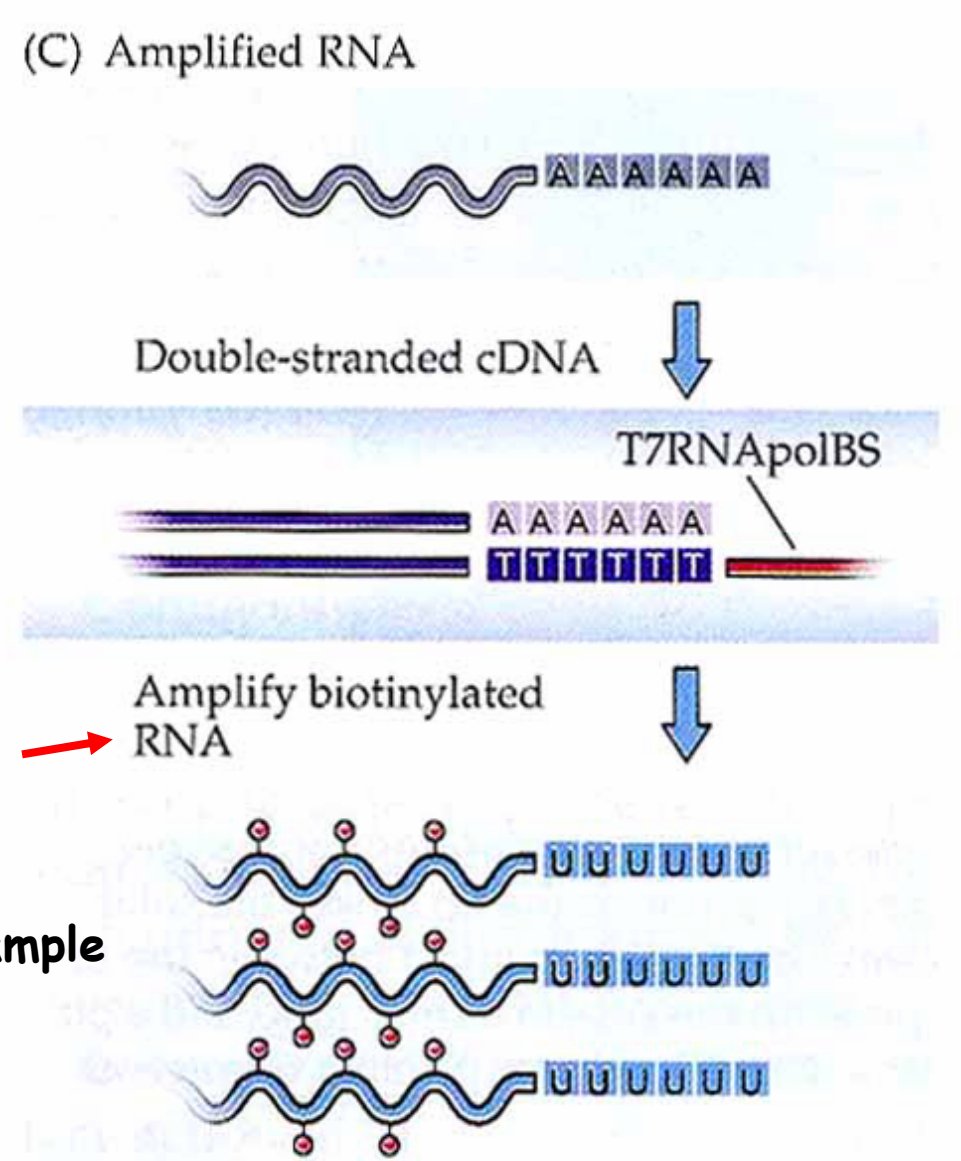
Affymetrix GeneChip® (4)

- Produced by synthesizing **tens of thousands** of short oligonucleotides ***in situ*** **onto** glass wafers
- **16-20 nucleotides** representing **each gene** on the array
- Each oligo on the chip is matched with an almost identical one, differing **only by one single base mismatch (PM vs. MM)**
  - Comparison of **target intensity** between the two partners oligonucleotides
- Measure of the ***absolute*** **level** of expression of genes

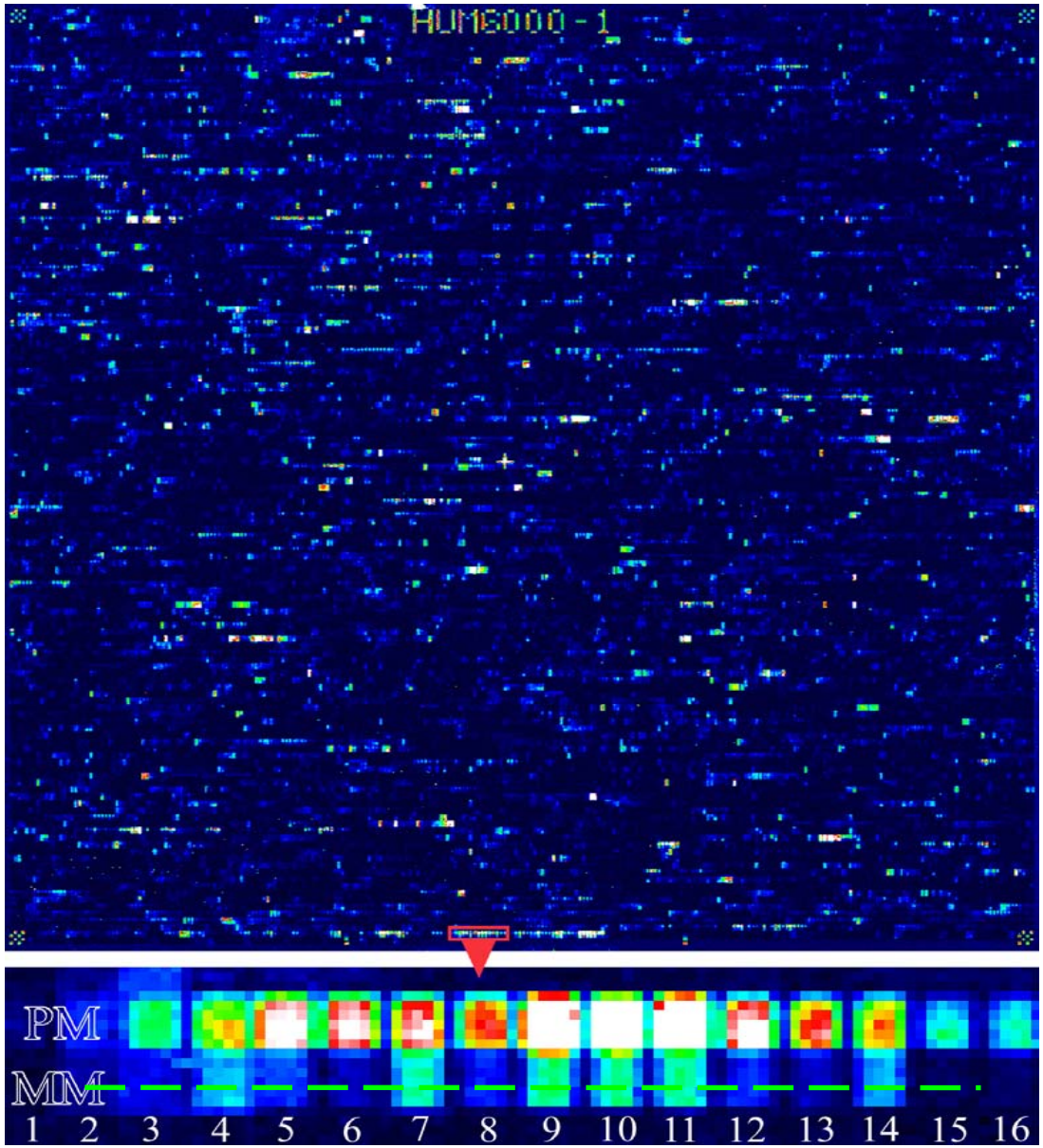
Principle of Oligonucleotide Arrays



From Gibson & Muse 2001



(C) In a RNA synthesis, **biotinylated, amplified RNA** is produced by *in vitro* **transcription** from a **T7 promoter** at the 5'-end of the primer used in **cDNA synthesis**. The biotin is then recognized by **fluorescently labeled streptavidin-phycoerythrin** compound **on the array** (not shown) used with Affymatrix® gene chips (from Gibson & Muse, 2001)

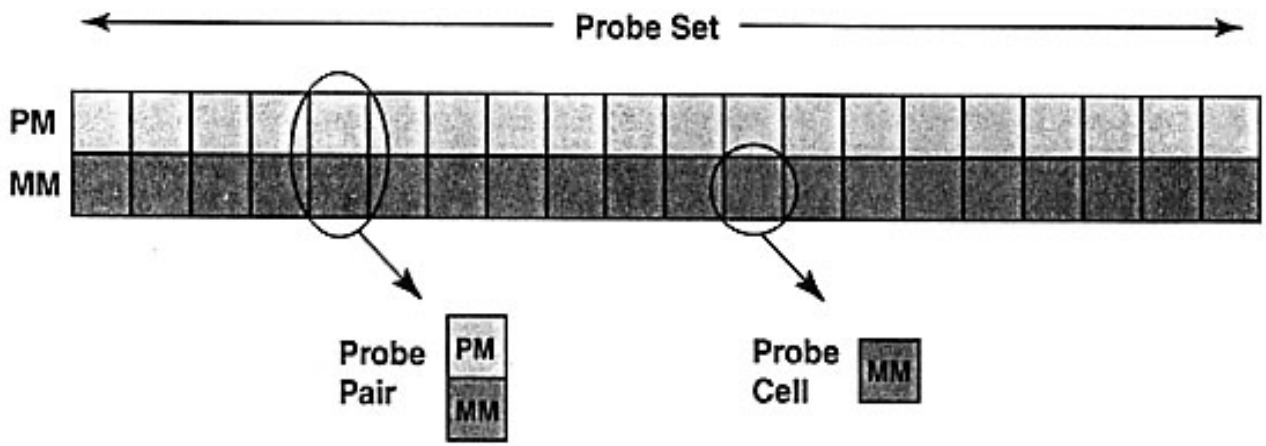


- The Affymetrix GeneChip®
- Probe: 25 bases long single stranded DNA oligos
- **Probe cell**: single square-shaped feature on an array containing one type of probe
  - 24-50 μm
- Millions of probe molecules

Dr. Karoly Mirnics

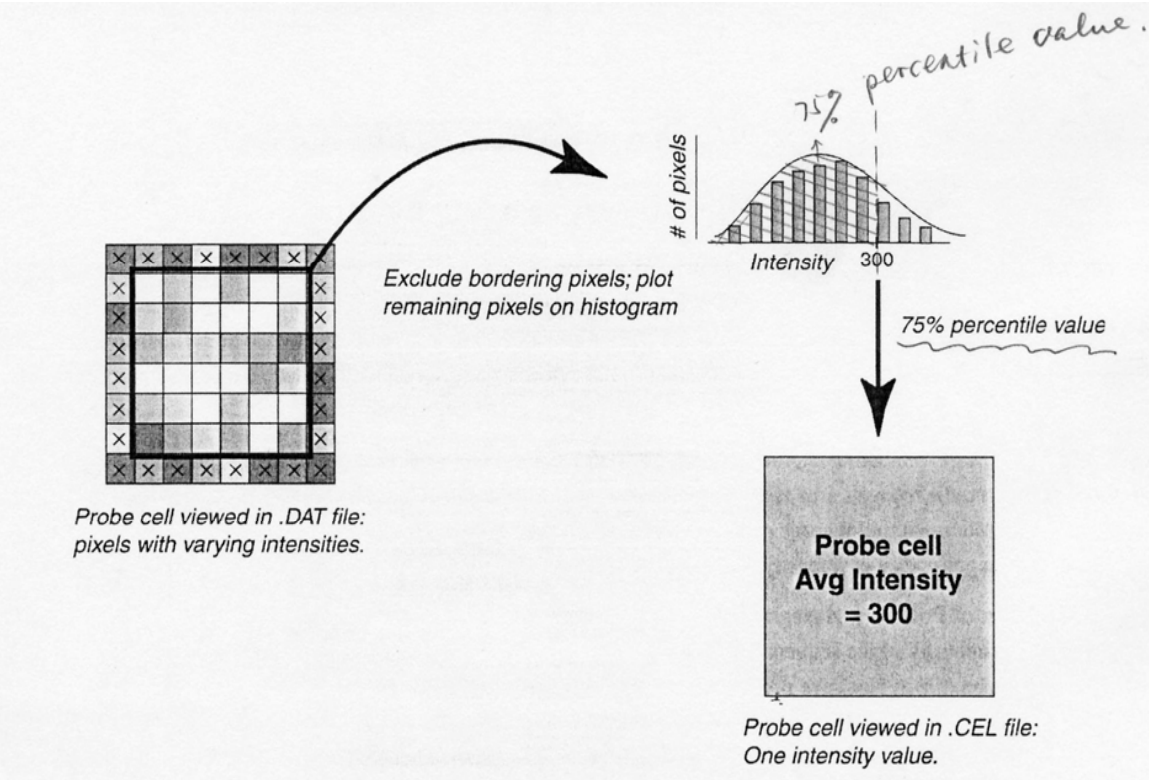
Affymetrix GeneChips® (5)

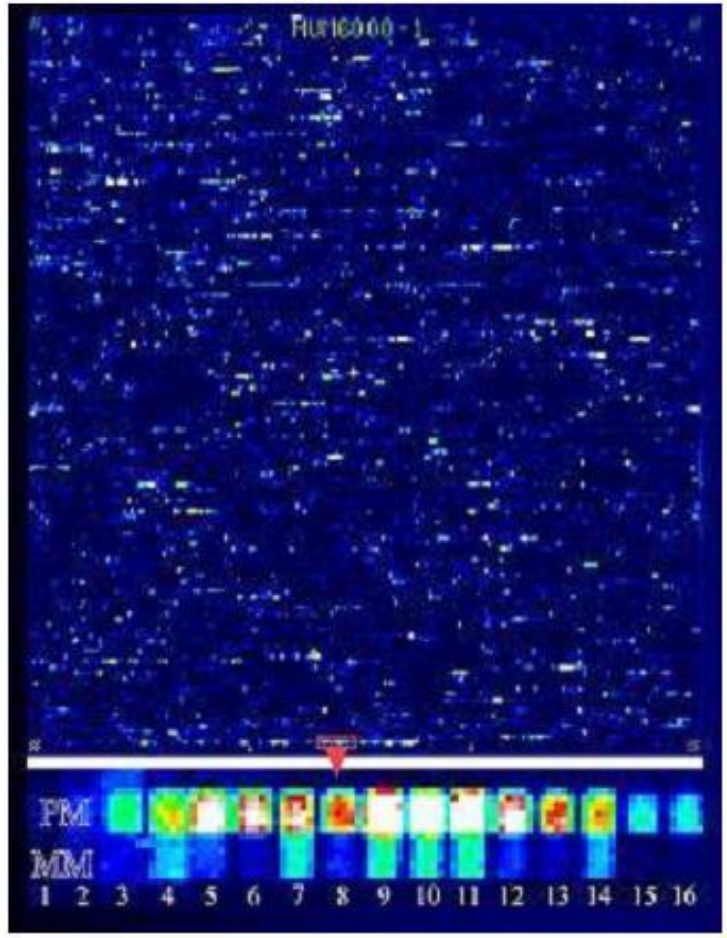
- **Probe set**
  - **16-20 probe pairs** that can **uniquely** identify **a transcript**
- **Probe pair**
  - One PM cell above MM cell
    - Perfect match probe/mismatch probe (**the 13th base change**)
    - MM probe designed to measures **non-specific binding** & noise
- Each probe: 25-mer; Probe cell: one million copies



Affymetrix GeneChips® (6)

.dat file: a huge image file
.cel file: cell intensity file





Absolute Call Metrics

- The number of positive & negative probe pairs, PM, MM intensities are used to derive absolute call metrics for **every transcript**
  - Positive Fraction = # positive probe pairs/# total probe pairs
  - Positive/Negative Ratio = # positive probe pairs /# negative probe pairs
- Averages **across probe set**, the Log of **PM/MM intensities for each probe pair** (describes hybridization performance)
  - Log Avg Ratio=10*[Σ log(PM/MM)]/Pairs in Ave)
- Each of the three metrics is entered into **a decision matrix** to determine **the status of a transcript**
  - Absent, Marginal or Present?
  - Default values for **thresholds** have been established through empirical testing

Average Difference – Relative Measure of Expression Level

- Avg Diff
  - Used to determine **change in expression** of **a given gene between two experiments**
- **Avg Diff** = Σ (PM-MM)/(**Pairs in Avg**)
  - Comparison of **experimental data** usually done to **baseline data**
  - **Pairs in avg**
    - = pairs used (if # of probe pairs $\leq$8)
    - = "trimmed" (if # of probe pairs >8)

Why use GeneChips®?

- Highly **sensitive** (250,000 mRNA copies)
- **Quantitative** data
- Small amount of starting material (~**50ng** mRNA)
- Redundant immobilized probe features

Affymetrix - Products - GeneChip® Arrays - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://www.affymetrix.com/products/arrays/index.affx 移至 連結

Google Affymetrix 搜尋 PageRank 20 已攔截 檢查 選項

http://www.affymetrix.com/products/arrays/index.affx

PRODUCTS & APPLICATIONS | SUPPORT | NETAFFX | SCIENTIFIC COMMUNITY | CORPORATE 日本語

APPLICATIONS

RESEARCH AREAS

DNA

-> Whole Genome

-> Targeted Genotyping

-> Sequence Analysis

Expression

-> Quantitation

-> Regulation

Clinical

-> Research & Trials

-> Molecular Diagnostics

Fields of Study

Cancer

PRODUCTS

-> GeneChip System

-> GeneChip Arrays

-> Assays & Reagents

-> Instruments

-> Software

-> NetAffx™ Analysis Center

TECHNOLOGY

-> Overview

-> Array Manufacturing

-> Combinatorial Chemistry

-> Probe Design and Selection

-> Molecular Inversion Probe

-> Mismatch Repair Detection

AFFYMETRIX PROGRAMS

GENECHIP® ARRAYS

Each high-density GeneChip array provides multiple, independent measurements for each transcript and genotyping call. Multiple probes mean that you get a complete data set with accurate, reliable, reproducible results from every experiment.

DNA Analysis Arrays

Catalog Arrays

-> Human Mitochondrial Resequencing Array 2.0

-> Mapping 100K Set

-> Mapping 10K 2.0 Array

-> Mapping 10K Array

-> Mapping 500K Array Set

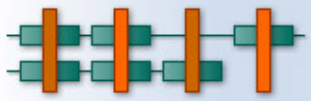
-> SARS Resequencing Array

Custom DNA Array Program

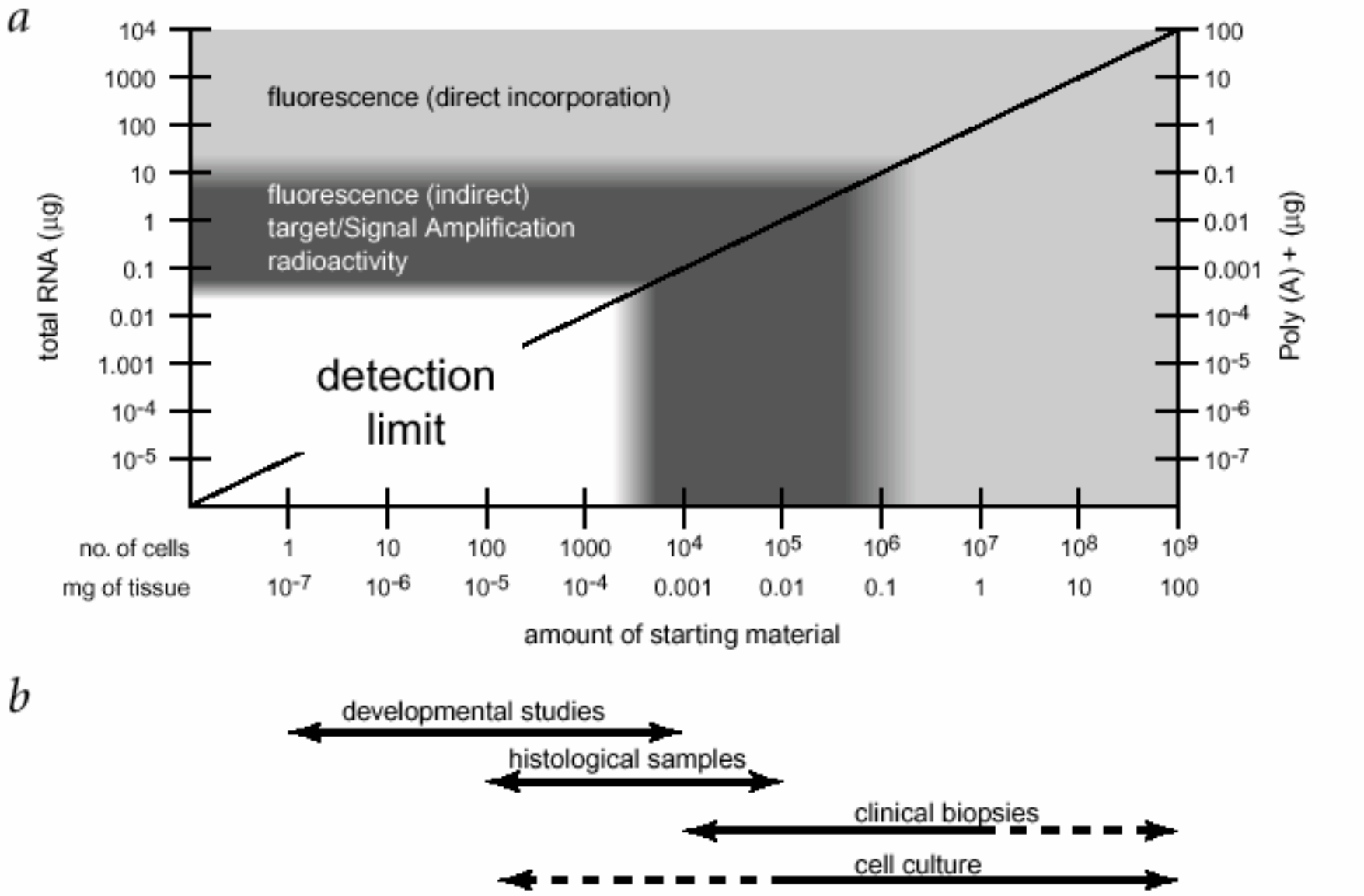
-> CustomSeq® Resequencing Arrays

Expression Analysis Arrays

Catalog Arrays

GENE	TRANSCRIPT	GENOME
		
Robust, simple representation focusing on the 3' ends.	Genome-wide, exon-level analysis on a single array — a survey of alternative splicing and gene expression.	High-density tiled microarrays for transcript mapping and chromatin immunoprecipitation.
Human Arrays...	-> Exon Arrays	Human Arrays...
Mouse Arrays...		

網際網路

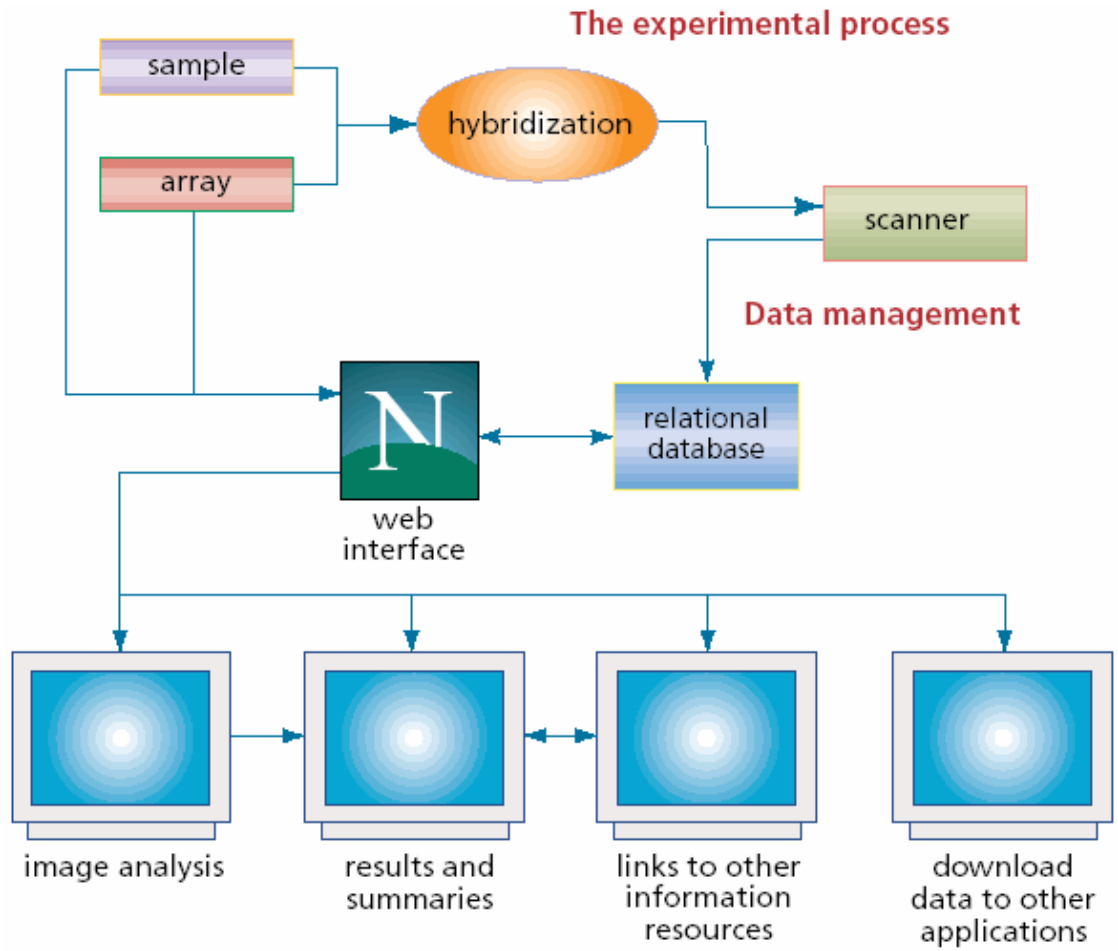

a
fluorescence (direct incorporation)
fluorescence (indirect)
target/Signal Amplification
radioactivity
detection
limit
total RNA (µg)
Poly (A) + (µg)
no. of cells
mg of tissue
amount of starting material
b
developmental studies
histological samples
clinical biopsies
cell culture

Microarray Data on the Web

- Many groups have made their **raw data** available, but in many formats
- Some groups have created **searchable** databases…
- There are several initiatives to create **"unified" databases**
  - EBI: ArrayExpress
  - NCBI: Gene Expression Omnibus (GEO)
- Companies are beginning to sell microarray expression data (*e.g.* Incyte)

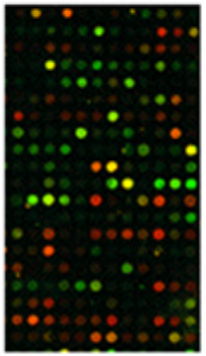
The Overall Process with Microarrays

- Microarray data has to be used in a larger frame of experimentation

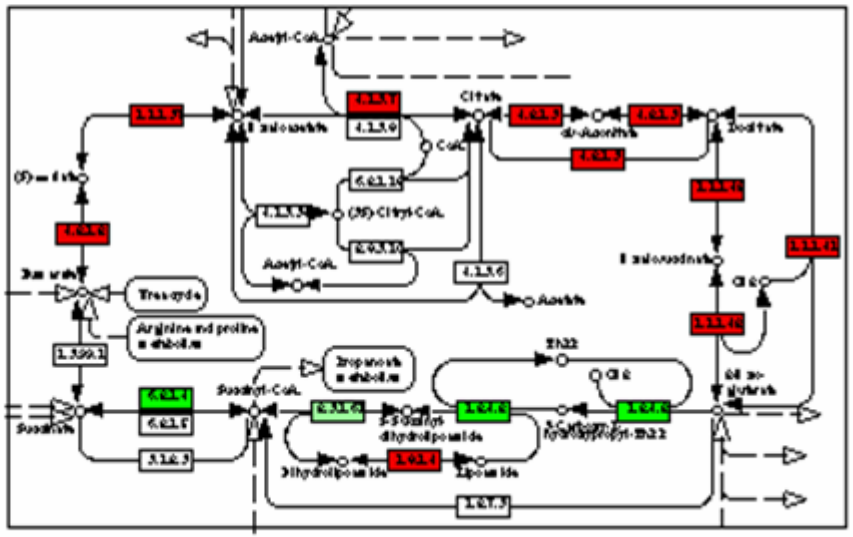


More Challenges?

Mapping Gene Expression Profiles on the KEGG Pathways



Microarray patterns



- The list of genes being **activated** or **inactivated** or that are unaffected when comparing **two samples** becomes more informative if the genes can be **mapped onto maps** from which **functions** can be deduced

Nature Biotechnology, 20:370-375, 2002

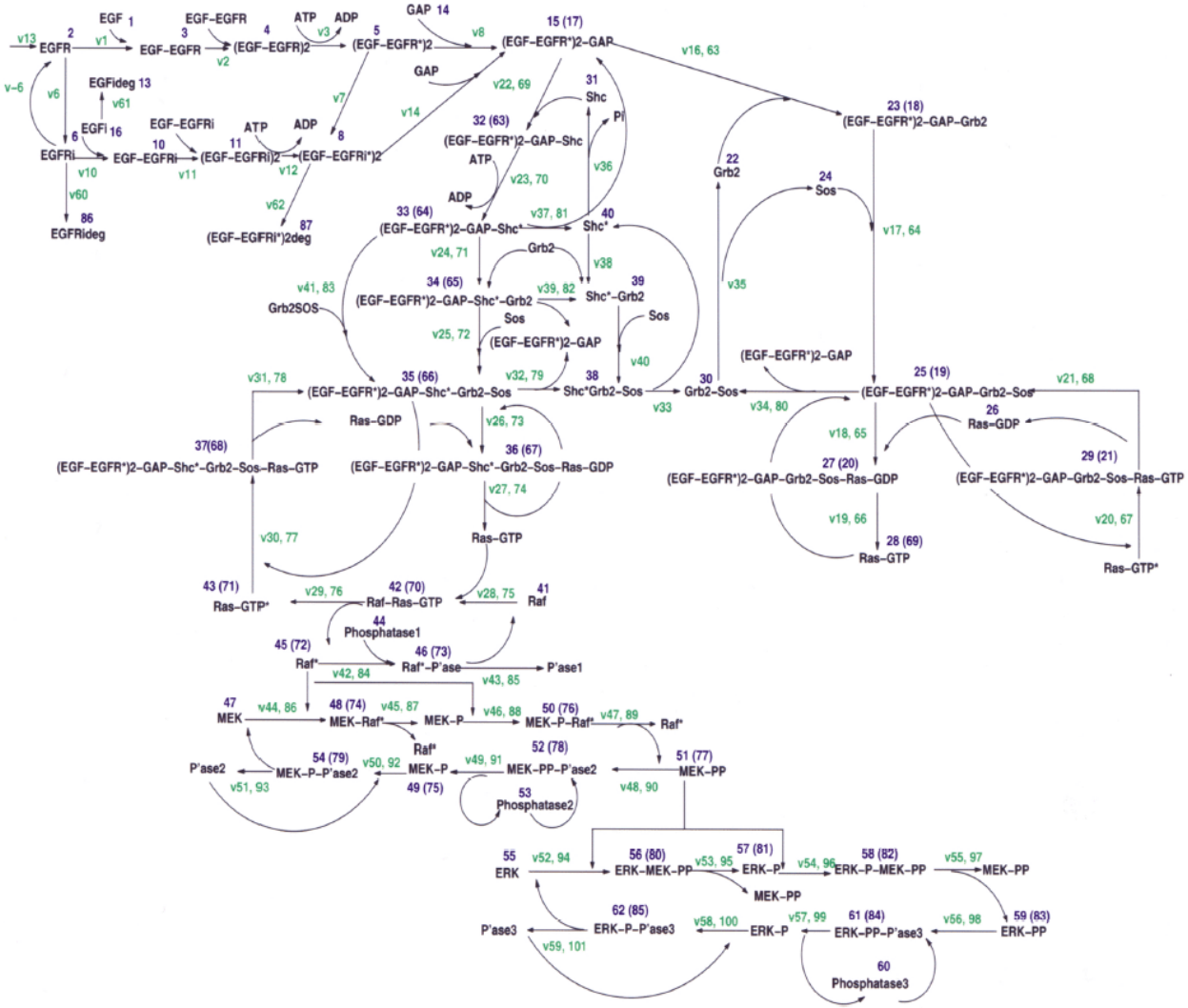
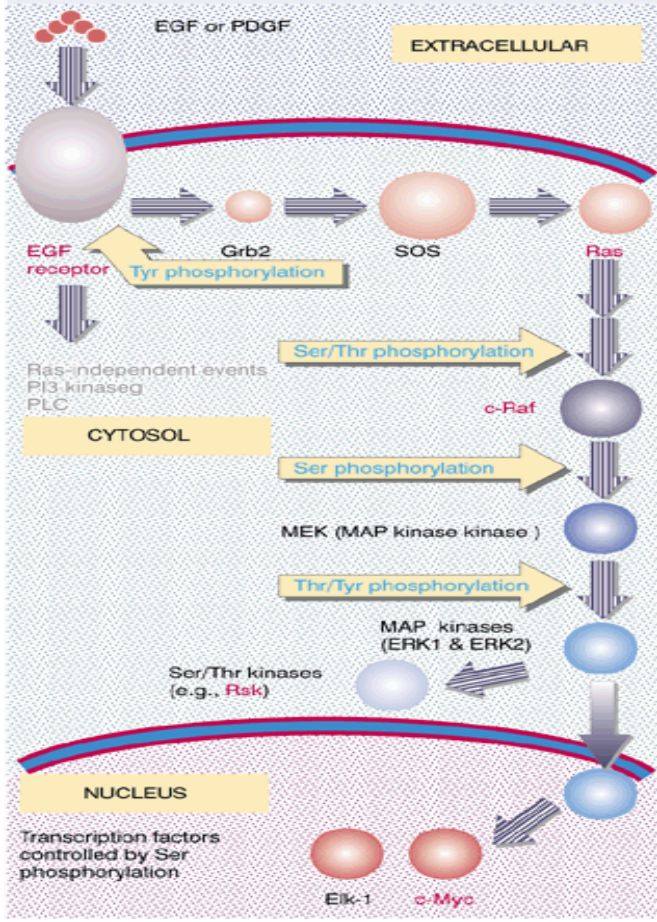


Figure 26.20 A common signal transduction cascade passes from a receptor tyrosine kinase through an adaptor to activate Ras, which triggers a series of Ser/Thr phosphorylation events. Finally, activated MAP kinases enter the nucleus and phosphorylate transcription factors. Missing components are indicated by successive arrows.

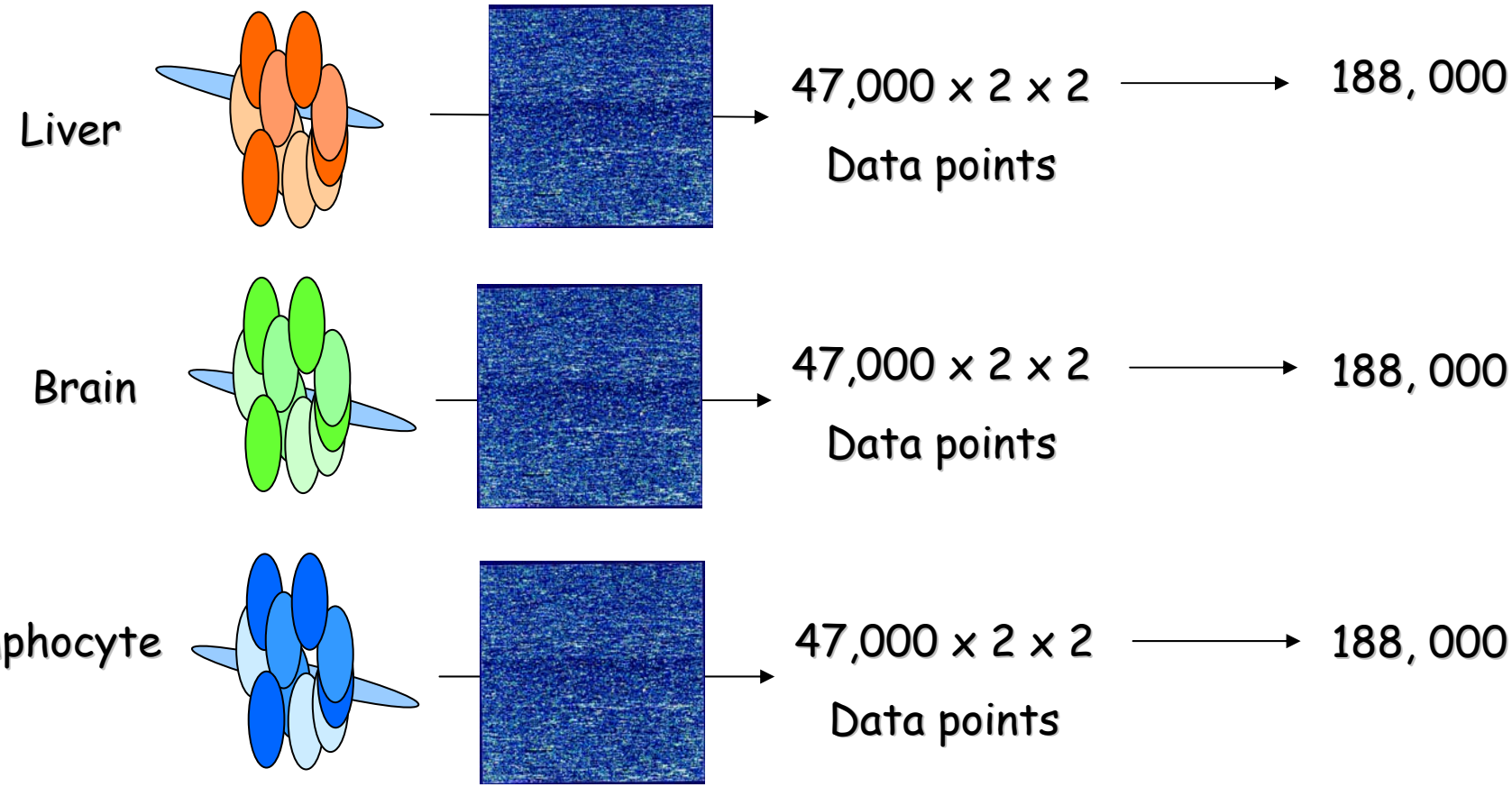


Microarray & GeneChip® Approaches

- **Advantages**
  - Rapid
  - Method & data analysis well described & supported
  - Robust
  - Convenient for directed and focused studies

- **Disadvantages**
  - **Closed system** approach
  - Difficult to correlated with **absolute transcript number**
  - Sensitive to **alternative splicing ambiguities**

Analysis (1)

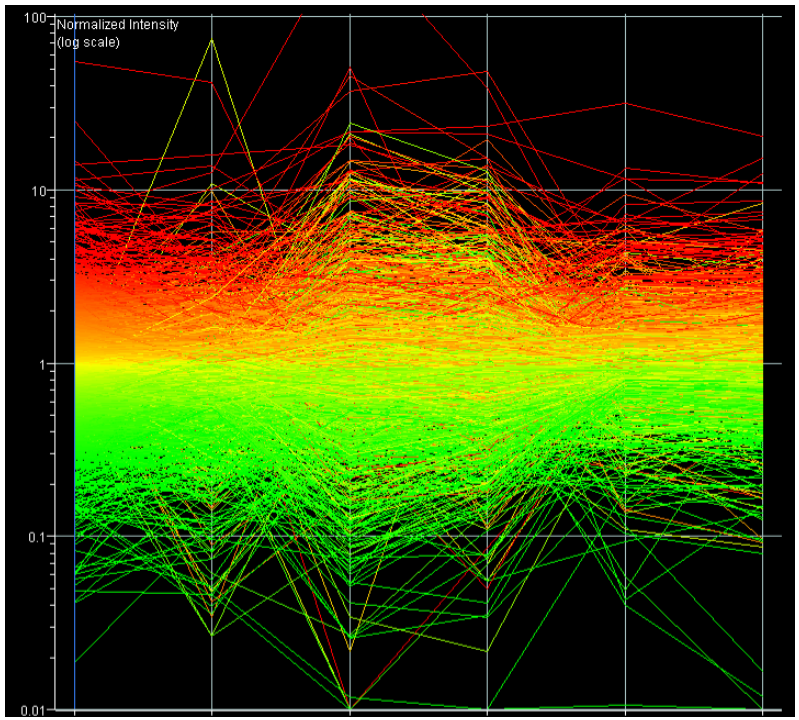

Liver
47,000 x 2 x 2
Data points
188, 000
Brain
47,000 x 2 x 2
Data points
188, 000
Lymphocyte
47,000 x 2 x 2
Data points
188, 000

Analysis (2)

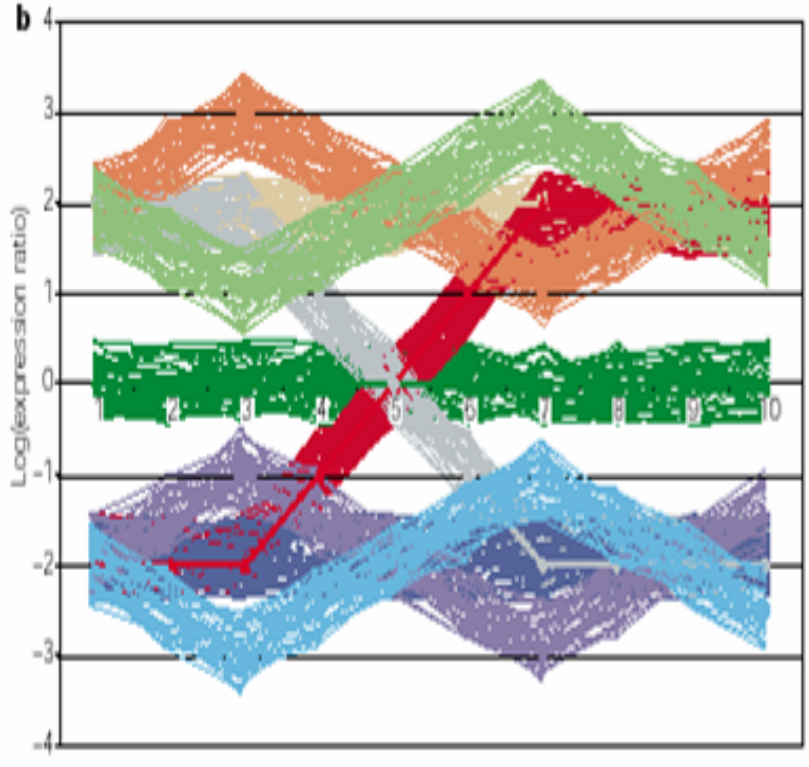
- **Essential problem**
  - Given a large dataset with technical and biological noise

- **To find**
  - **Transcripts**: **patterns** (common themes or **differences**) measures of robustness or some idea of uncertainty

  - Sample: **similarities** or **differences** between samples on global/multi-gene level



Liver Brain Lymphocytes



Which transcripts are **different?**

What are the patterns?

Biologists Nightmare: Statisticians Playground

- Characteristics of the expression profile data
  - High dimensionality
  - Sample number (n) low and observation number high (p)

  - Non-independence of observations
  - **Complex pattern**: visualization & extraction
  - Incorporation of contextual information

  - **Standardization** & data sharing
  - Integration of & with other data types

Analysis Methods (1)

- Classical **parametric** & **non-parametric** statistical tests for hypothesis testing

- **Unsupervised clustering**
  - Hierarchical clustering
  - K-means and Self-Organizing Maps (SOMs)

- **Classification (Supervised)**
  - *E.g.,* Machine learning & linear discriminant analysis

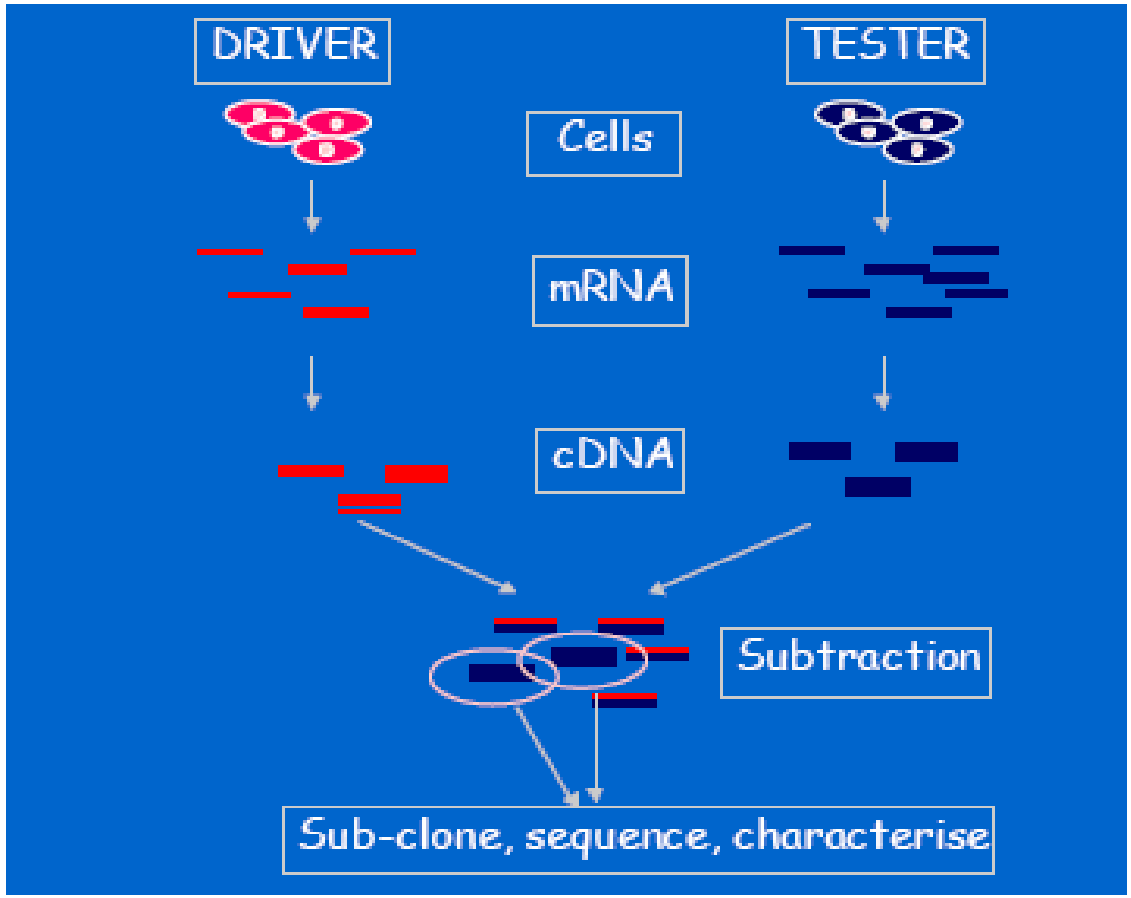
Analysis Methods (2)

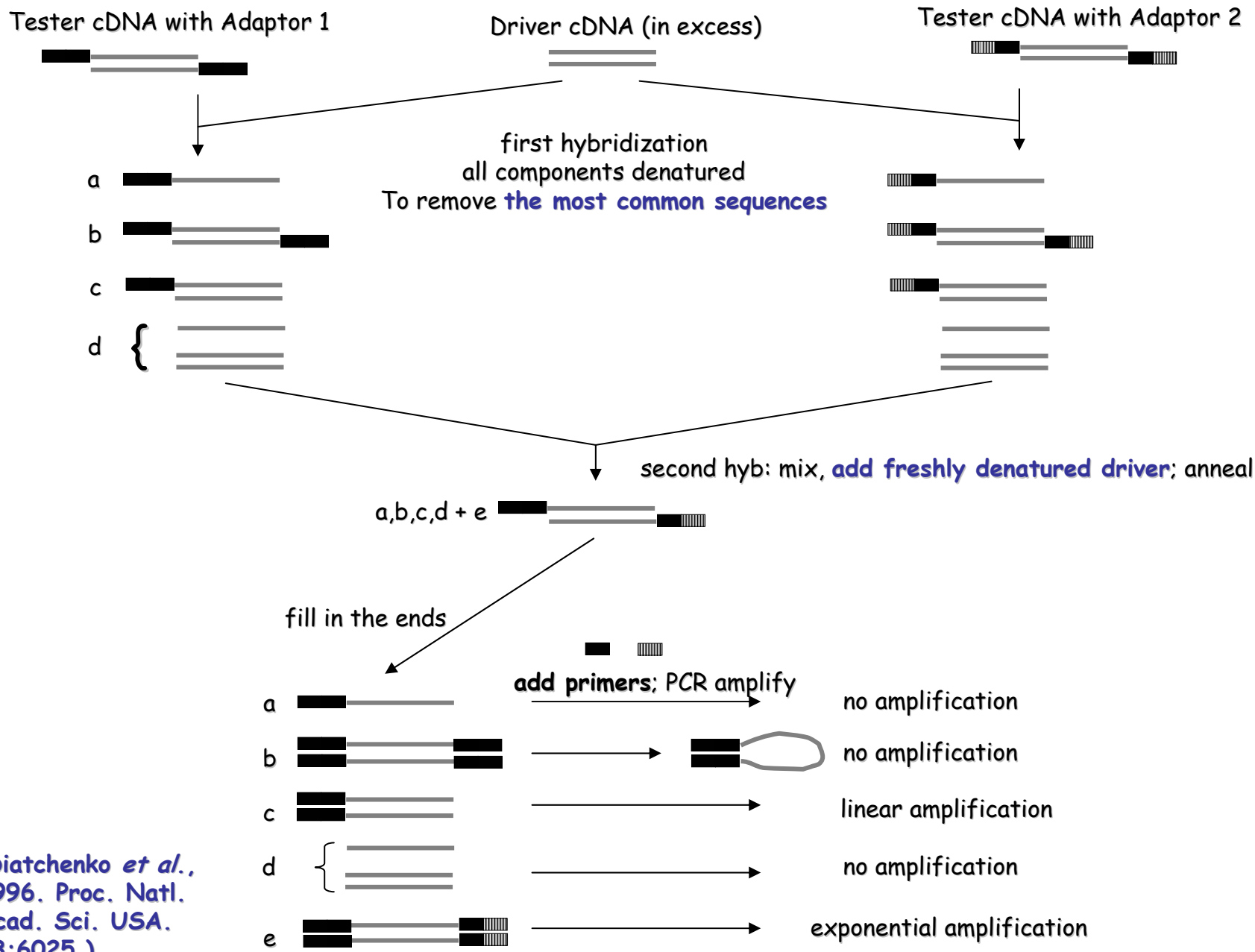
- Dimensionality reduction or principle component analysis
  - Gene shaving & multi-dimensional scaling
- **Probabilistic modeling**
  - Dynamic Bayesian networks
  - Markov Models
- **Statistical significance does not equal biological significance**

Software Tools

- GeneSpring (SiliconGenetics)
- Expressionist (GeneData)
- GeneTraffic (lobion)
- Spotfire (Spotfire)
- Cluster and TreeView (free)
- …

Suppressive Subtractive Hybridization cDNA libraries





(Diatchenko *et al.*, 1996. Proc. Natl. Acad. Sci. USA. 93:6025)

Efficacy of SSH
Ji et al. 2002 BMC Genomics 3:12

- Diatchenko *et al.* 1996 (PNAS **93:6025)**
  - Could detect as little as **0.001% target**
- Critical factor is **relative concentration** of target in **tester** and **driver** populations
- Effective enrichment when
  - Target present at **>= 0.01%**
  - Concentration ratio**>= 5-fold**

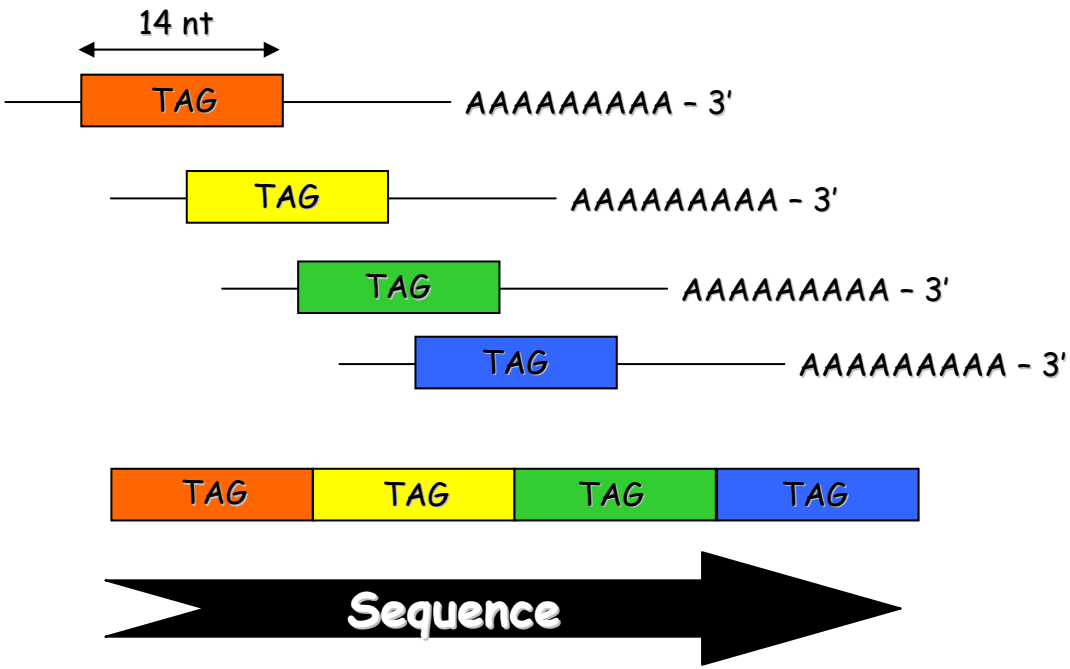
SSH Advantages & Drawbacks

- **Advantages**
  - **Normalization** of transcript levels
  - Detects small (2-fold) differences in transcript levels
  - Identify previously uncharacterized genes (**novel genes**)
  - Generates subtracted libraries **rapidly**

- **Drawbacks**
  - Isolating & sequencing transcripts **slow & laboratories**
  - **Many clones may contain the same sequences**
  - All transcripts must be verified by Northern or **quantitative RT-PCR**

Serial Analysis of Gene Expression (SAGE)

- Velculescu et al. Science 1995
- A transcript (new or novel) can be recognised by a small subset (*e.g.* 14) of its nucleotides – a tag
- Linking tags allows for rapid sequencing
- Open system for transcript profiling

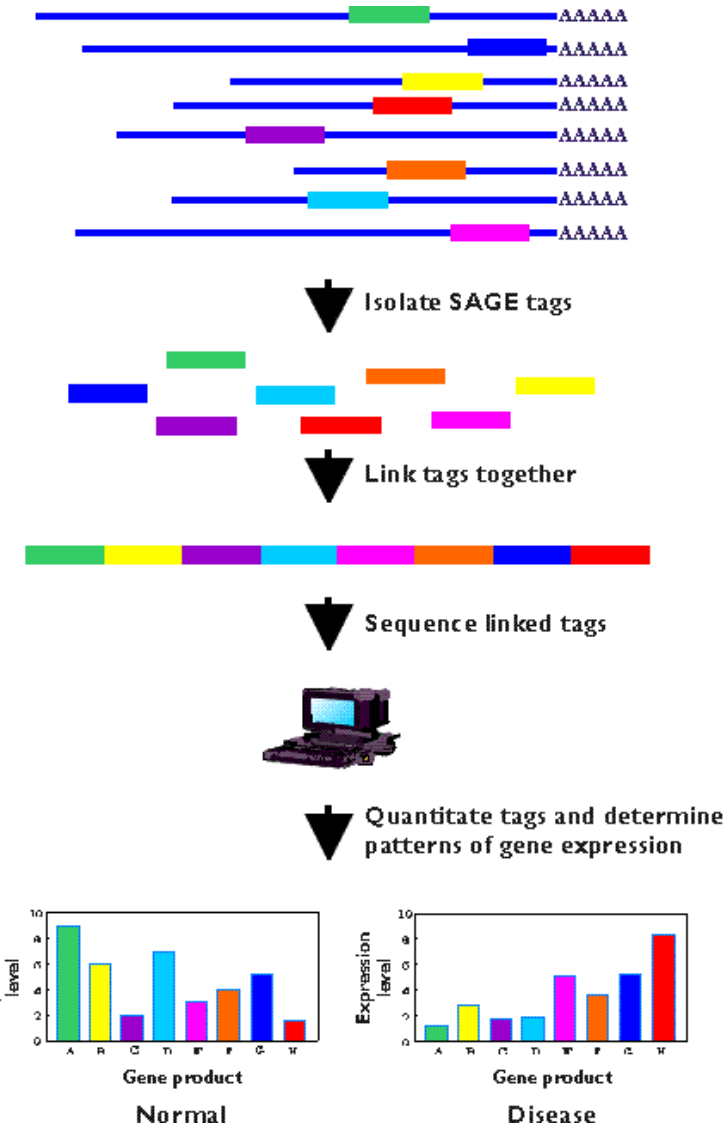


AGCTTGAACCGTGACA
TCATGGCCATTGGCCCC
AATTGAGACAGTGAGT
TCAATGC

Modified SAGE

- LongSAGE (21 nt)
- SAGE-lite, micro-SAGE, mini-SAGE
- RASL/DASL methods (5' and 3' Tags)

1. Trap RNAs with beads
2. Convert the RNA into cDNA
3. Make a cut in each cDNA so that there is a broken end sticking out
4. Attach a "docking module" to this end; here a new enzyme can dock, reach over to cut off a short tag
5. Combine two tags into a unit, a di-tag
6. **Make billions of copies of the di-tags (using a method called PCR)**
7. Remove the modules and glue the di-tags together into long concatamers
8. Put the concatamers into bacteria and copy them millions of times
9. Pick the best concatamers and sequence them
10. Use software to identify different cDNAs there are, and count them;
11. Match the sequence of each tag to the gene that produced the RNA.



Expected Number of Tag Hits

- **Number of unique SAGE tags** grows exponentially with **tag length**

$$N_{tag} = 4^{L}$$

- **Poisson distribution** for probability of finding k hits **at random**

$$P(k \mid \lambda) = \frac{\lambda^{k}}{k!} \exp(-k)$$

- **Unique tag** means no other hits in the genome

$$P_{unique} = P(0 \mid \lambda) = \exp\left(-\frac{L_{genome}}{N_{tag}}\right)$$

SAGE Tag Mapping Statistics

Table 1A. Theoretical matching of tags to genome

Tag length (*n* base pairs)	Complexity[a]	Tag uniqueness probability[b]
14	1,048,576	0.00%
15	4,194,304	0.08%
16	16,777,216	16.73%
17	67,108,864	63.95%
18	268,435,456	89.43%
19	1,073,741,824	97.24%
20	4,294,967,296	99.30%
21	17,179,869,184	99.83%

SAGE

- **Advantages**
  - Potential **"open" system method** - new transcripts can be identified
  - Accuracy of unambiguous transcript observation
  - Digital output of data
  - **Quantitative** & qualitative information

- **Disadvantages**
  - Characterizing **novel transcripts** is often computationally difficult from **short sequences**
  - Tag specificity (recently increased length to **21 nt**)
  - Length of tags can vary (RE enzyme activity variable with temperature)
  - A subset of transcripts **do not contain enzyme recognition sequence**
  - Sensitive of a subset of alternative spice variants