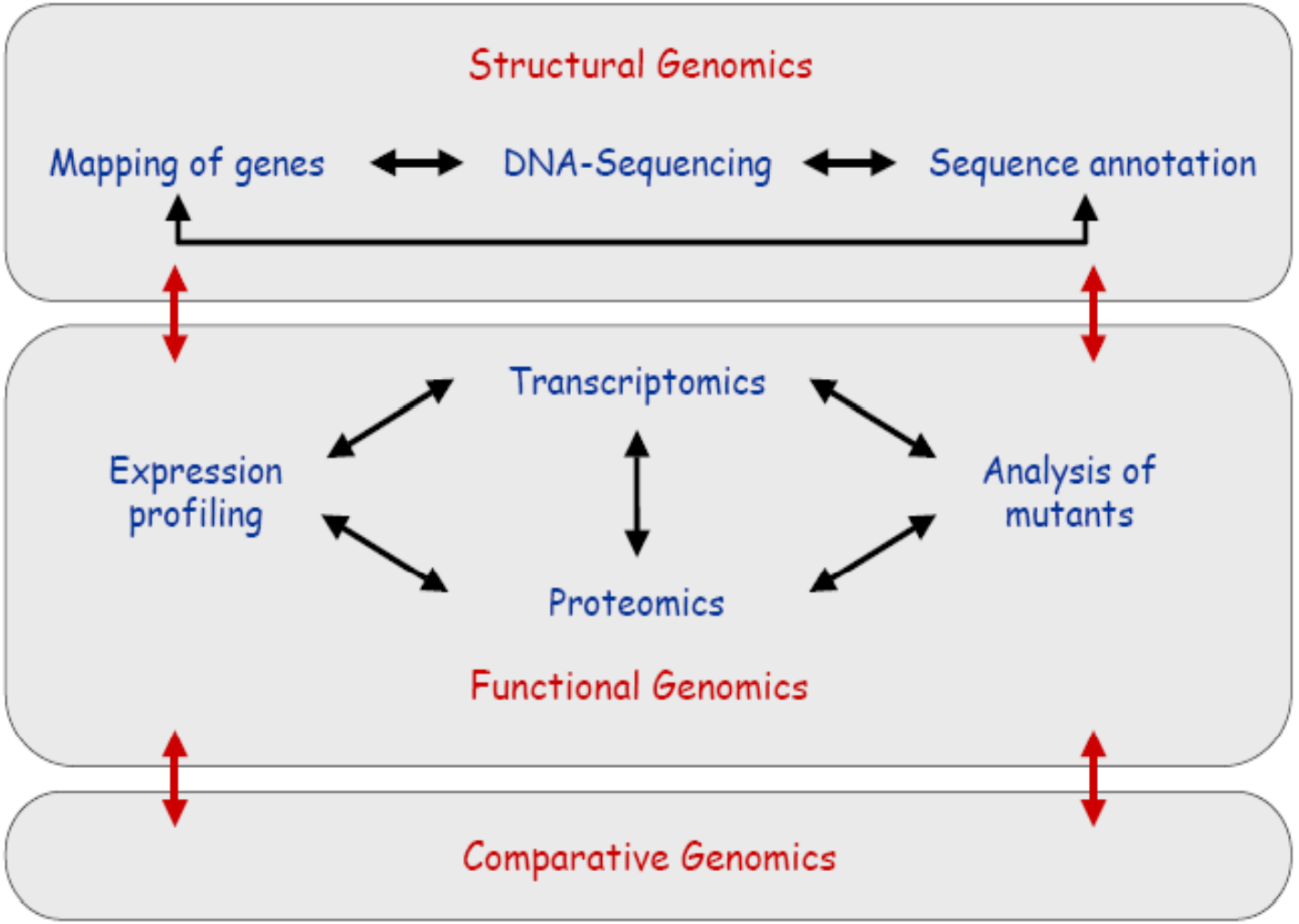
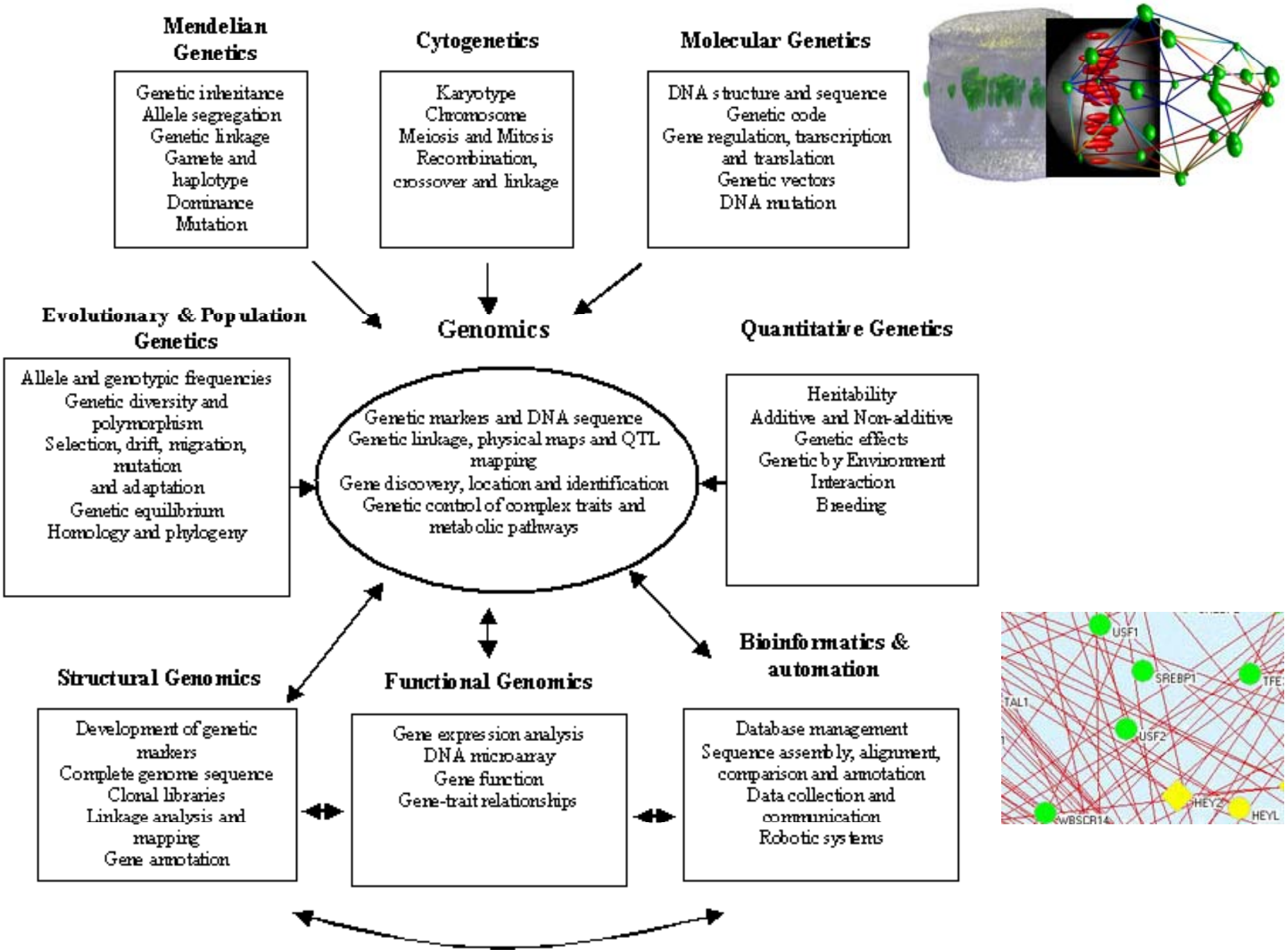
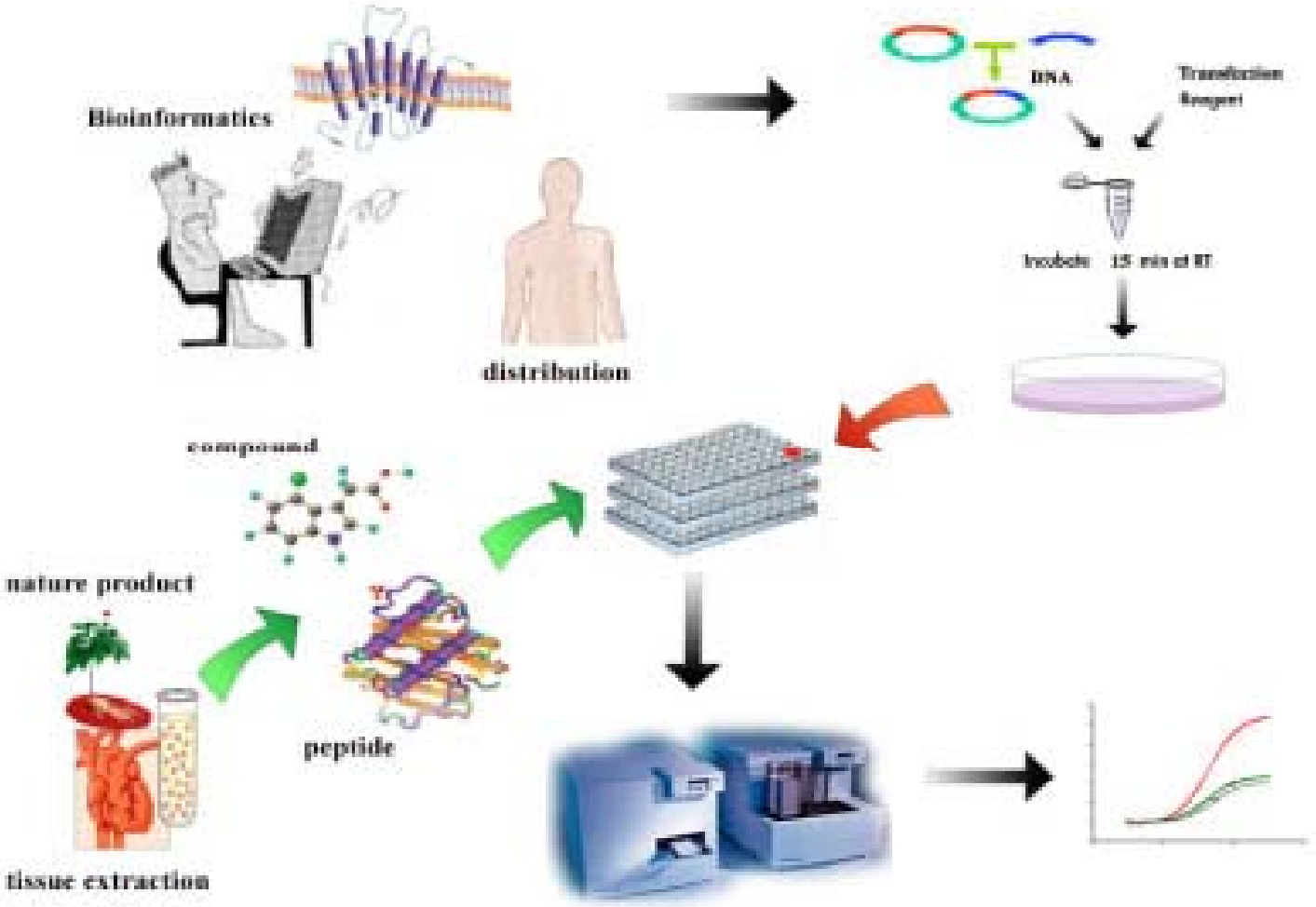
Functional Genomics (1)

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

Three levels of genome research




Mendelian Genetics
Genetic inheritance
Allele segregation
Genetic linkage
Gamete and haplotype
Dominance
Mutation
Cytogenetics
Karyotype
Chromosome
Meiosis and Mitosis
Recombination, crossover and linkage
Molecular Genetics
DNA structure and sequence
Genetic code
Gene regulation, transcription and translation
Genetic vectors
DNA mutation
Evolutionary & Population Genetics
Allele and genotypic frequencies
Genetic diversity and polymorphism
Selection, drift, migration, mutation and adaptation
Genetic equilibrium
Homology and phylogeny
Genomics
Genetic markers and DNA sequence
Genetic linkage, physical maps and QTL mapping
Gene discovery, location and identification
Genetic control of complex traits and metabolic pathways
Quantitative Genetics
Heritability
Additive and Non-additive Genetic effects
Genetic by Environment Interaction
Breeding
Structural Genomics
Development of genetic markers
Complete genome sequence
Clonal libraries
Linkage analysis and mapping
Gene annotation
Functional Genomics
Gene expression analysis
DNA microarray
Gene function
Gene-trait relationships
Bioinformatics & automation
Database management
Sequence assembly, alignment, comparison and annotation
Data collection and communication
Robotic systems
USF1
SREBP1
TFE
TAL1
USF2
HEY2
HEYL
WBSCR14


Bioinformatics
distribution
DNA
Transfection
Reagent
Incubate 15 min at RT
compound
nature product
peptide
tissue extraction

Steps of Genome Analysis

- Genome sequence assembled, markers
  - Identify repetitive sequences – mask out, filter
- Gene location/gene map (mapping)
- Gene prediction – train a model for each genome (including EST & cDNA sequences)
- Genome annotation
- **Functional genomics**
  - http://www.ornl.gov/sci/techresources/Human_Genome/research/function.shtml
- Comparative genomics & Integrative genomics

Functional Genomics Technology Goals

- Generate sets of **full-length cDNA clones** and sequences that represent human **genes** and model organisms
- Support research on **methods** for studying functions of nonprotein-coding sequences
- Develop technology for comprehensive analysis of **gene expression**
- Improve methods for **genome-wide mutagenesis**
- Develop technology for **large-scale protein analyses**
- http://www.ornl.gov/sci/techresources/Human_Genome/research/function.shtml

Definition (1) – Hieter & Boguski 1997

- The development & application of **global**
  - **Genome-wide or**
  - **System-wide experimental** approaches to assess **gene function** by making use of the **information** & **reagents** provided by **structural genomics**

- It is characterized by **high-throughput** or **large-scale** experimental methodologies
  - Combined with **statistical** or **computational analysis** of the results

Definition (2) – UC Davis Genome Center

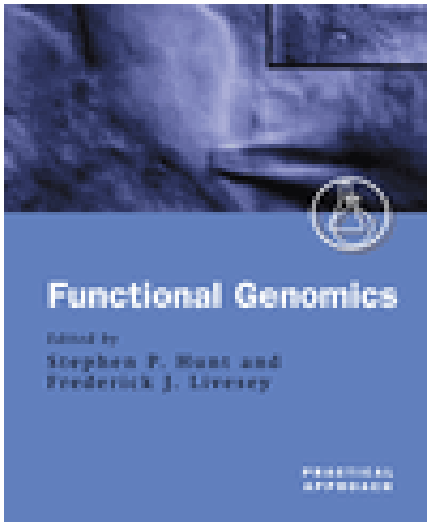
- A means of assessing **phenotype differs** from more **classical approaches** primarily with respect to
  - **The scale & automation** of biological investigations
    - ***A classical investigation*** of gene expression might examine how the expression of ***a single gene*** varies with the development of an organism *in vivo*

- **Modern** functional genomics approaches, however, would examine **1,000-10,000 *genes*** are expressed as a function of development

http://genomics.ucdavis.edu/index_html.html

Definition (3) – Hunt & Livesey (ed.)

- Subtracted cDNA libraries
- Differential display (DD)
- Representational difference analysis
- Suppression subtractive hybridization
- cDNA microarrays
- 2-D gel electrophoresis



http://www.oup.co.uk/isbn/0-19-963774-1

Functional Genomics

- **What** to know
  - Gene **expression**
  - Gene **regulation**
  - Genome-wide **mutagenesis**
- **How** to do
  - Data-mining
  - [SAGE]
  - **Microarray analysis**
  - Subtractive cDNA libraries
  - **Yeast-two hybrids**
  - Transgenics
  - Transposon targeting
  - RNAi & miRNA

Tools for Data Mining

NCBI

PubMed | Entrez | BLAST | OMIM | Books | TaxBrowser | Structure

Search Entrez for Go

Nucleotide Sequence Analysis | Protein Sequence Analysis | Structures | Genome Analysis | Gene Expression

NCBI

Site Map
Guide to NCBI resources

Tools for Programmers

BLAST
Standard tool for sequence analysis

BLink
BLAST Link

CDART
Conserved Domain Architecture Retrieval Tool

Tools - Nucleotide Sequence Analysis

BLAST The **Basic Local Alignment Search Tool** (BLAST) for comparing gene and protein sequences against others in public databases, now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

electronic PCR 00110101011AG Electronic PCR - allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.

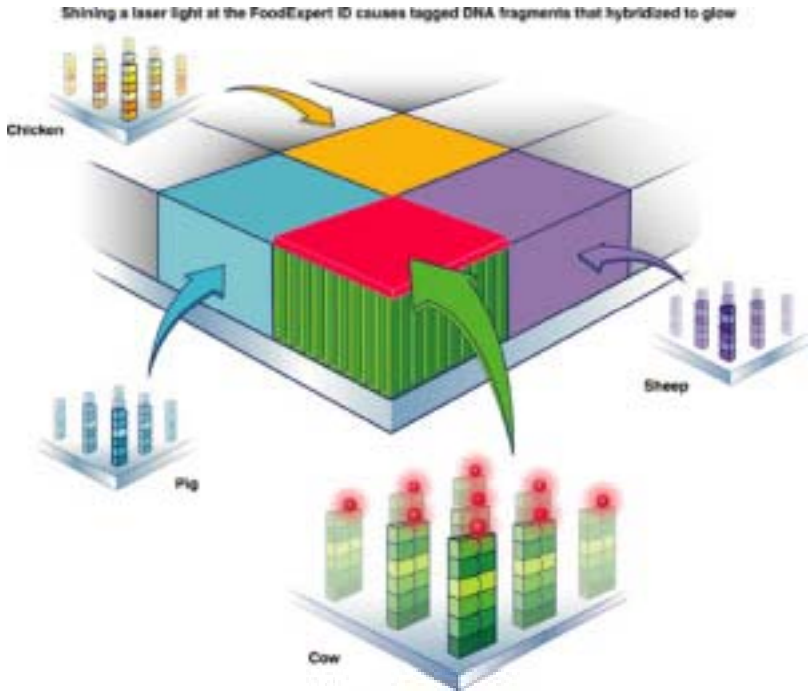
Entrez Gene - each Entrez Gene record encapsulates a wide range of information for a given gene and organism. When possible, the information includes results of analyses that have been done on the sequence data. The amount and type of information presented depend on what is available for a particular gene and organism and can include: (1) graphic summary of the genomic context, intron/exon structure, and flanking genes, (2) link to a graphic view of the mRNA sequence, which in turn shows biological features such as CDS, SNPs, etc., (3) links to gene ontology and phenotypic information, (4) links to corresponding protein sequence data and conserved domains, (5) links to related resources, such as mutation databases. Entrez Gene is a successor to LocusLink.

MM Model Maker - allows you to view the evidence (mRNAs, ESTs, and gene predictions) that was aligned to assembled genomic sequence to build a gene model and to edit the model by selecting or removing putative exons. You can then view the mRNA sequence and potential ORFs for the edited model and save the mRNA sequence data for use in other programs. Model Maker is accessible from sequence maps that were analyzed at NCBI and displayed in Map Viewer.

http://www.ncbi.nlm.nih.gov/Tools/

Expression Arrays - Microarray

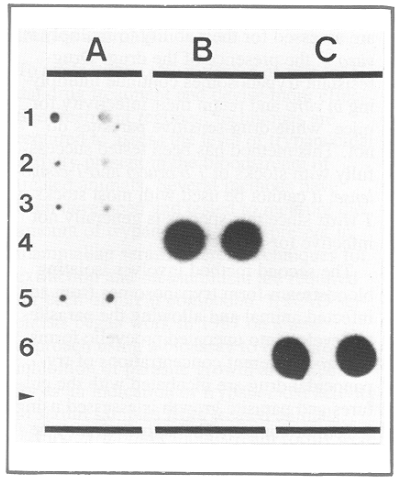
- Cell growth in **different** environments, treatments *etc.*
- Isolate RNA ⇨ cDNAs
- Measure **expression** using array technology
- Create **database** of **expression information**
- **Data Analysis**
  - Display information in an easy-to-use format
  - Show **ratio of expression** under **different conditions**



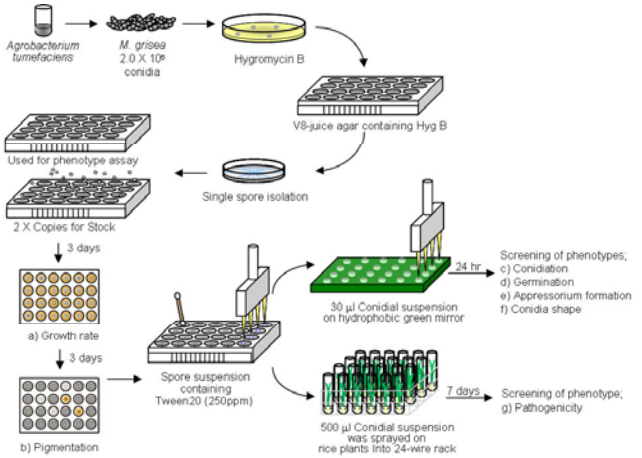
Affymetrix® food chip

Historical Perspective

- **DNA hybridization** (1960s)
  - **Detection** of hybrids
    - Hydroxyapatite $Ca_5(PO_4)_3OH$
    - **Radioactive** labeling
    - Enzyme-linked detection
    - **Fluorescent** labeling
  - Fixing sample on **solid support**
    - Southern blots (1970s)
    - Northern blots
    - Dot blots



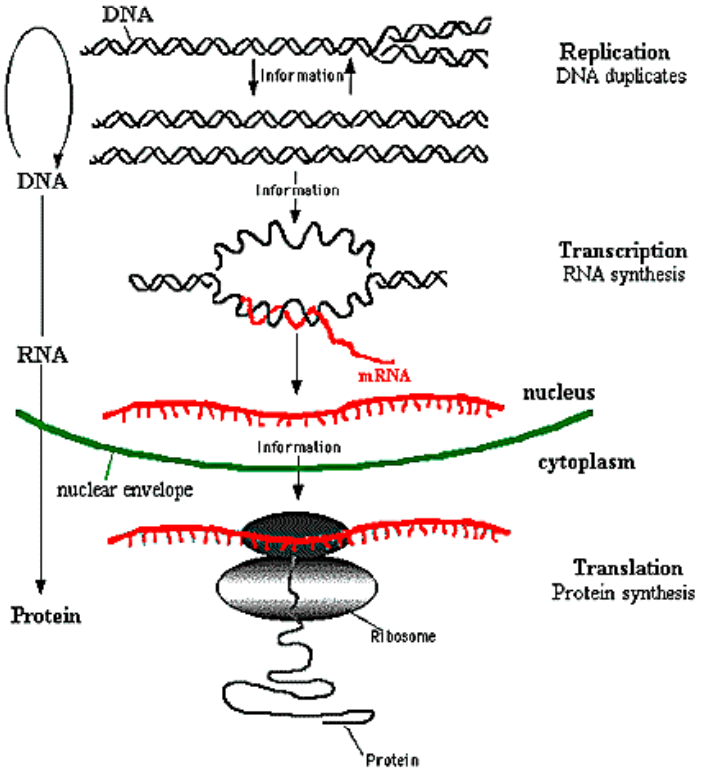
Basic Principles



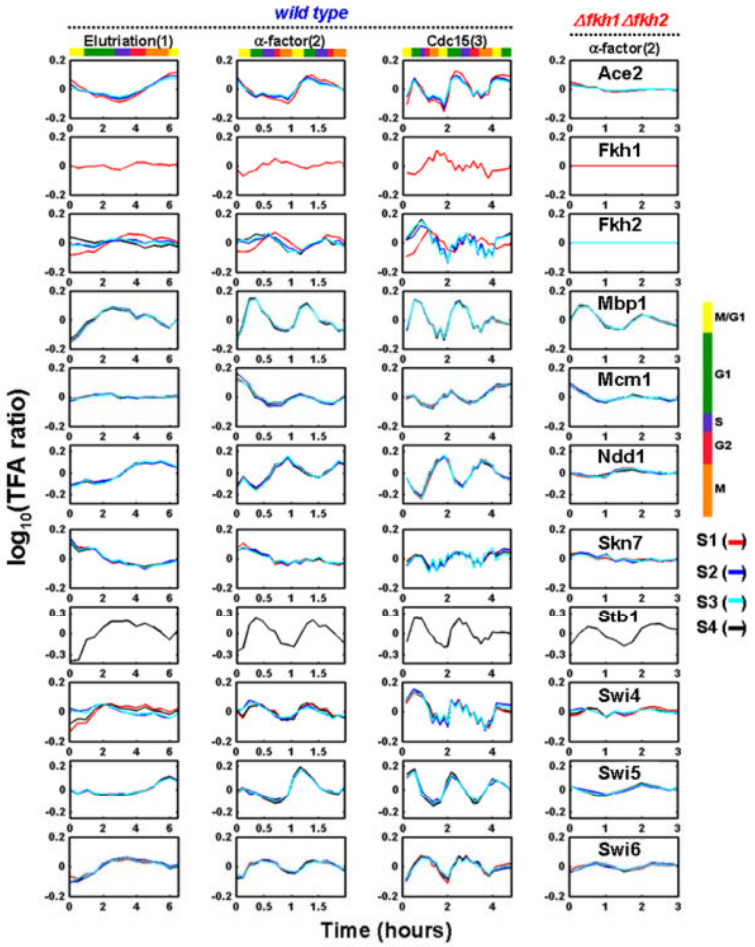
- Main novelty is one of **scale**
  - **Hundreds or thousands** of probes rather than tens
- Probes are attached to **solid supports**
- **Robotics** are used extensively
- **Informatics** is **a central component** at all stages

Gene Expression Analysis (Whole Genome)

- Quantitative Analysis of Gene Activities
  - **Transcription Profiles**

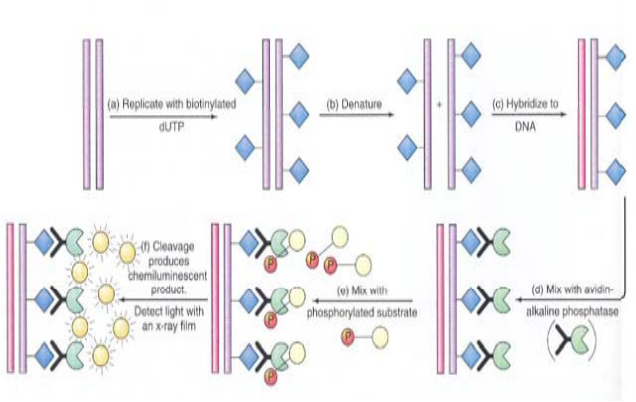


The Central Dogma of Molecular Biology



Yang *et al. BMC Genomics* 2005
6:90 doi:10.1186/1471-2164-6-90

Major Technologies



- **cDNA probes (> 200 nt),** usually produced by **PCR**, attached to either **nylon** or **glass** supports
- **Oligonucleotides (25-80 nt)** attached to glass support
- **Oligonucleotides** (25-30 nt) synthesized ***in situ*** on silica wafers (***Affymetrix***)
- Probes attached to **tagged beads**

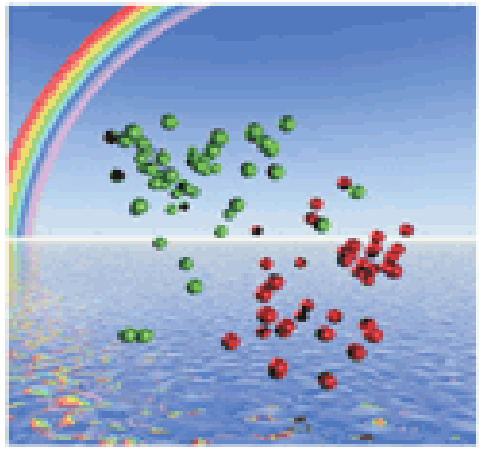
Principal Uses of Chips (1)

- Genome-scale **gene expression** analysis
  - Differentiation
  - Responses to environmental factors
  - Disease processes
  - Effects of drugs

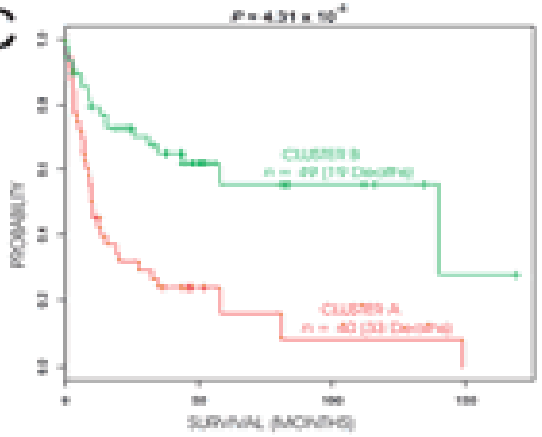
4187 genes; 91 samples

A

B

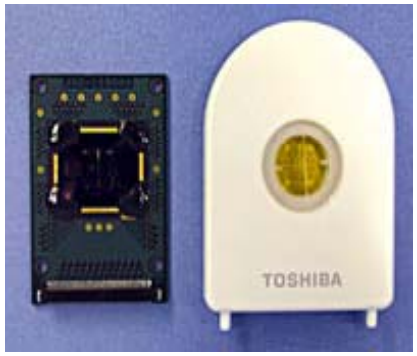


C



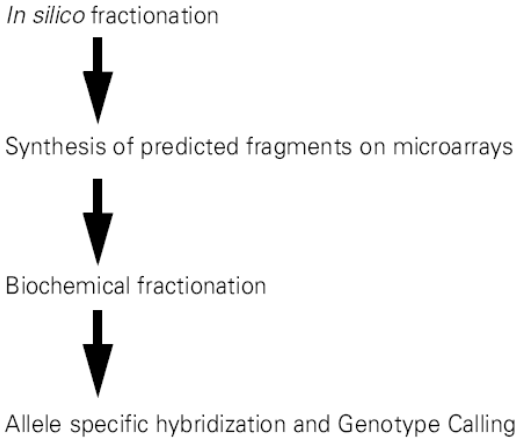
- Genome-scale profiling of gene expression in hepatocellular carcinoma: classification and survival prediction
  - CCR Frontiers in Science (2006); Lee et al. Hepatology 40:667-76 (2004)

Principal Uses of Chips (2)



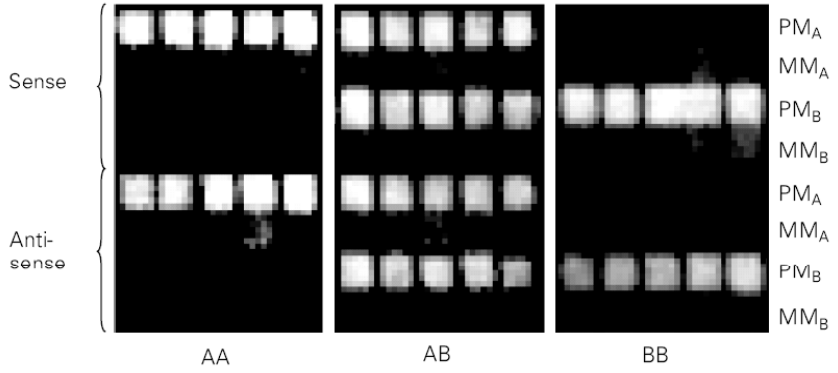
Toshiba's hepatitis C SNP typing chip

- Detection of **sequence variation**
  - **Genotyping**
  - Detection of somatic mutations (*e.g.* in oncogenes)
  - Direct sequencing

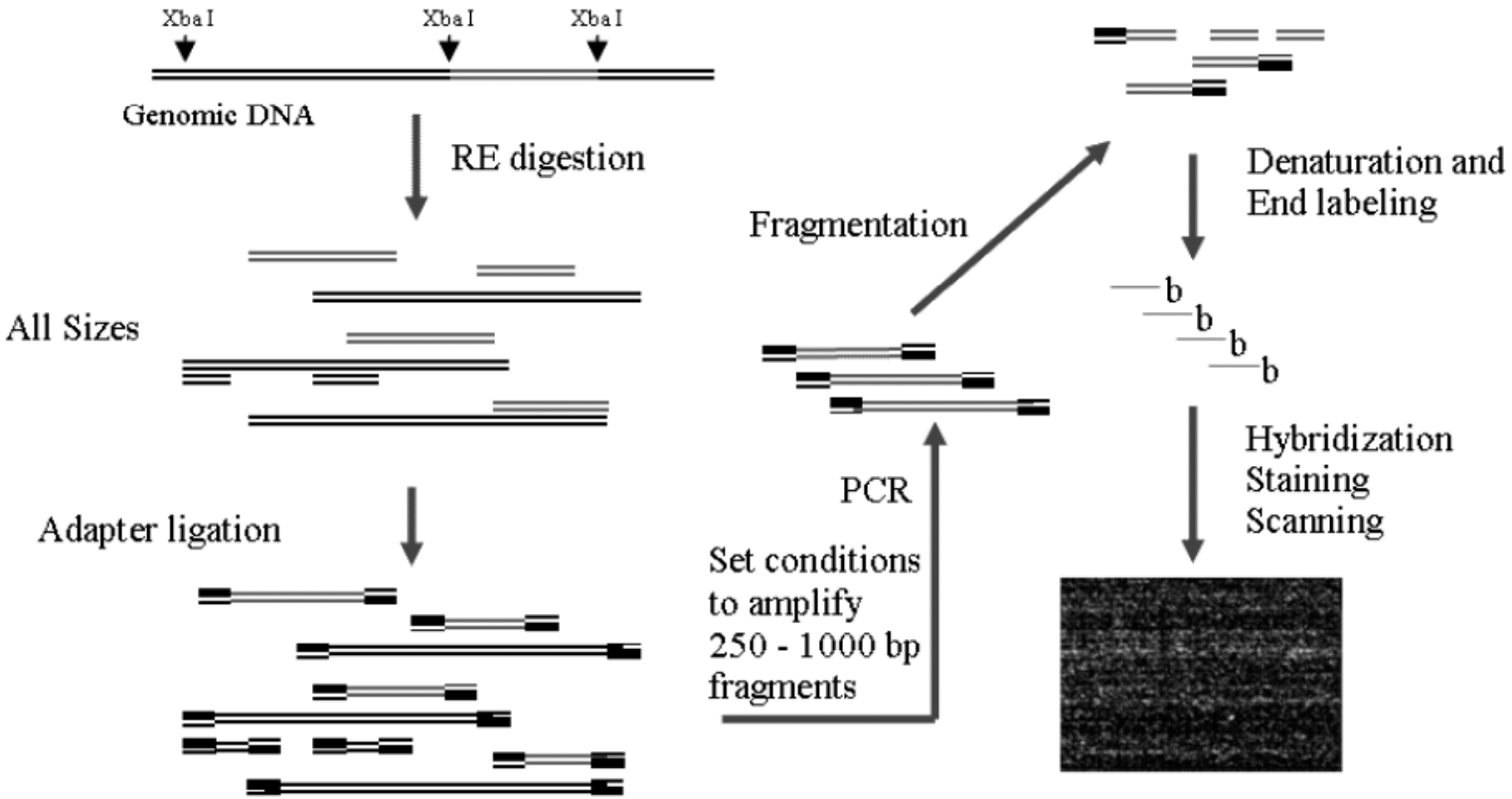


Allele-specific hybridization (ASH)

Chee et al. 1996; Wang et al. 1998; Lindblad-Toh et al. 2000; **40 different, 25-bp oligos**

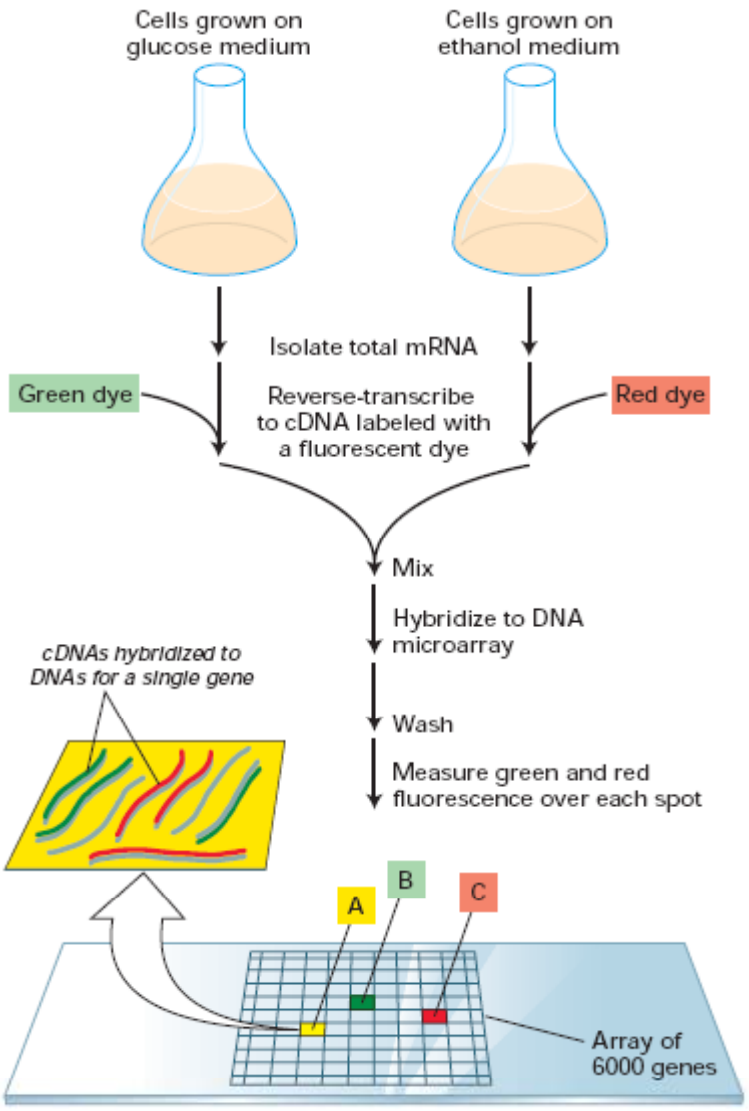


SNP Strategy - "GeneChip Mapping Assay



cDNA Chips

- Probes are **cDNA fragments**, usually **amplified by PCR**
- **Probes** are deposited on a solid support, either **positively charged nylon** or **glass** slide
- Samples (normally poly*A*+ RNA) are labeled using **fluorescent dyes**
- At least **two samples** are hybridized to chip
- **Fluorescence** at different **wavelengths** measured by a scanner



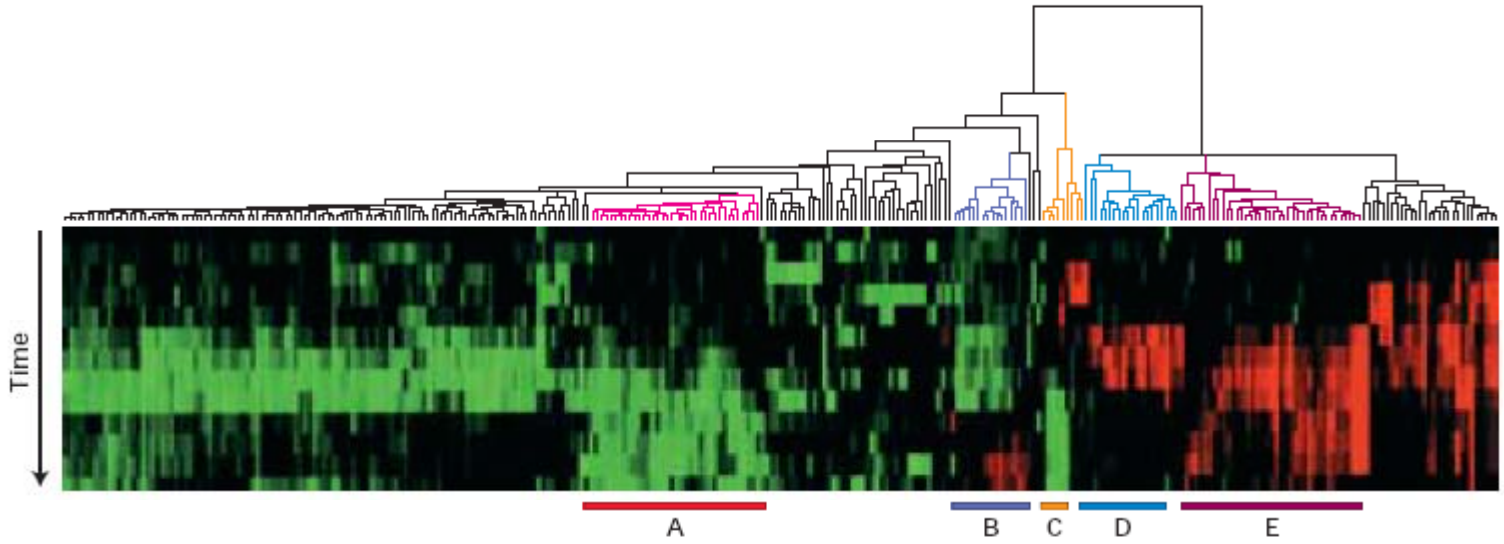
A If a spot is yellow, expression of that gene is the same in cells grown either on glucose or ethanol

B If a spot is green, expression of that gene is greater in cells grown in glucose

C If a spot is red, expression of that gene is greater in cells grown in ethanol

▲ **EXPERIMENTAL FIGURE 9-35 DNA microarray analysis can reveal differences in gene expression in yeast cells under different experimental conditions.** In this example, cDNA prepared from mRNA isolated from wild-type *Saccharomyces* cells grown on glucose or ethanol is labeled with different fluorescent dyes. A microarray composed of DNA spots representing each yeast gene is exposed to an equal mixture of the two cDNA preparations under hybridization conditions. The ratio of the intensities of red and green fluorescence over each spot, detected with a scanning confocal laser microscope, indicates the relative expression of each gene in cells grown on each of the carbon sources. Microarray analysis also is useful for detecting differences in gene expression between wild-type and mutant strains.

Molecular Cell Biology,
Lodish 5th Ed.



▲ EXPERIMENTAL FIGURE 9-36 Cluster analysis of data from multiple microarray expression experiments can identify co-regulated genes. In this experiment, the expression of 8600 mammalian genes was detected by microarray analysis at time intervals over a 24-hour period after starved fibroblasts were provided with serum. The cluster diagram shown here is based on a computer algorithm that groups genes showing similar changes in expression compared with a starved control sample over time. Each column of colored boxes represents a single gene, and each row represents a time point. A red box indicates an increase in expression relative to the control; a green box, a decrease in expression; and a black box, no significant change in expression. The "tree" diagram at the top shows how the expression patterns for individual genes can be organized in a hierarchical fashion to group together the genes with the greatest similarity in their patterns of expression over time. Five clusters of coordinately regulated genes were identified in this experiment, as indicated by the bars at the bottom. Each cluster contains multiple genes whose encoded proteins function in a particular cellular process: cholesterol biosynthesis (A), the cell cycle (B), the immediate-early response (C), signaling and angiogenesis (D), and wound healing and tissue remodeling (E). [Courtesy of Michael B. Eisen, Lawrence Berkeley National Laboratory.]

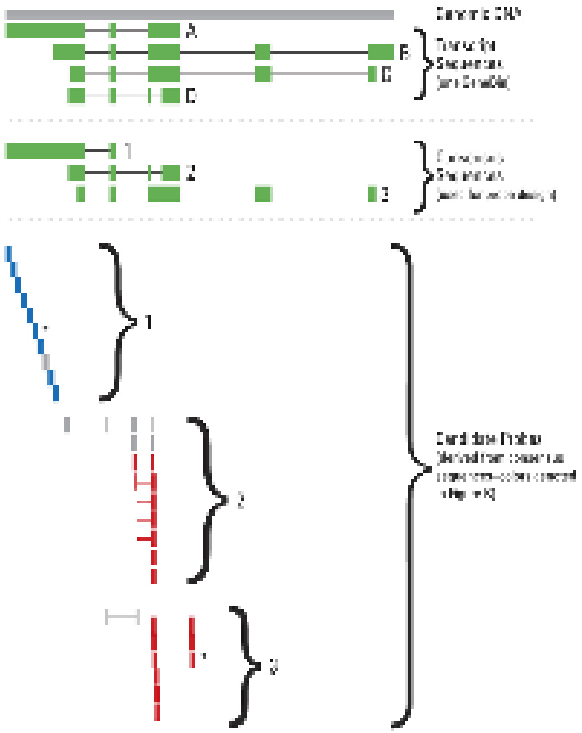
cDNA Chip Design

- **Probe selection**
  - **Non-redundant** set of probes
  - Includes genes of **interest** to project
  - Corresponds to physically **available** clones

- **Chip layout**
  - Grouping of probes by **function**
  - Correspondence between wells in **microtiter plates** and spots on the chip

Probe Selection

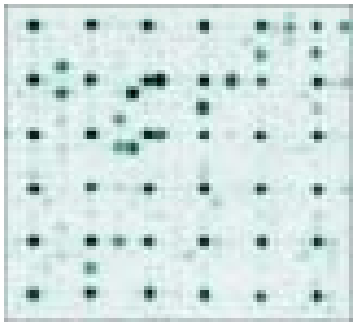
- Make sure that database entries are **cDNA**
  - Preference for **RefSeq** entries
    - Criteria for **non-redundancy**
    - **>98% identity over >100 nt**
    - Accession number is unique
- Mapping of sequence to **clone**
  - Use **Unigene clusters**
  - Directly use data from **sequence verified collection** (*e.g.* Research Genetics)
  - **Independently verify sequence**



Agilent Technology: 60-mer probe selection; GeneBin

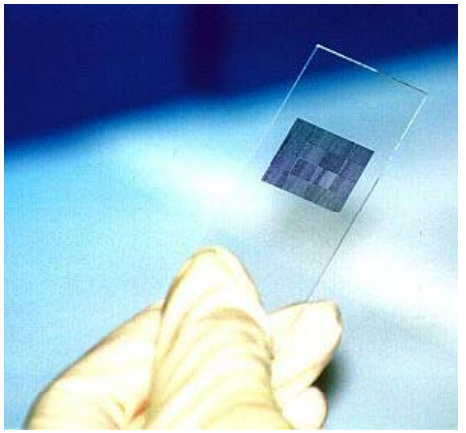
cDNA Arrays on Nylon and Glass

- **Nylon arrays**
  - Up to about **1,000 probes** per filter
  - Use radiolabeled cDNA target
  - Can use **phosphorimager** or X-ray film

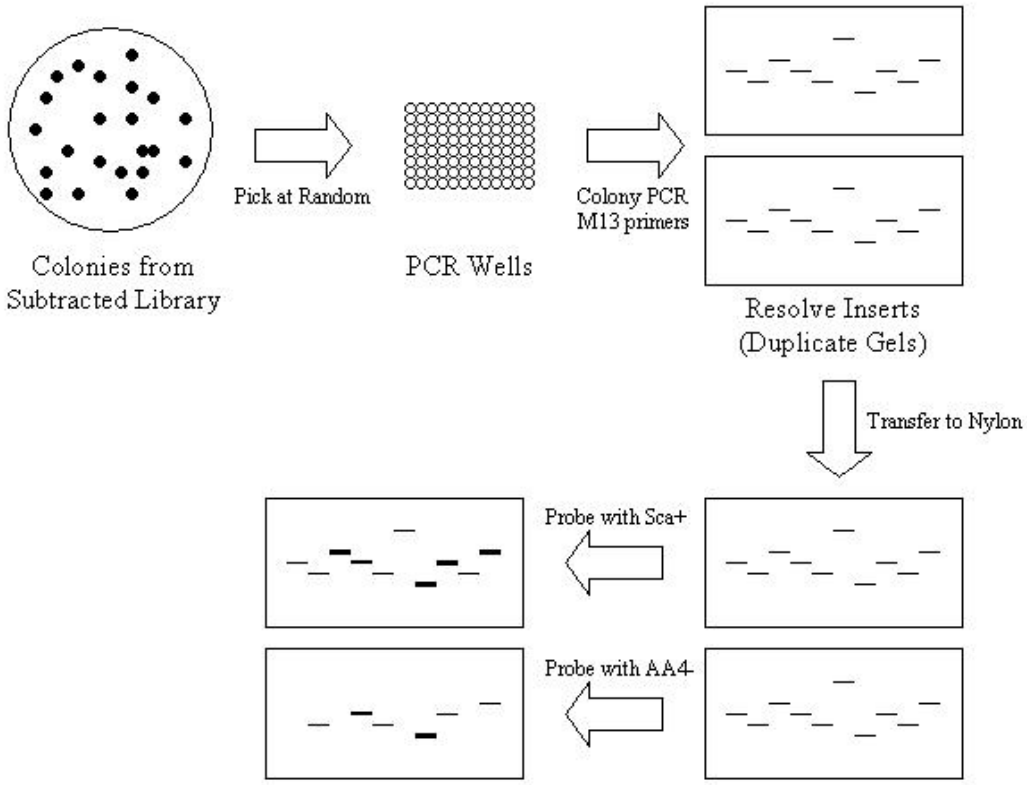


RZPD Nylon array

- **Glass arrays**
  - Up to about **40,000** probes per slide, or 10,000 per $2cm^2$ area (limited by arrayer's capabilities)
  - Use **fluorescent targets**
  - Require specialized scanner



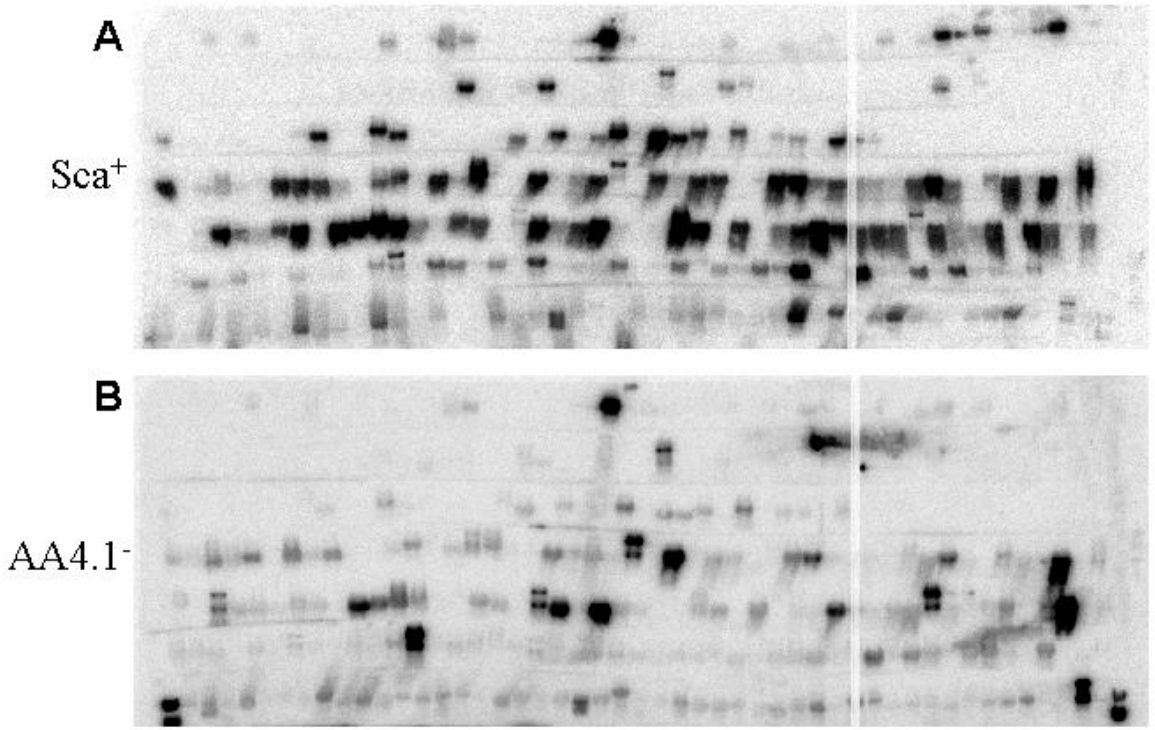
Overview of the Production of a Pair of Cheap, Low-density Nylon Arrays of PCR Products



StemCellDB:
library ID

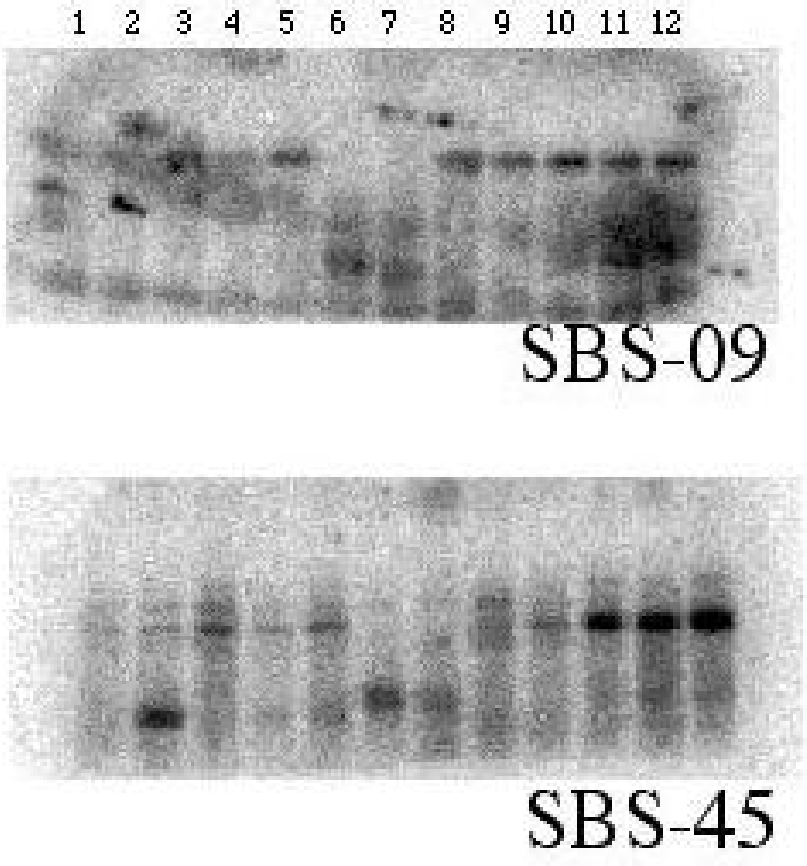
http://stemcell.princeton.edu/v1/sbs_screen.html

Actual image of two duplicate arrays of 332 clones each, probed with Sca+ (-) AA4- (top) or AA4- (-) Sca+ (bottom) subtracted probe populations



http://stemcell.princeton.edu/v1/sbs_screen.html

Northern Blotting Confirmation

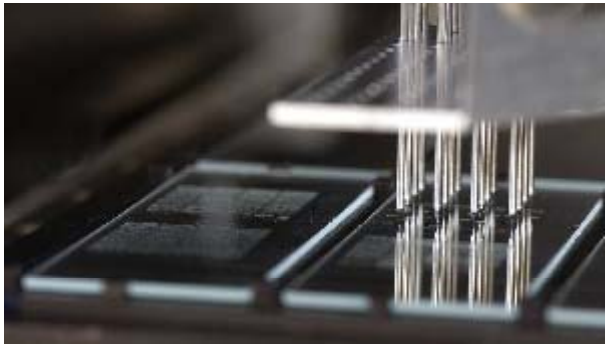


http://stemcell.princeton.edu/v1/sbs_screen.html

Array Type & Spot Density

Array Type	*Spot Density (per cm^2)*	*Probe*	*Target*	*Labeling*
Nylon Macroarrays	< 100	cDNA	RNA	Radioactive
Nylon Microarrays	< 5000	cDNA	mRNA	Radioactive/Flourescent
Glass Microarrays	< 10,000	cDNA	mRNA	Flourescent
Oligonucleotide Chips	<250,000	oligo's	mRNA	Flourescent

Glass Chip Manufacturing

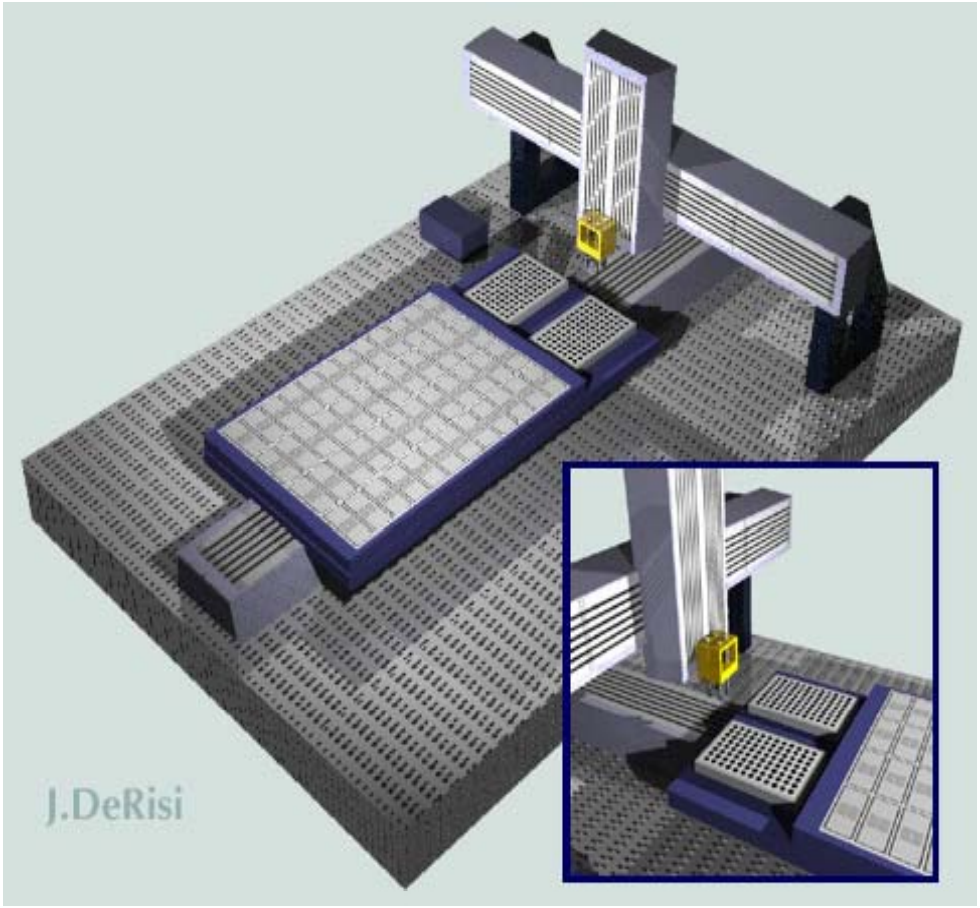


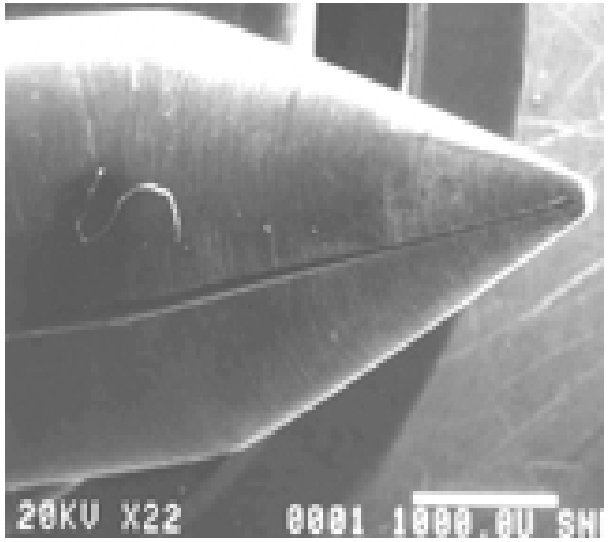
- Choice of **coupling** method
  - **Physical** (charge), non-specific chemical, specific chemical (modified PCR primer)
- Choice of **printing** method
  - Mechanical pins: flat tip, split tip, pin & ring
  - Piezoelectric (壓電的) deposition ("**ink-jet**")
- **Robot design**
  - Precision of movement in 3 axes
  - Speed and **throughput**
  - Number of pins, numbers of spots per pin load

CHIP 1000,

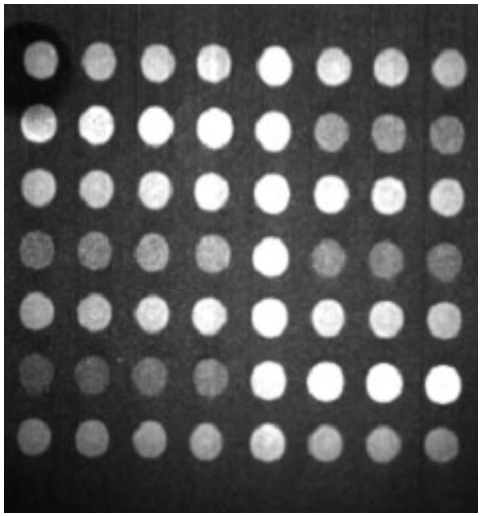
Shimadzu Biotech

Physical Spotting





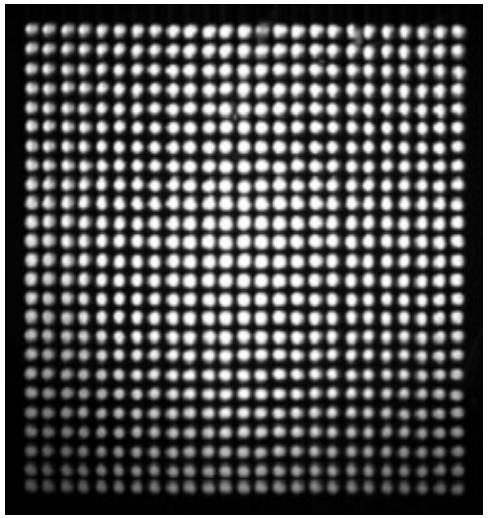
Typical Ink Jet Spot Deposition Results



Volume per spot: 250 nl

Spot size: 1,100 μm

Spot density: 70/cm^2



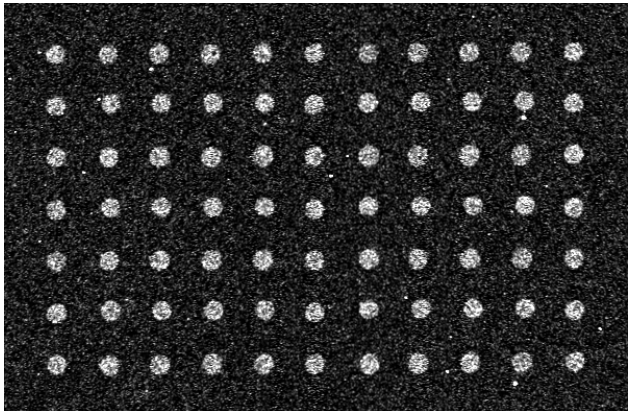
Volume per spot: 0.5 nl

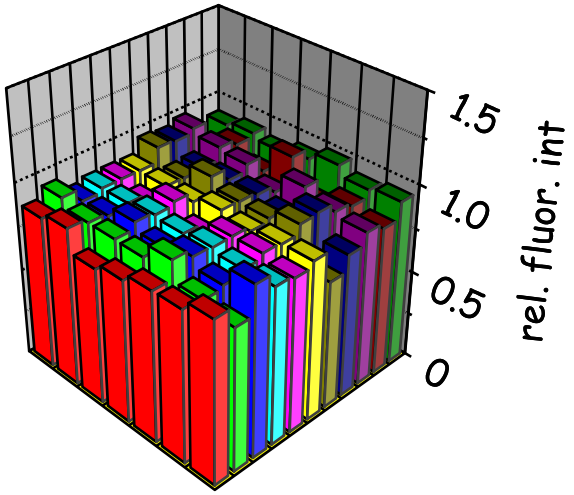
Spot size: 115 μm

Spot density: 4,800/cm^2

Labelled BSA (Cy5)

Typical Pin Spot Deposition Microarray Results

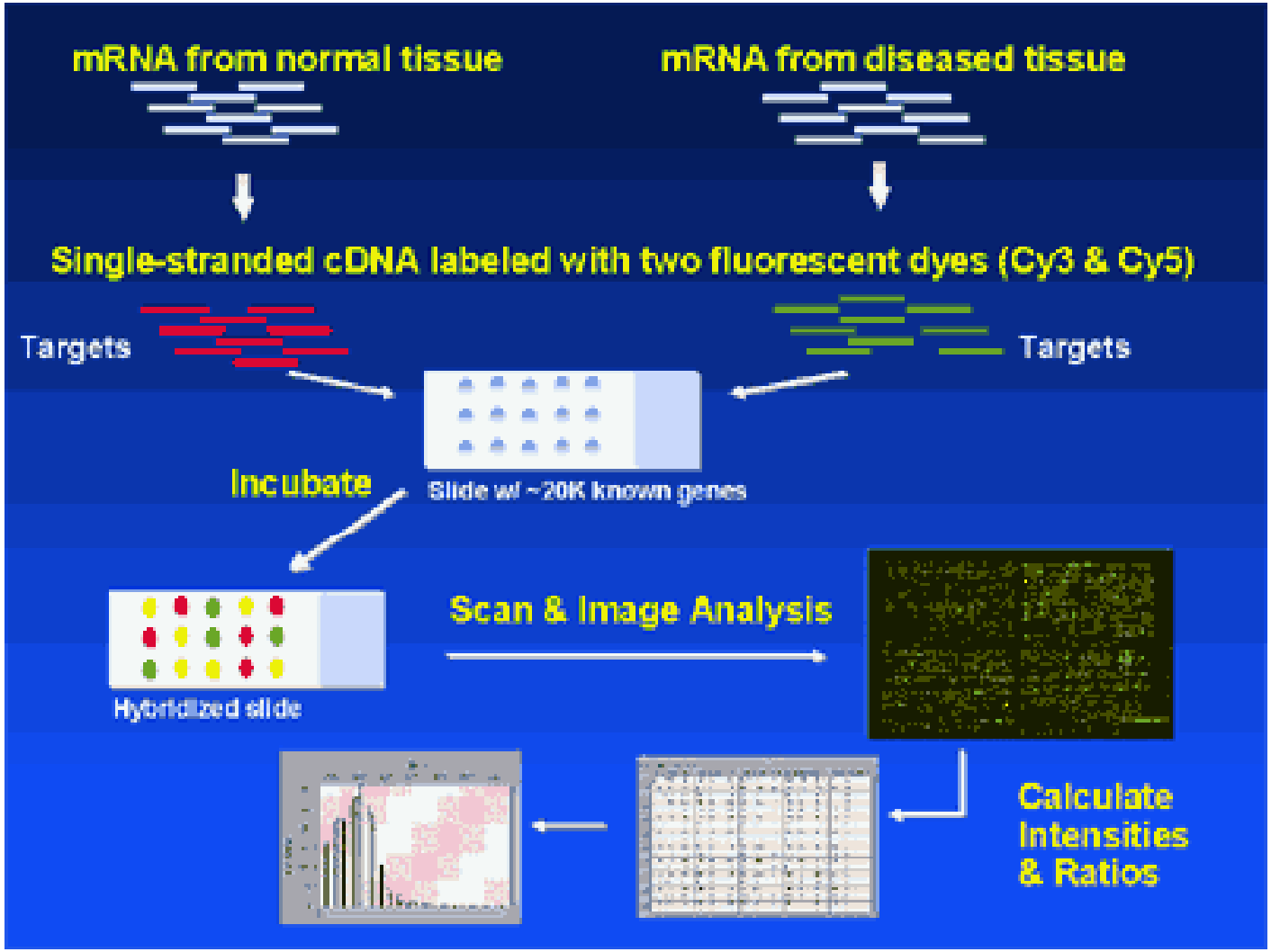




7x11 microarray consisting of
identical Cy5-BSA spots (pitch 500 mm)

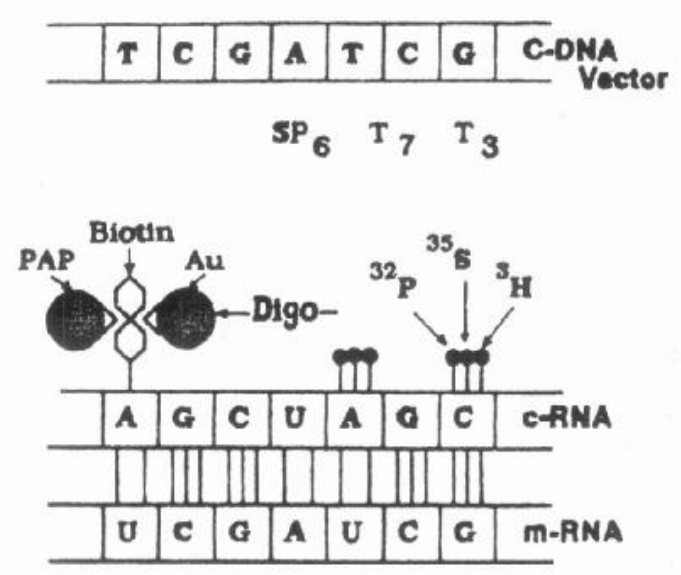
Typical CV: ≤ 5%

Protocol


mRNA from normal tissue
mRNA from diseased tissue
Single-stranded cDNA labeled with two fluorescent dyes (Cy3 & Cy5)
Targets
Targets
Incubate
Slide w/ ~20K known genes
Hybridized slide
Scan & Image Analysis
Calculate
Intensities
& Ratios

Labeling and Hybridization

- **Targets** are normally prepared by **oligo(dT)** primed cDNA synthesis
  - Probes should contain 3' end of mRNA
  - Need **CoT1 DNA** as **competitor** (*esp.* LINE)
- **Alternative protocol** is to make **ds cDNA** containing bacterial promoter, then **cRNA**
  - Can work with **smaller amount of RNA**
    - Less quantitative
- Hybridization usually **under coverslips**

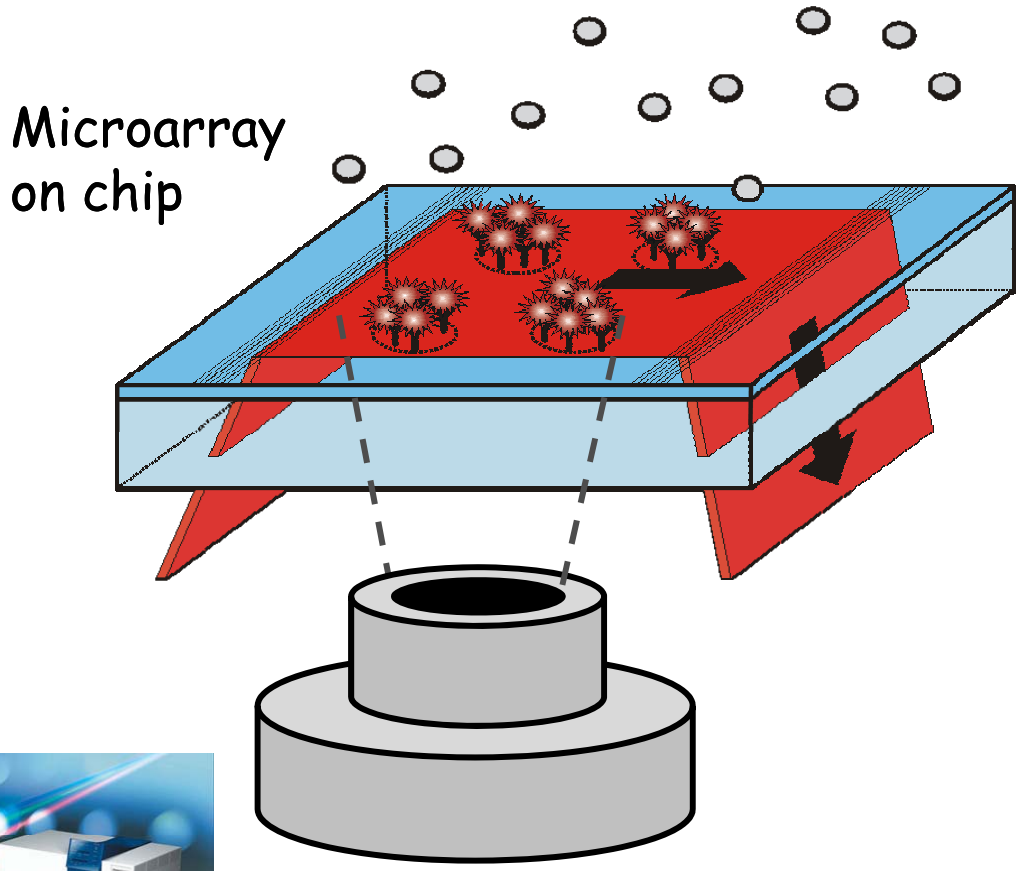


Scanning the Arrays

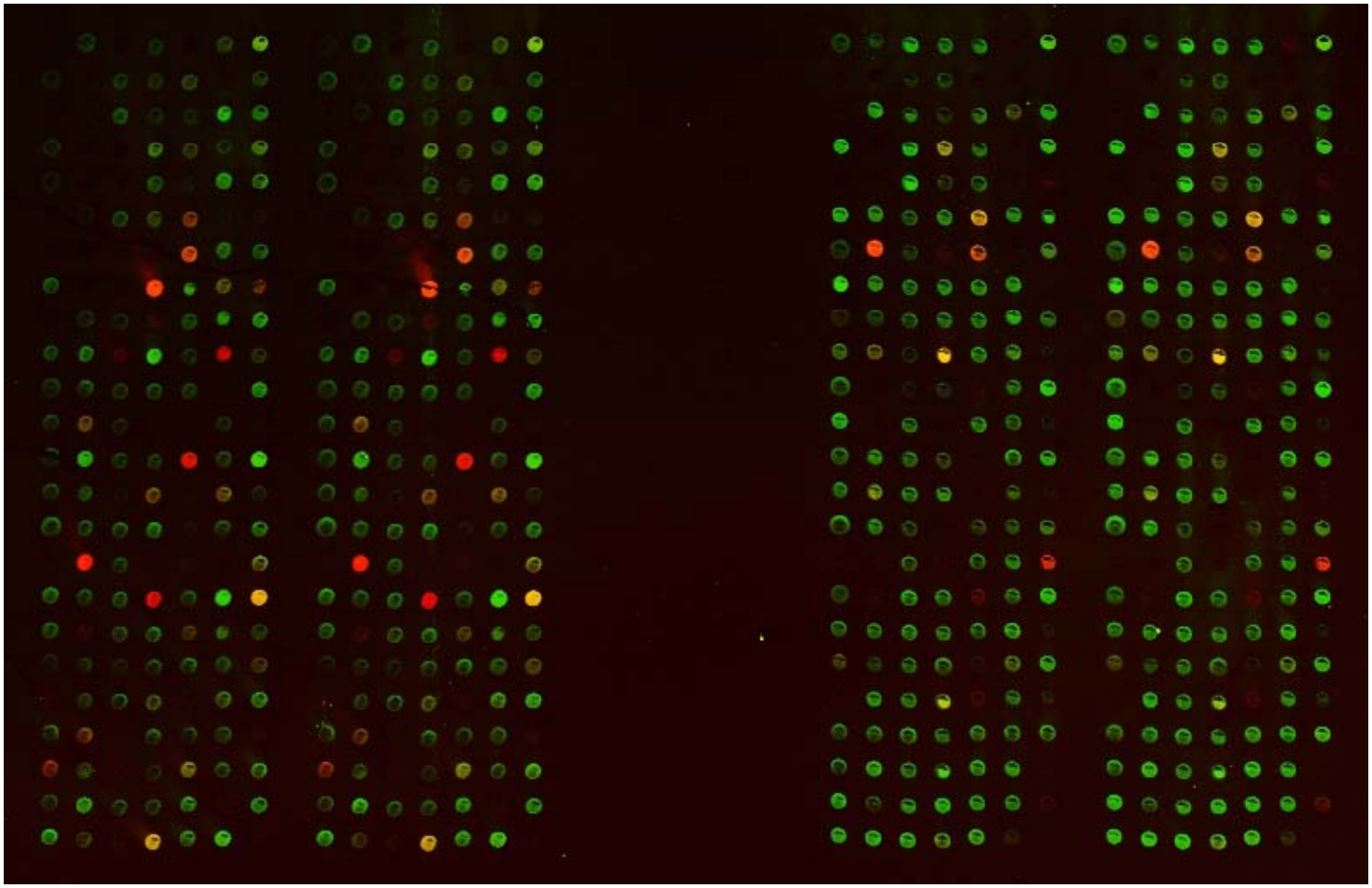
- **Laser scanners**
  - Excellent spatial resolution
  - **Good sensitivity**, but can **bleach** fluorochromes
  - Still rather **slow**
- **CCD (Charged-Coupled Device) scanners**
  - Spatial resolution can be a problem
  - **Sensitivity** easily adjustable **(exposure time)**
  - **Faster** and cheaper than lasers
- In all cases, raw data are **images** showing **fluorescence** on surface of chip

BioRad: VersArray ChipReader ™ laser confocal scanners

Example: Zeptosens Planar Waveguide Principle – for High Sensitivity Fluorescence Microarray Detection



Glass Microarray –
326 Rat Heart Genes, 2X spotting



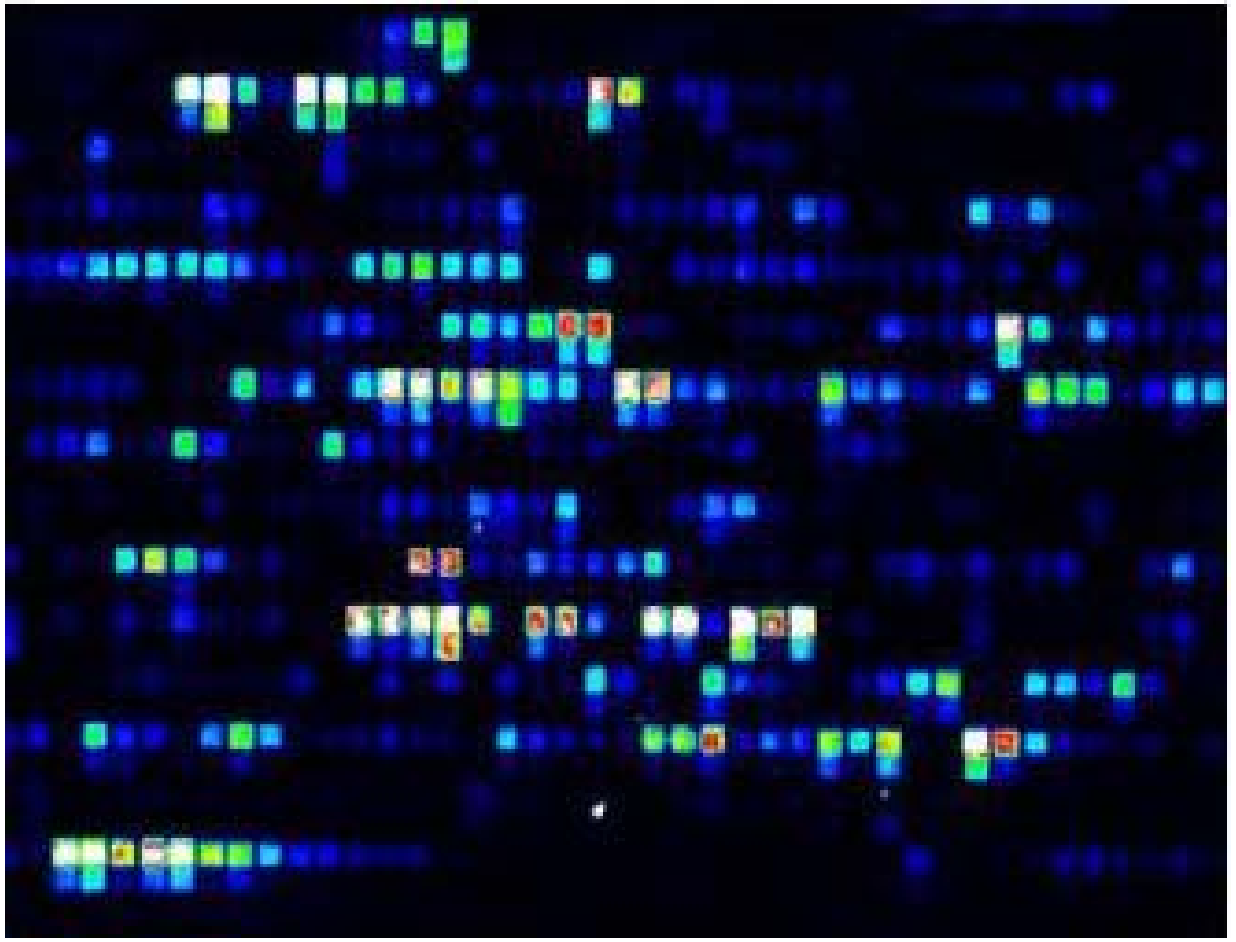
Coffee Break

- What did a Math book says to the other?
- I have a lot of problems!

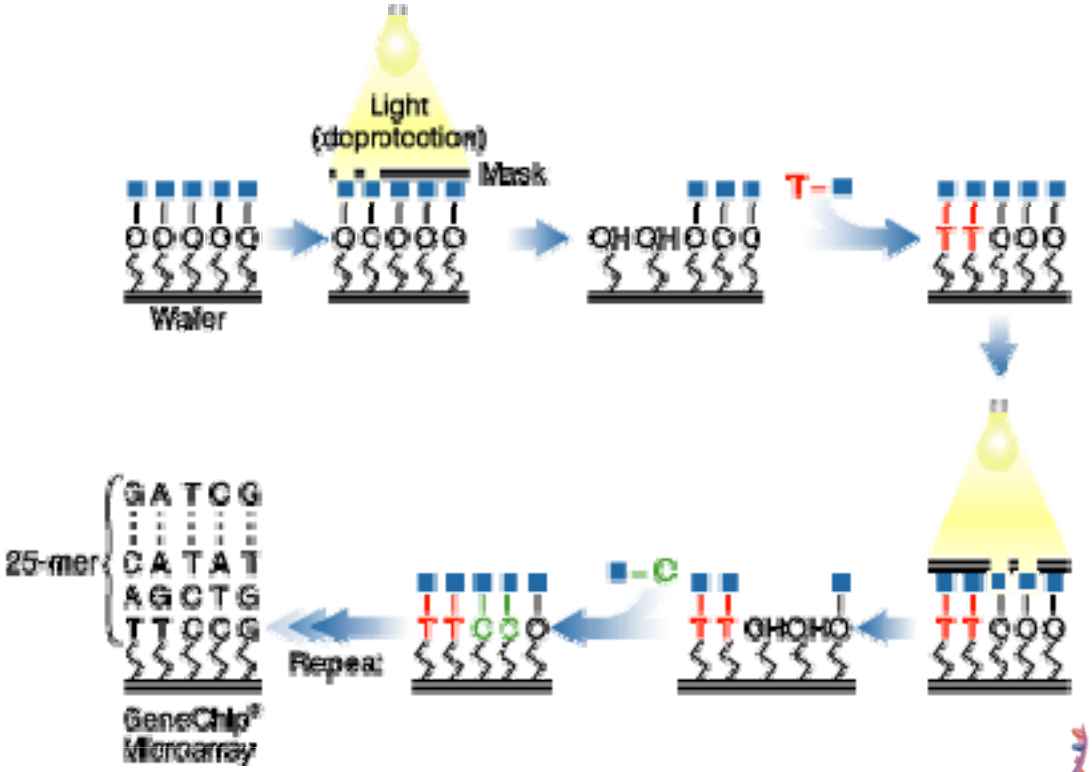
The Affymetrix Approach

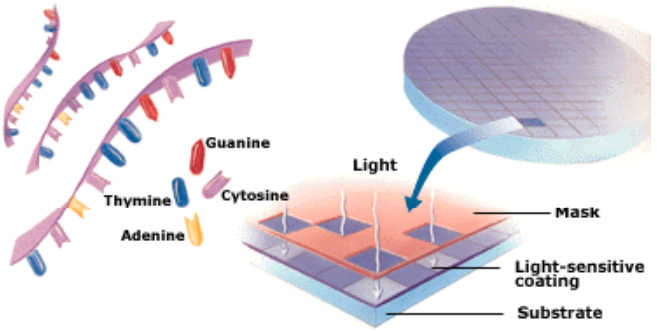
- Probes are **oligos** synthesized *in situ* using a **photolithographic** approach
- There are at least **13-16 oligos per gene (PM)**, plus **an equal number** of **negative controls (MM)**
- The apparatus requires **a fluidics station** for hybridization and a **special scanner**
- Only **a single fluorochrome** is used per hybridization
- It is **very** expensive !

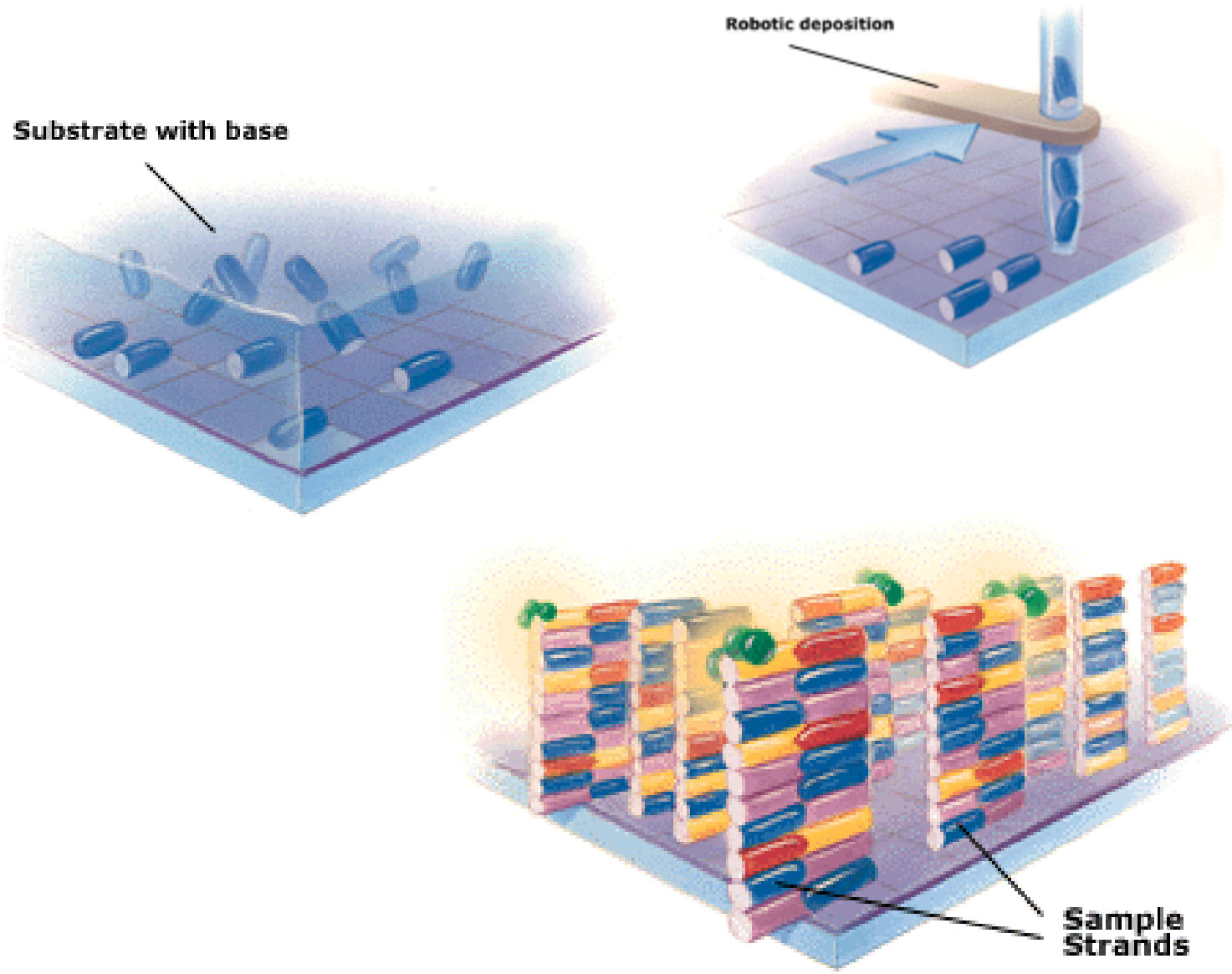
Affymetrix GeneChip®

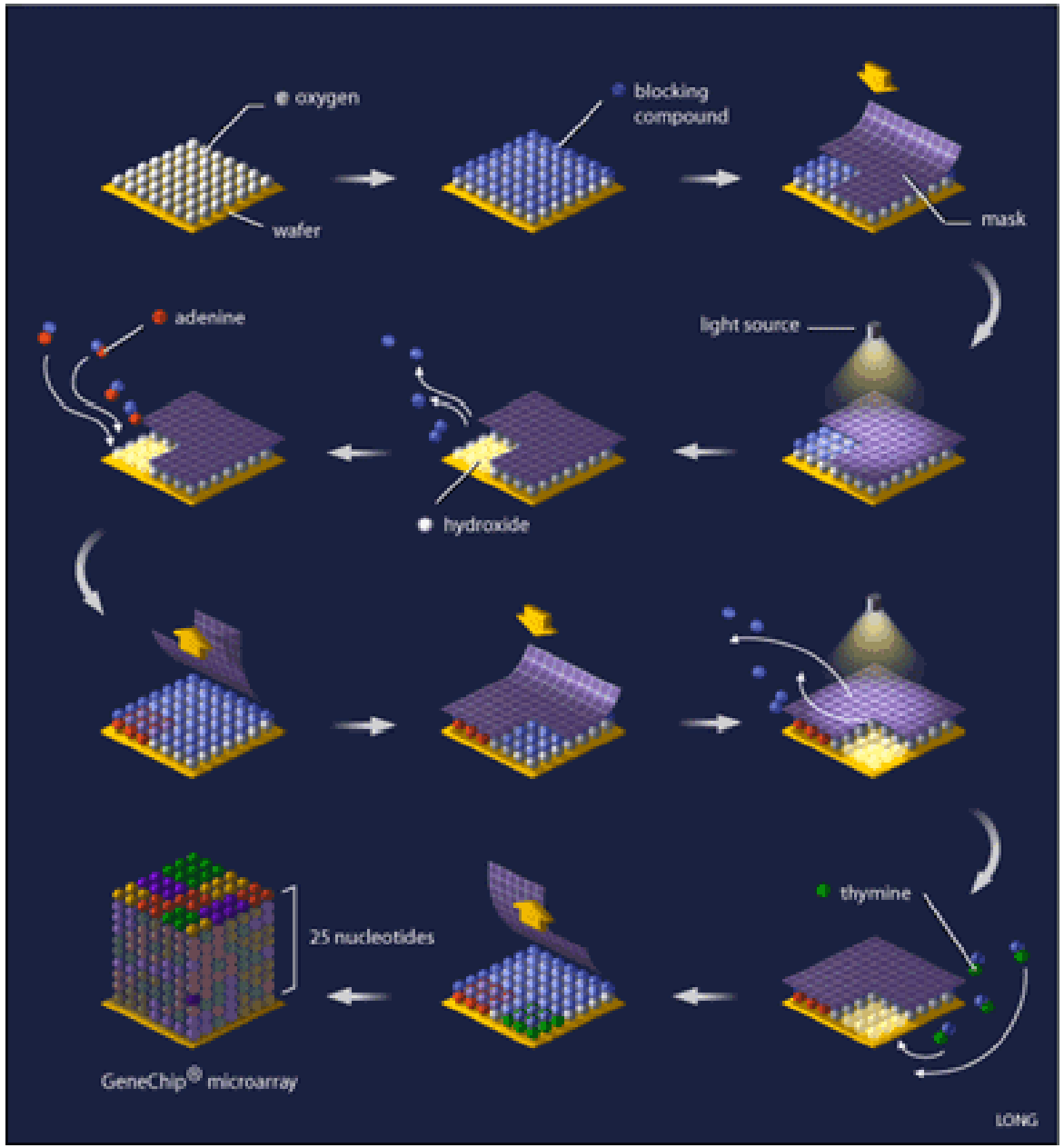


Affymetrix Chip Production - GeneChip® (Photolithography)

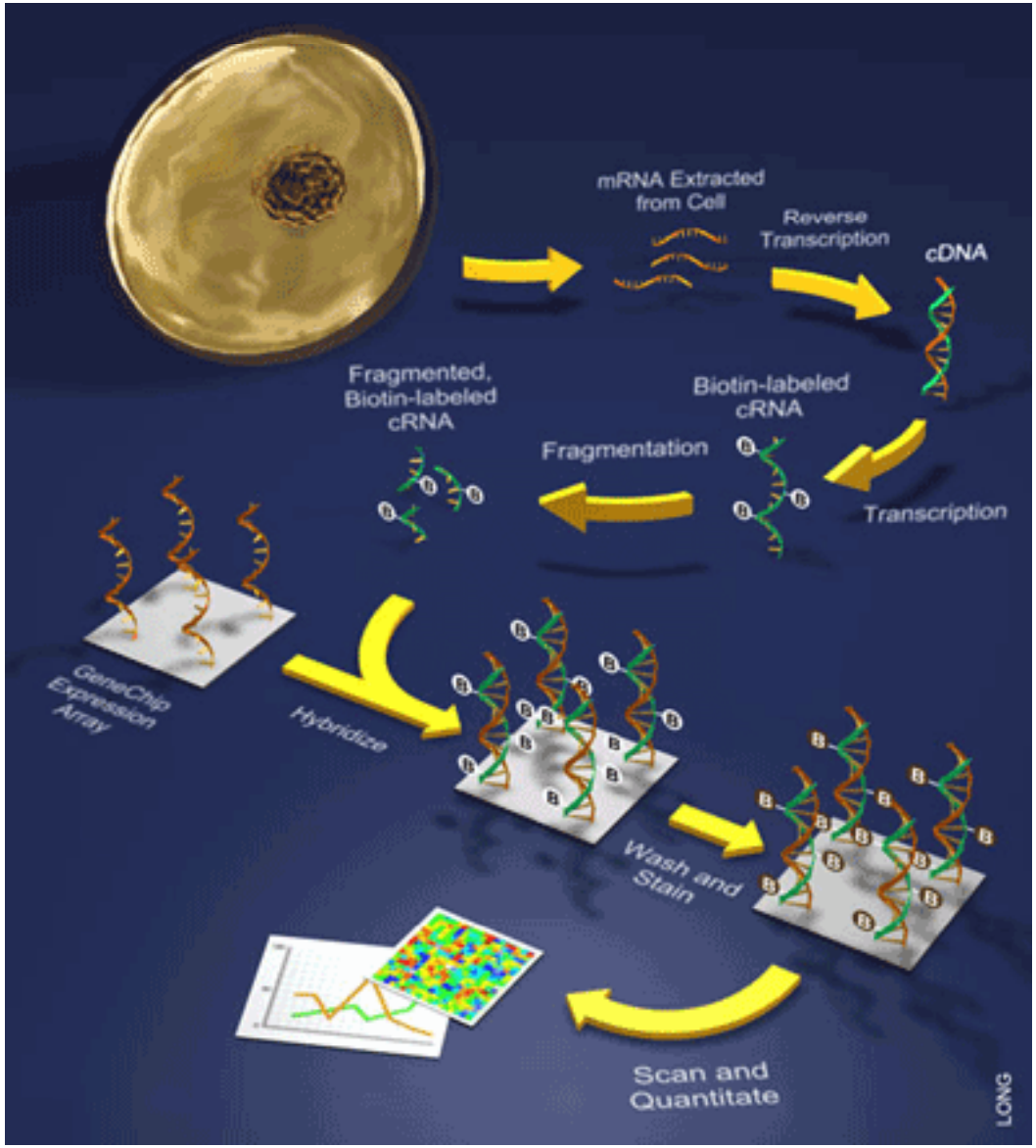





Substrate with base
Robotic deposition
Sample
Strands

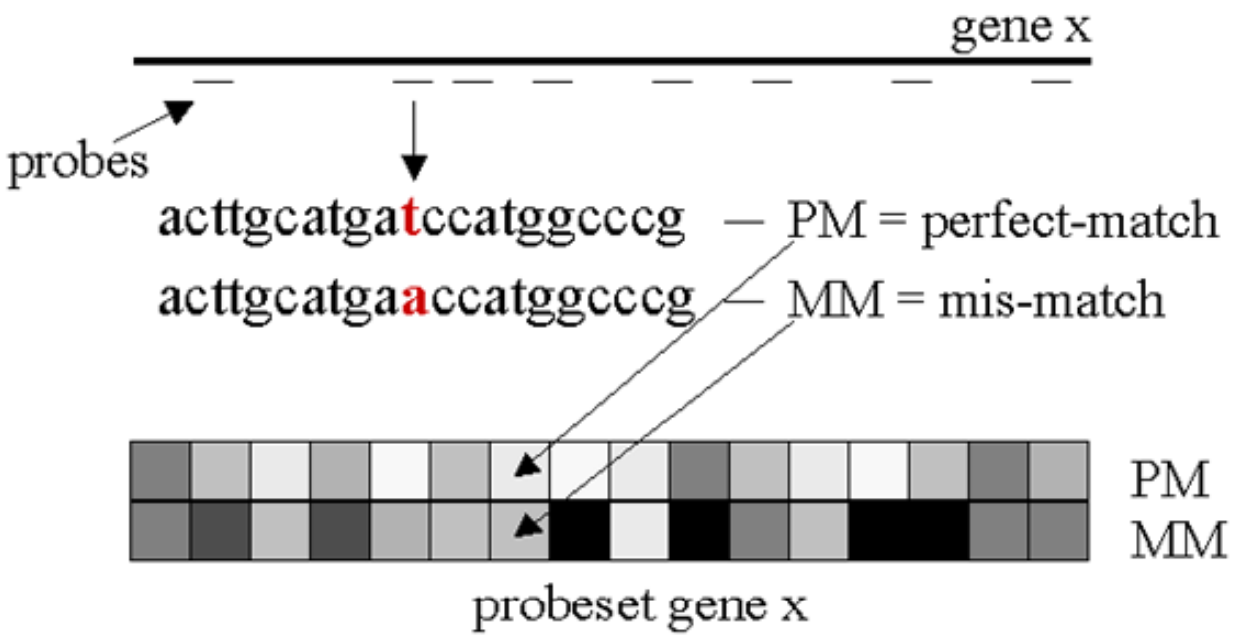


Production of an Affymetrix GeneChip: through the use of photolithography & combinaotrial chemistry specific DNA probes are constructed on the chip surface (Coe & Antler 2004)



The use of **oligonucleotide** arrays. mRNA is extracted from cells and amplified through a process that labels the RNA for analysis. The sample is then applied to an array & and bound RNA stained (Coe & Antler 2004).

Probe Design



R=Discrimination Score = (PM-MM)/(PM+MM)

http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf

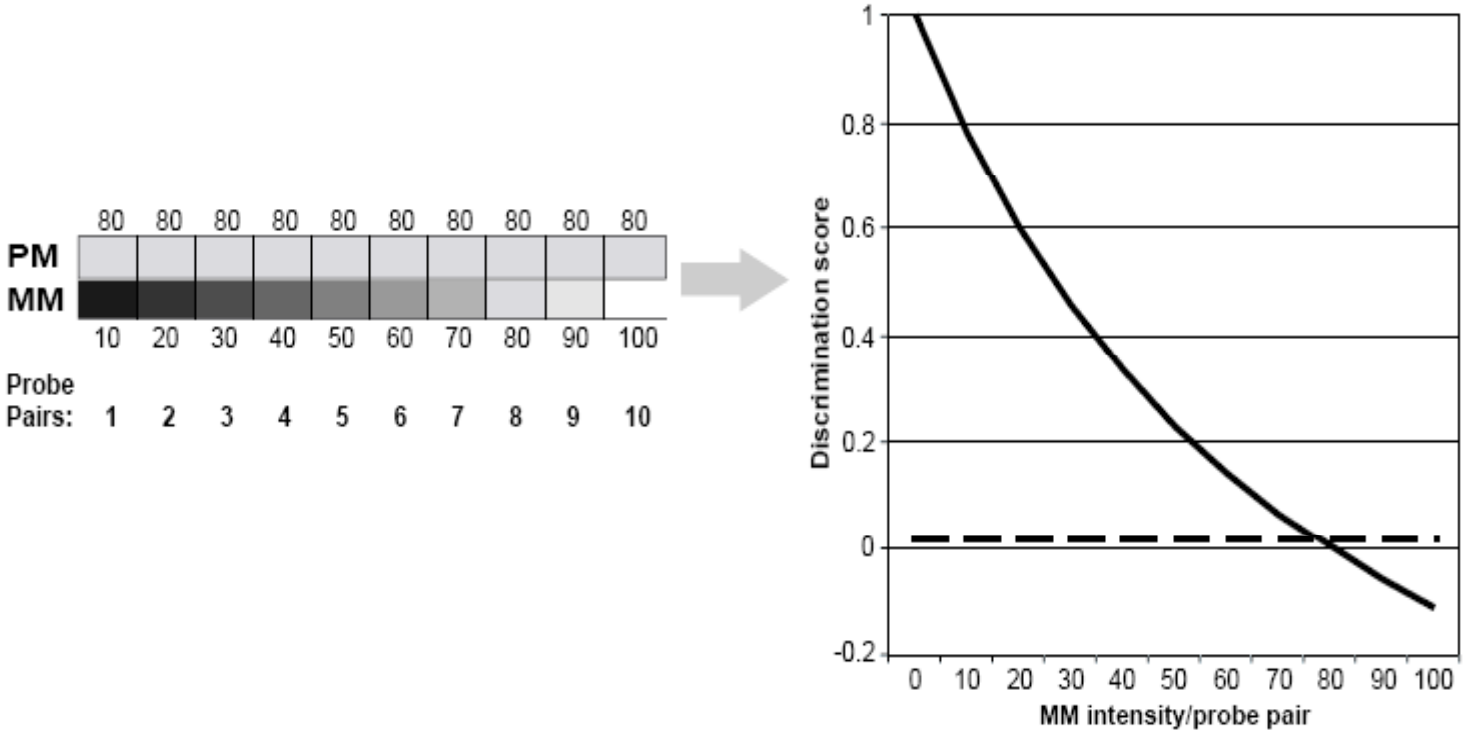
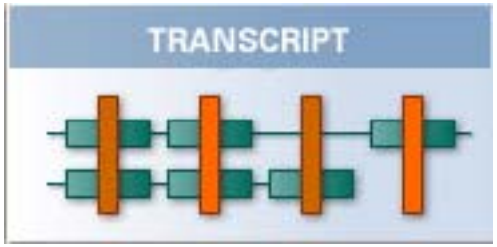


Figure 2. In this hypothetical probe set, the Perfect Match (PM) intensity is 80 and the Mismatch (MM) intensity for each probe pair increases from 10 to 100. The probe pairs are numbered from 1 to 10. As the Mismatch (MM) probe cell intensity, plotted on the x-axis, increases and becomes equal to or greater than the Perfect Match (PM) intensity, the Discrimination score decreases as plotted on the y-axis. More specifically, as the intensity of the Mismatch (MM) increases, our ability to discriminate between the PM and MM decreases. The dashed line is the user-definable parameter Tau (default = 0.015).

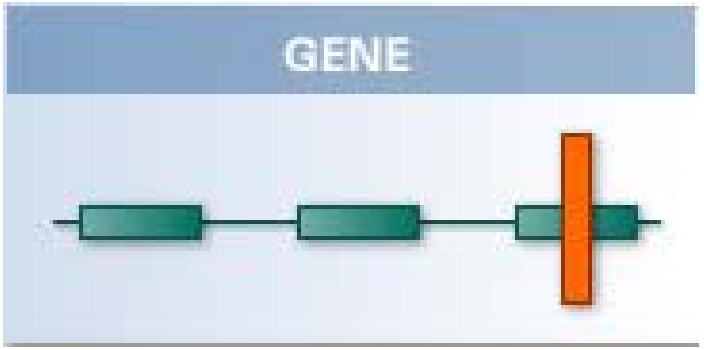
Commercial Chips

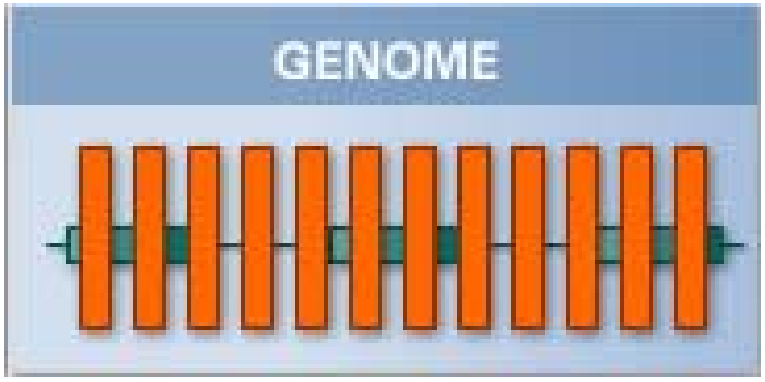
- Clontech, Incyte, Research Genetics
  - **Filter-based arrays** with **up to** about 8,000 clones
- Incyte/Synteni
  - 10,000 probe chips, not distributed (have to send them target RNA)
- Affymetrix
  - **Oligo-based chips** with **12,000 genes** of known function **(13-16 oligos/gene)** and 4x10,000 from ESTs
    - http://www.affymetrix.com/products/arrays/index.affx

Incyte microarray



Affymetrix Designs



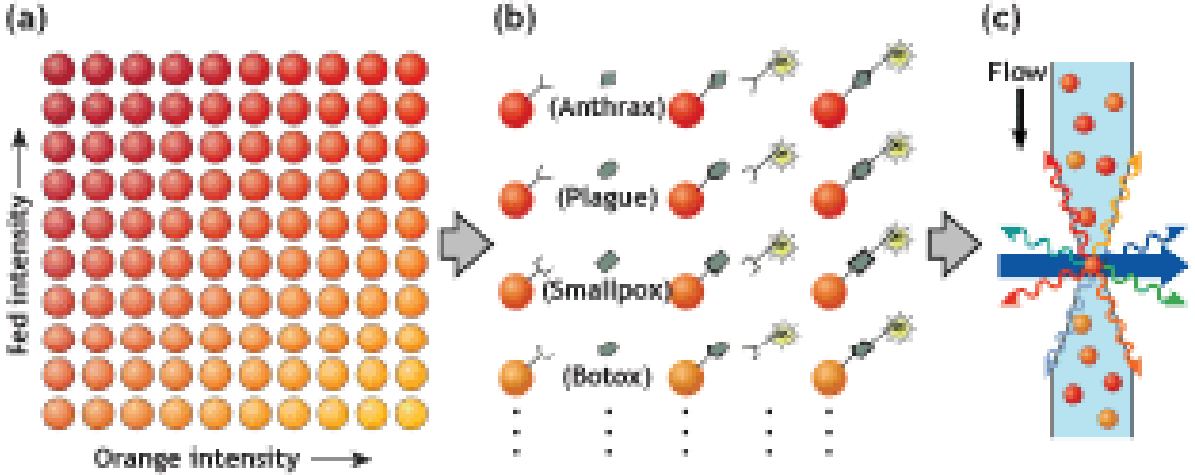


Alternative Technologies

- Synthesis of **probes** on **microbeads**
  - Hybridization **in solution**
  - Identification of beads by **fluorescent bar** coding by embedding transponders
    - **Readout** using **micro-flow cells** or **optic fibers**

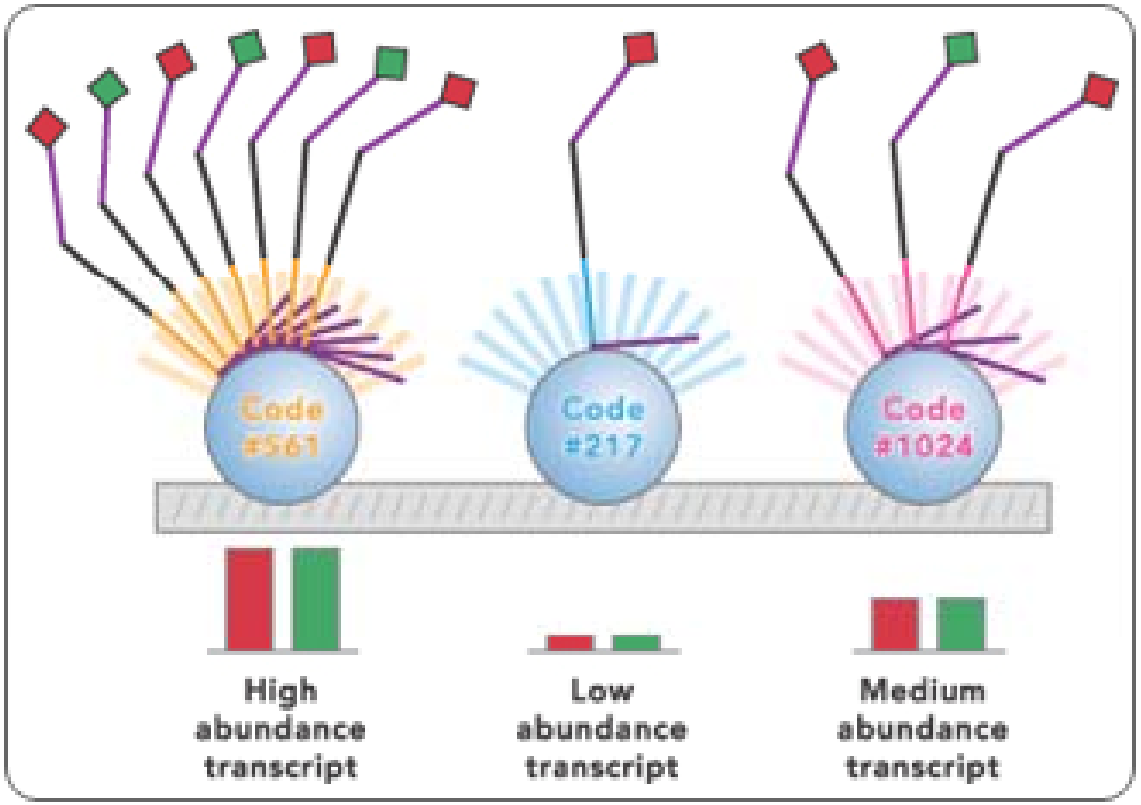
535 Multi-purpose Cell

- Production of **"universal" arrays**
  - Array uses a unique combination of oligos, and **probes** containing **the proper complements**



(a) 100-plex Luminex™ liquid array generated by embedding varying ratios of two different dyes into polystyrene latex microbeads. Each optically encoded microbead has a unique spectral address. (b) Beads are coated with antibodies specific for target antigens. After incubation with the antigens, secondary or detector antibodies are added, followed by addition of a fluorescent molecule, to complete the "antigen sandwich." (c) The beads are analyzed in the flow cytometer. Beads are interrogated one at a time. A red laser classifies the bead, identifying the bead type. A green laser measures the amount of pathogen on the bead surface. The signal is proportional to the antigen concentration.

Two-color Assay: DASL Hybridization of Labeled Amplicons to Bead-based Address Code Sequences on Sentrix Universal Arrays



http://www.illumina.com/products/arraysreagents/universal_arrays.ilmn

Illumina© Universal array

Nucleic Acids Research, 2002, Vol. 30 No. 16 e87

DNA probes on beads arrayed in a capillary, 'Bead-array', exhibited high hybridization performance

Yoshinobu Kohara[1,2,*], Hideyuki Noda[1], Kazunori Okano[1] and Hideki Kambara[1]

[1]Central Research Laboratory, Hitachi Ltd, 1-280 Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, Japan and
[2]Department of Biotechnology and Life Science, Tokyo University of Agriculture and Technology, 2-24-16 Nakacho, Koganei, Tokyo 184-0012, Japan

Received February 11, 2002; Revised June 18, 2002; Accepted June 27, 2002

ABSTRACT

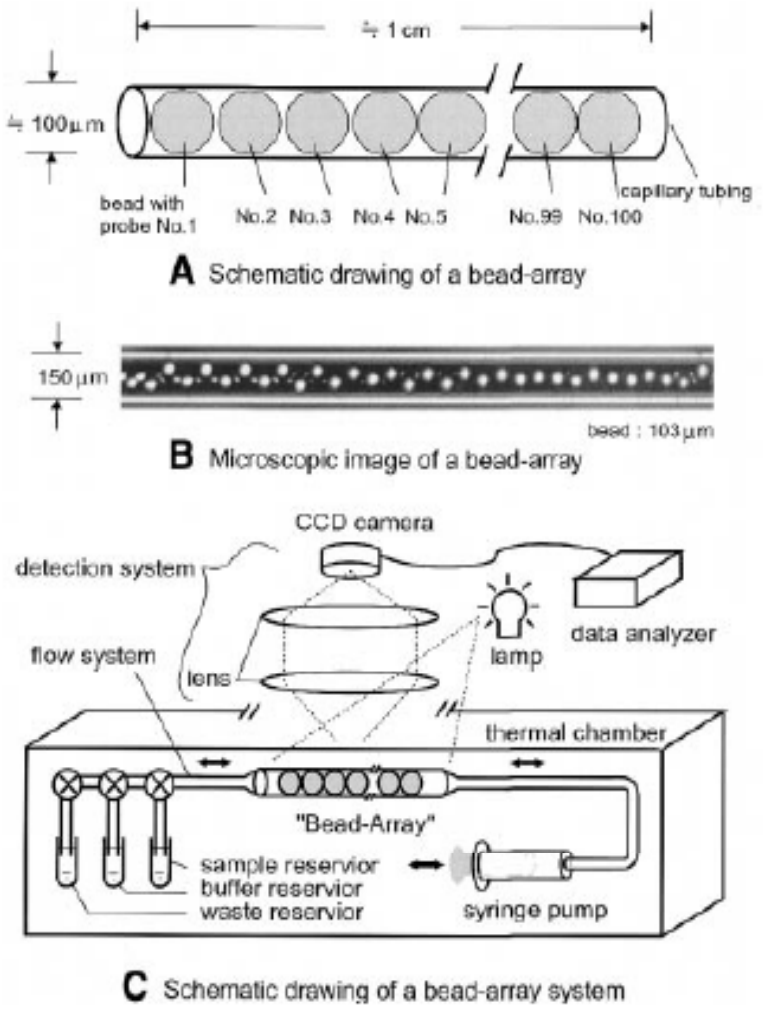
A DNA analysis platform called 'Bead-array' is presented and its features when used in hybridization detection are shown. In 'Bead-array', beads of 100-μm diameter are lined in a determined order in a capillary. Each bead is conjugated with DNA probes, and can be identified by its order in the capillary. This probe array is easily produced by just arraying beads conjugated with probes into the capillary in a fixed order. The hybridization is also easily completed by introducing samples (1–300 μl) into the capillary with reciprocal flow. For hybridization detection, as little as 1 amol of fluorescent-labeled oligo DNA was detected. The hybridization reaction was completed in 1 min irrespective of the amount of target DNA. When the number of target molecules was smaller than that of probe molecules on the bead, 10 fmol, almost all targets were captured on the bead. 'Bead-array' enables reliable and reproducible measurement of the target quantity. This rapid and sensitive platform seems very promising for various genetic testing tasks.

Although it is a very powerful and attractive device, it is very expensive and a practical fabrication method for producing a cost-effective device is still required. In addition, it is impossible to rearrange any of the probes in the array in accordance with changes in the analysis target. This requirement seems to be overcome by micro-spheres having DNA probes. The combination of color-coded microbeads and a flow cytometer (11,12), massive parallel signature sequence, which uses microbeads for cDNA cloning and parallel sequencing reactions (13,14), and fixed microbeads mounted on the terminal wells of optical fibers (15,16) have been reported.

In this study, we demonstrate the excellent characteristics of a new DNA probe array format using beads with DNA probes. This format, 'Bead-array', is an array of DNA probes on beads in a capillary with an order determined by the probe species and hybridization is performed by the reciprocal flow of the sample.

Outline of the probe array in a capillary ('Bead-array')

A schematic view of 'Bead-array' is shown in Figure 1. It is an array of beads conjugated with DNA probes in a narrow capillary in a determined order according to the probe species (Fig. 1A). Each bead has a different DNA probe to capture different DNA targets. The bead-array is designed so as to decrease the volume of reaction space to <0.1 μl to enable fast hybridization. The bead size was determined to be 100 μm,



A: 100 beads with **different probe DNA** are arrayed in a capillary in the intended order

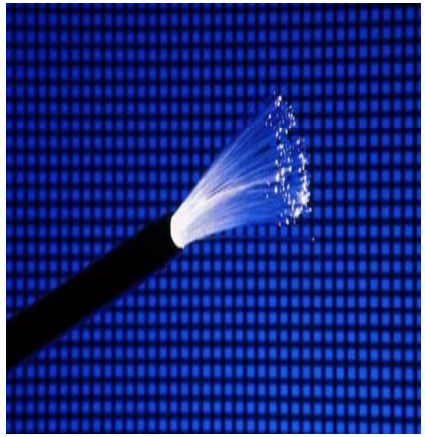
B. Microscopic image

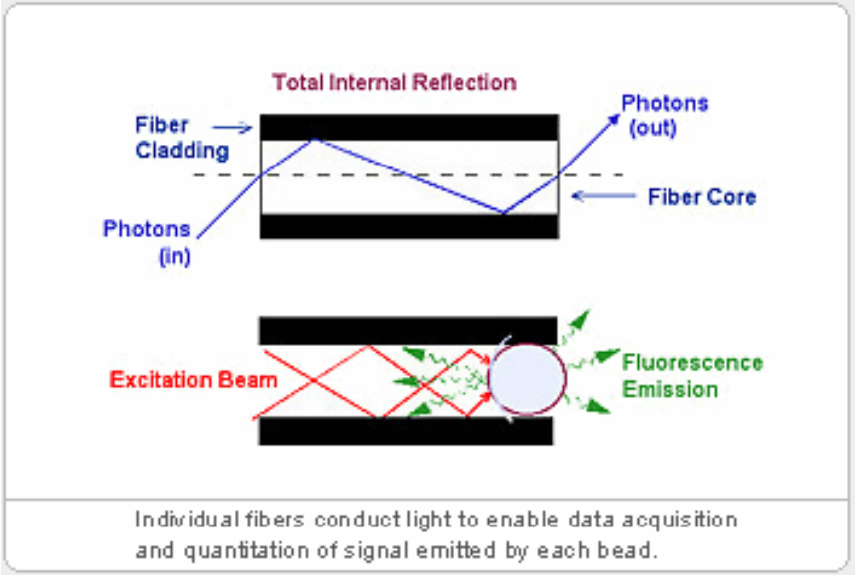
C. A bead-array system

Sample, buffer & waste reservoir

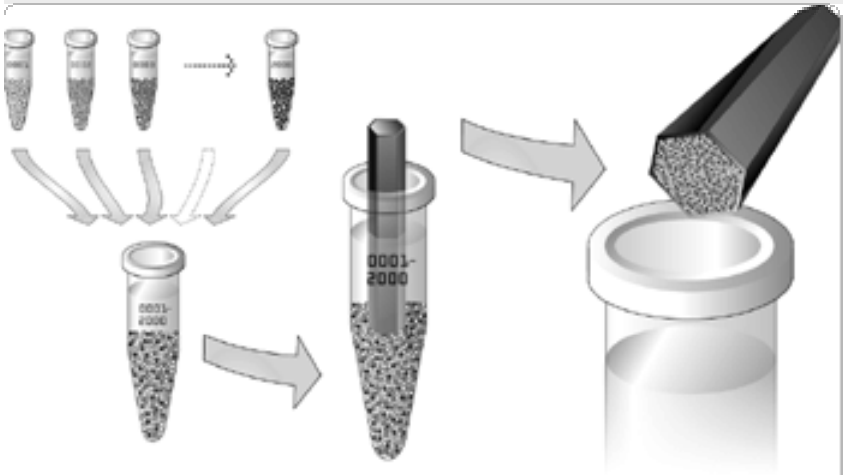
•Sample solution from the sample reservoir **moves back & forth inside the bead-array** during hybridization & buffer solution from the buffer reservoir is introduced during washing

Fiber Optics Technology





Individual fibers conduct light to enable data acquisition and quantitation of signal emitted by each bead.



Bead Pooling

To learn more: Illumina's Web site

Arrays for Genetic Analysis

- **Mutation detection**
  - Molecular Inversion Probe Technology for **SNP Genotyping** (next slide)
    - 20,000 SNPs in a single array
  - PCR followed by primer extension, with **detection of alleles** by **MALDI-TOF** mass spectroscopy (MS) (Sequenom)
- **Gene loss & amplification**
  - Measure **gene dosage** in **genomic DNA** by hybridization to **genomic probes**

Methods

Highly multiplexed molecular inversion probe genotyping: Over 10,000 targeted SNPs genotyped in a single tube assay

Paul Hardenbol,[1] Fuli Yu,[2] John Belmont,[2] Jennifer MacKenzie,[1] Carsten Bruckner,[1] Tiffany Brundage,[1] Andrew Boudreau,[1] Steve Chow,[1] Jim Eberle,[1] Ayca Erbilgin,[1] Mat Falkowski,[1] Ron Fitzgerald,[1] Sy Ghose,[2] Oleg Iartchouk,[1] Maneesh Jain,[1] George Karlin-Neumann,[1] Xiuhua Lu,[2] Xin Miao,[1] Bridget Moore,[1] Martin Moorhead,[1] Eugeni Namsaraev,[1] Shiran Pasternak,[2] Eunice Prakash,[1] Karen Tran,[1] Zhiyong Wang,[1] Hywel B. Jones,[1] Ronald W. Davis,[3] Thomas D. Willis,[1,4] and Richard A. Gibbs[2]

[1]ParAllele BioScience, Inc., South San Francisco, California 94080, USA; [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA; [3]Stanford Genome Technology Center, Stanford University, California 94305, USA

Large-scale genetic studies are highly dependent on efficient and scalable multiplex SNP assays. In this study, we report the development of Molecular Inversion Probe technology with four-color, single array detection, applied to large-scale genotyping of up to 12,000 SNPs per reaction. While generating 38,429 SNP assays using this technology in a population of 30 trios from the Centre d'Etude Polymorphisme Humain family panel as part of the International HapMap project, we established SNP conversion rates of ~90% with concordance rates >99.6% and completeness levels >98% for assays multiplexed up to 12,000plex levels. Furthermore, these individual metrics can be "traded off" and, by sacrificing a small fraction of the conversion rate, the accuracy can be increased to very high levels. No loss of performance is seen when scaling from 6,000plex to 12,000plex assays, strongly validating the ability of the technology to suppress cross-reactivity at high multiplex levels. The results of this study demonstrate the suitability of this technology for comprehensive association studies that use targeted SNPs in indirect linkage disequilibrium studies or that directly screen for causative mutations.

Genome Research 2005 15, 269-75.

Molecular Inversion Probes

Molecular Inversion Probes (Flash Demo) are so named because the oligonucleotide probe central to the process undergoes a unimolecular rearrangement from a molecule that cannot be amplified (step 1), into a molecule that can be amplified (step 6). This rearrangement is mediated by hybridization to genomic DNA (step 2) and an enzymatic "gap fill" process that occurs in an allele-specific manner (step 3). The resulting circularized probe can be separated from cross-reacted or unreacted probes by a simple exonuclease reaction (step 4). Figure 1 shows these steps.

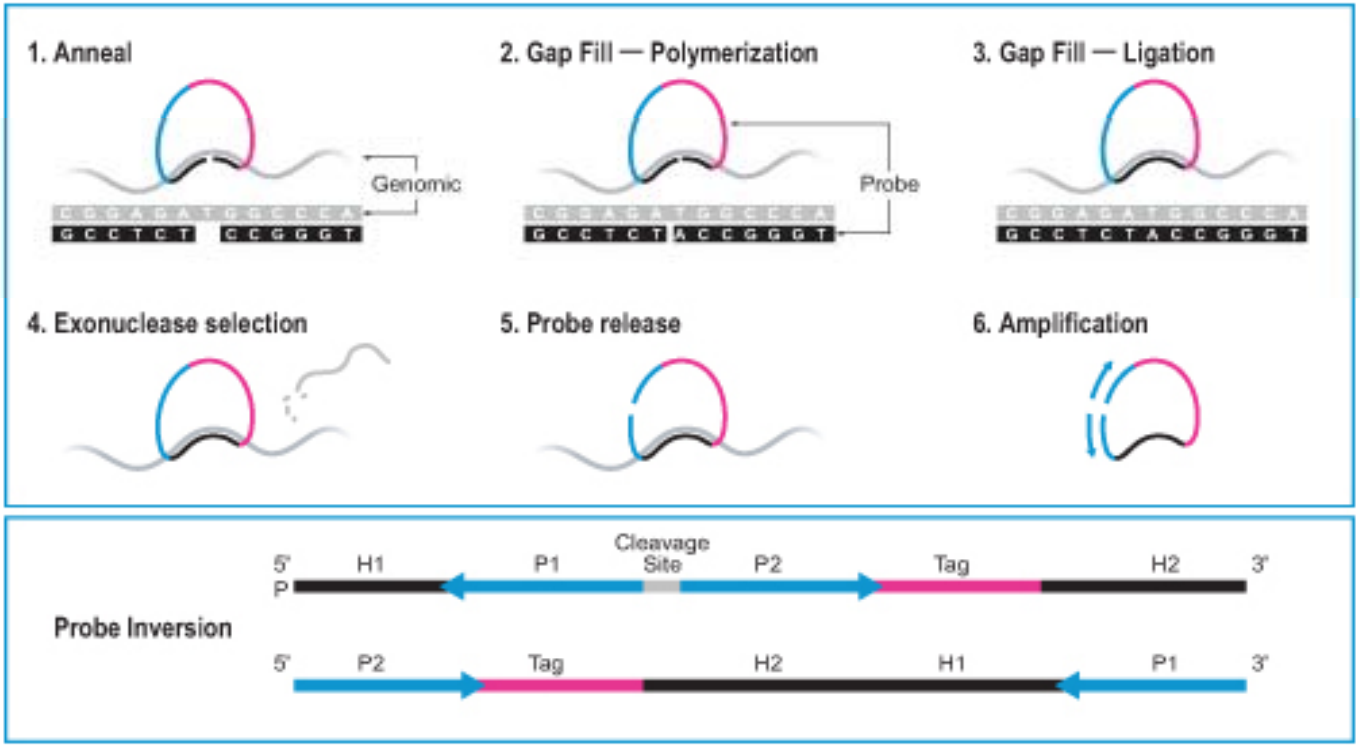


Figure 1: Schematic of the Molecular Inversion Probe

http://www.affymetrix.com/technology/mip_technology.affx#snp

SNP Genotyping Using Molecular Inversion Probes

The SNP genotyping process using molecular inversion probes is outlined diagrammatically in Figure 2a below.

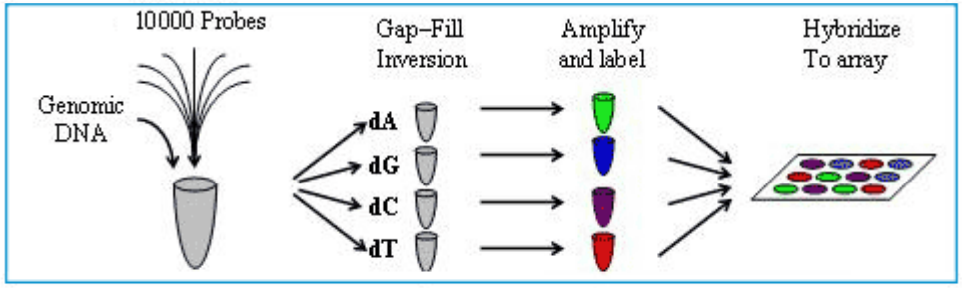
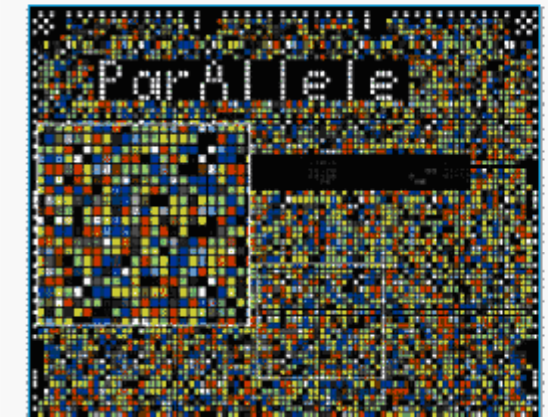


Figure 2a: 10,000 multiplex MIP assay detected on Tag Microarray

- Four-color single array technology; **up to 12, 000** SNPs per reaction
- Amplification with **universal PCR primer pair**

Molecular Inversion probe detection is possible using multiple different detection platforms. Four-color data obtained from the Affymetrix GeneChip shown below.

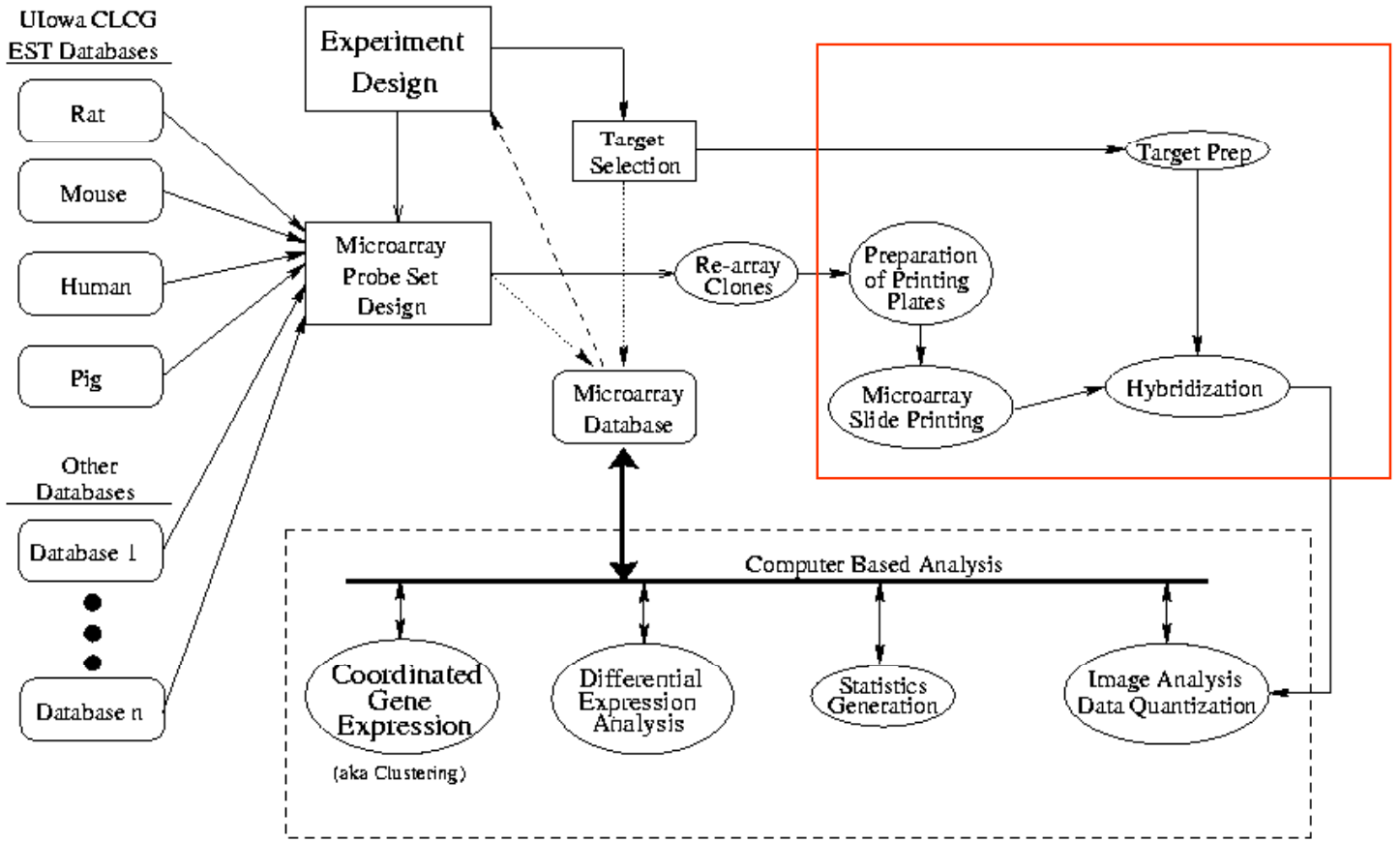


- Each amplified probe contains **a unique tag sequence** that is complementary to a sequence on **the universal tag array**
- **Tags** have been selected to have a **similar Tm & base composition** & to be maximally orthogonal in sequence complementarity

Bioinformatics of Microarrays

- **Array design**: choice of **sequences** to be used as probes
- Analysis of **scanned images**
  - Spot detection, normalization, quantitation
- **Primary analysis** of hybridization data
  - Basic statistics, reproducibility, data scattering, *etc.*
- Comparison of **multiple samples**
  - Clustering, SOMs, k-mean classification …
    - SOMs= self-Organizing Maps (a subtype of artificial neural network, low-dimensional viwes of high-dimensional data)
      - Unsupervised learning
- Sample tracking and databasing of results

Microarray Data Pipeline



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

- Companies are beginning to **sell** microarray expression data (*e.g.* Incyte)

Other Web Links

- Leming Shi's Gene-Chips.com page
  - Very rich source of **basic information** and commercial and academic links
  - DNA chips for dummies animation
- The Big Leagues: Pat Brown and NHGRI microarray projects

SEQUENOM®
Application Note

SNP Discovery Using the MassARRAY™ System

Mathias Ehrich*, Devan Correll, and Dirk van den Boom

SEQUENOM, Inc.
3595 John Hopkins Court
San Diego, California 92121
*correspondence: mehrich@sequenom.com

http://www.coactivepr.com/assets/pdf/writing_samples/sequenom/Genotyping%20Brochure_v8.pdf

Introduction

MassARRAY™ Discovery-RT (SNP Discovery) is a comparative sequence analysis tool based on matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) analysis of nucleic acids cleaved at specific bases. Using the speed and accuracy of the MassARRAY™ system, this innovative method opens new routes for high-throughput discovery and localization of single nucleotide polymorphisms (SNPs). Reference sequences are used to construct *in silico* cleavage patterns and enable cross-correlation of theoretical and experimental mass signal patterns. Observed signal pattern differences are indicators of sequence variations and form the basis for SNP analysis.

A 300-700 base pair (bp) sequence stretch of interest is amplified by PCR. Two PCRs are performed. One reaction introduces a T7-promoter tag in the forward strand of the amplification product. The other PCR introduces the T7-promoter tag in the reverse strand of the product. PCR amplification is followed by *in vitro* transcription, where each PCR product is split into two cleavage reactions (T Cleavage and C Cleavage). Introduction of modified nucleotides during transcription mediates base-specific cleavage in each of the four reactions during the subsequent RNase A treatment. Resulting cleavage products are measured by MALDI-TOF MS, generating a characteristic signal pattern based on the fragment masses.

2004

Genome Research
www.genome.org

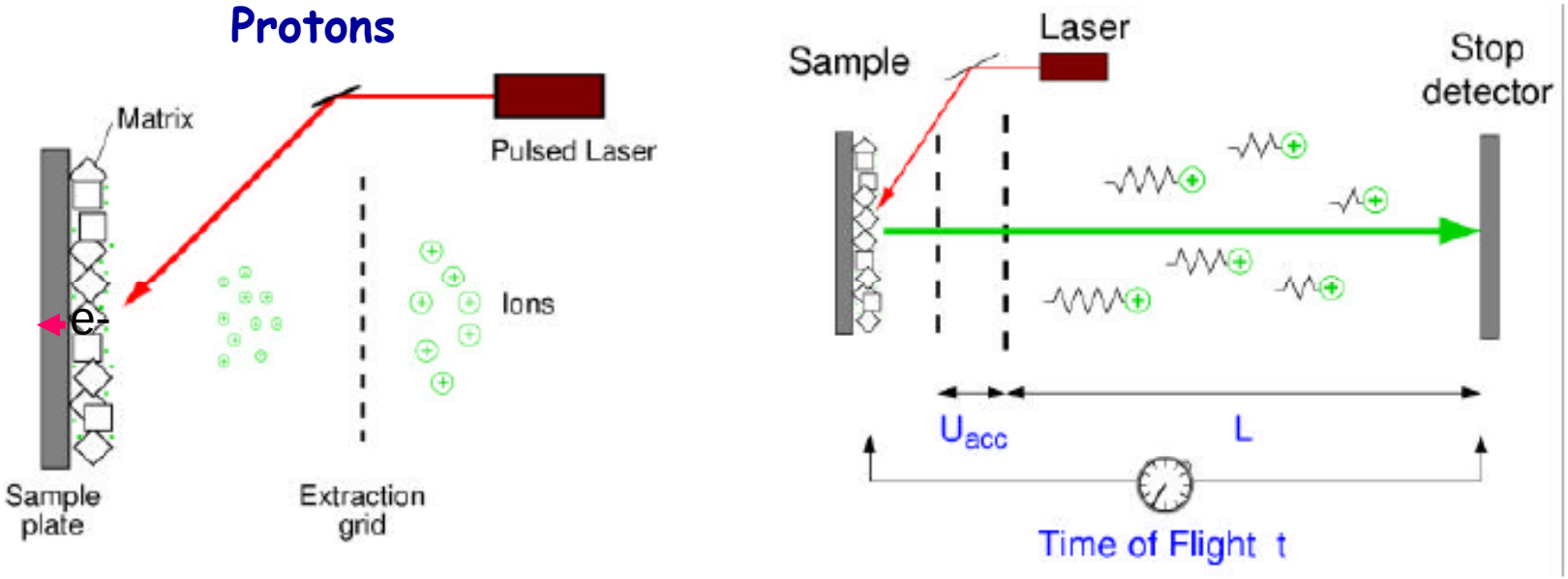
14:126–133 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

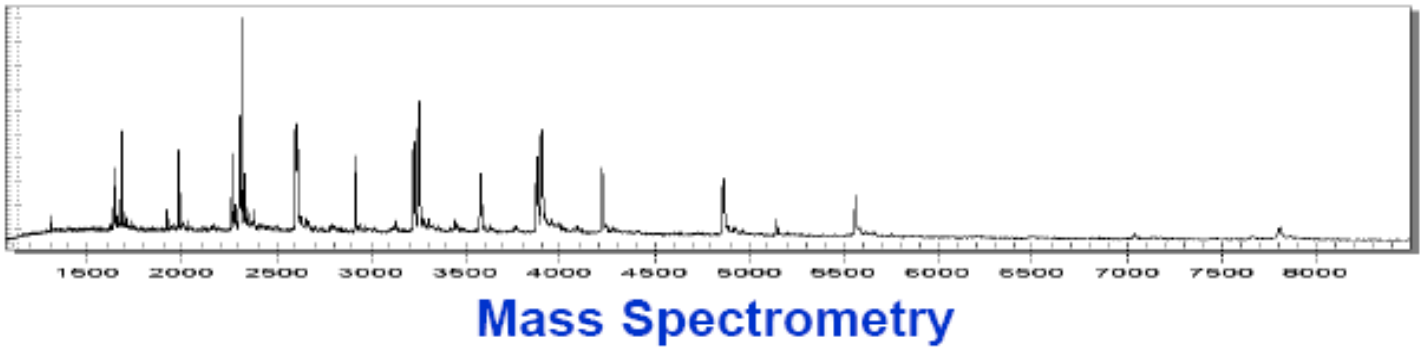
High-Throughput MALDI-TOF Discovery of Genomic Sequence Polymorphisms

Patrick Stanssens,[1,3] Marc Zabeau,[1,3] Geert Meersseman,[1] Gwen Remes,[1] Yannick Gansemans,[1] Niels Storm,[2] Ralf Hartmer,[2] Christiane Honisch,[2] Charles P. Rodi,[2] Sebastian Böcker,[2] and Dirk van den Boom[2,4]

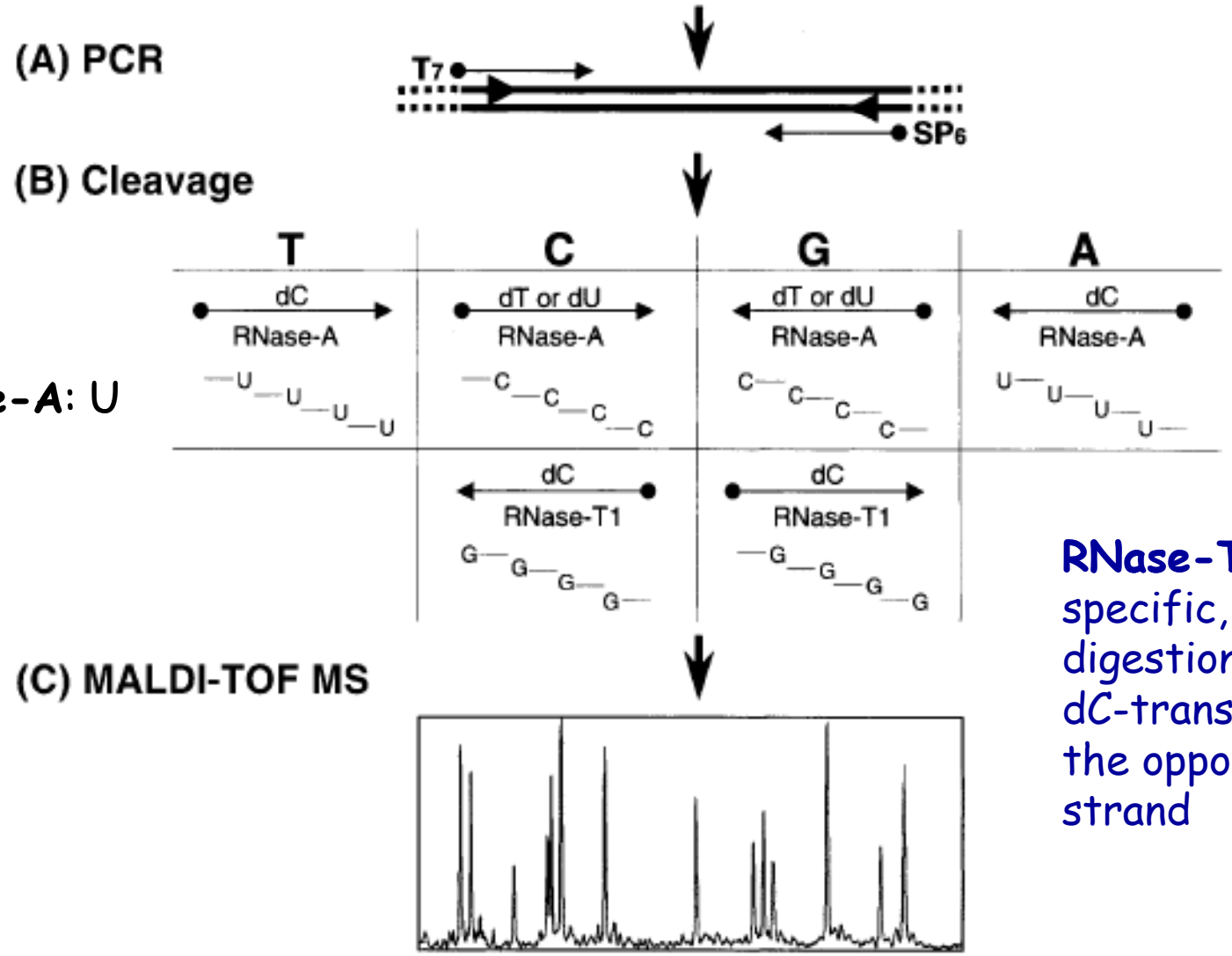
[1]*Methexis Genomics NV, B-9052 Zwijnaarde, Belgium;* [2]*SEQUENOM, Inc., San Diego, California 92121, USA*

We describe a comparative sequencing strategy that is based on matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) analyses of complete base-specific cleavage reactions of a target sequence. The target is converted to a DNA/RNA mosaic structure after PCR amplification using in vitro transcription. Cleavage with defined specificity is achieved by ribonucleases. The set of cleavage products is subjected to mass spectrometry without prior fractionation. The presented resequencing assay is particularly useful for single-nucleotide polymorphism (SNP) discovery. The combination of mass spectra from four complementary cleavage reactions detects approximately 98% of all possible homozygous and heterozygous SNPs in target sequences with a length of up to 500 bases. In general, both the identity and location of the sequence variation are determined. This was exemplified by the discovery of SNPs in the human gene coding for the cholesteryl ester transfer protein using a panel of 96 genomic DNAs.





matrix-assisted laser desorption/inoization



RNase-A: U and C

RNase-T1: G-specific, digestion of the dC-transcript of the opposite strand

Figure 1 Schematic outline of the resequencing strategy using chip-based mass spectrometry. (*A*) The target region is first PCR-amplified with primers bearing bacteriophage T7 and SP6 RNA polymerase promoter sequences (dashed lines). (*B*) A mosaic transcript, with dCMP or dTMP/dUMP replacing the regular nucleoside, is derived from each strand of the amplicon (represented by the arrows) and base-specifically cleaved (see text for details). (*C*) Finally, the set of cleavage products, as a group, is analyzed by an array mass spectrometer.

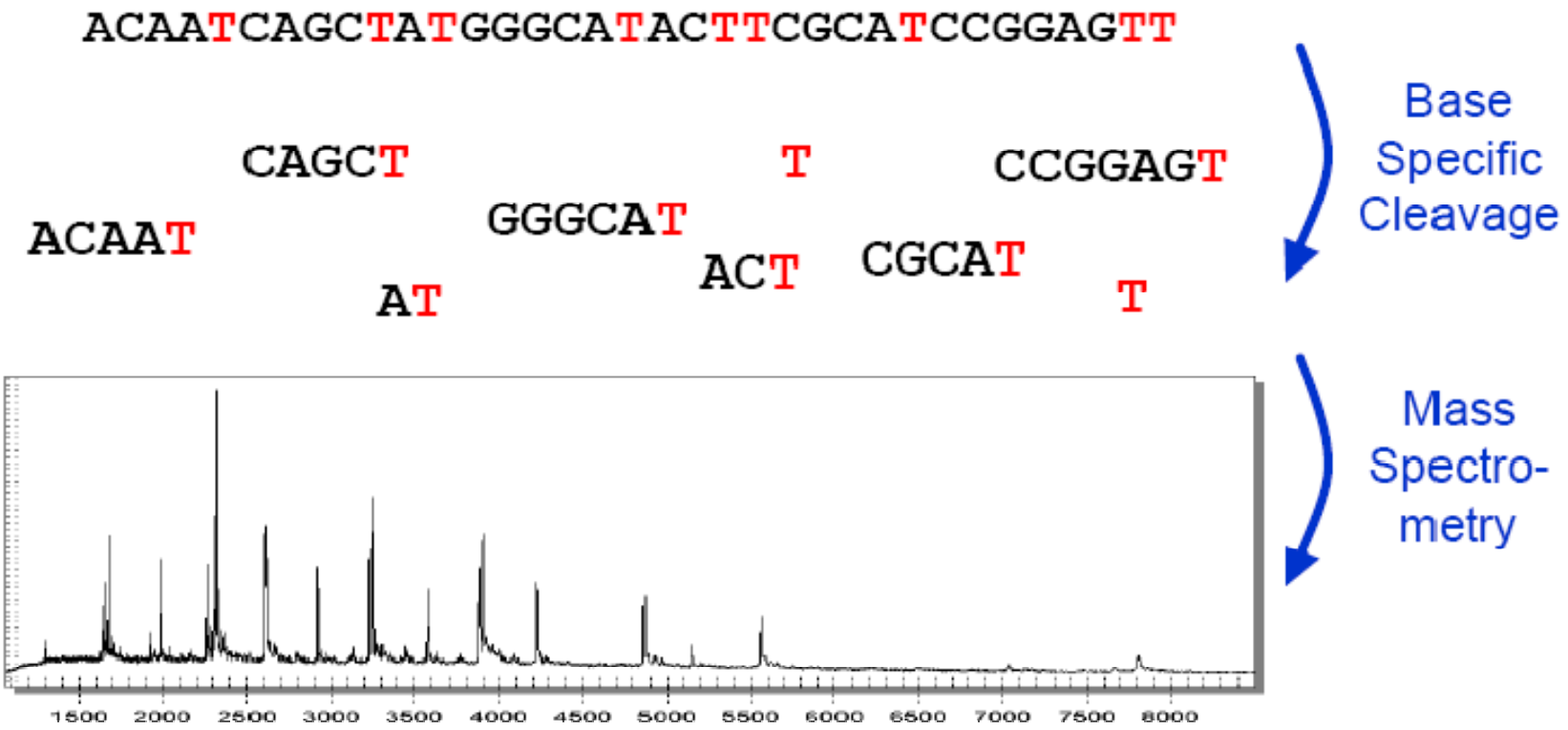
Single Nucleotide Polymorphisms

```
5′...ACCATGCT[A/C]ACAATCGAG...3′
3′...TGGTACGA[T/G]TGTTAGCTC...5′
```

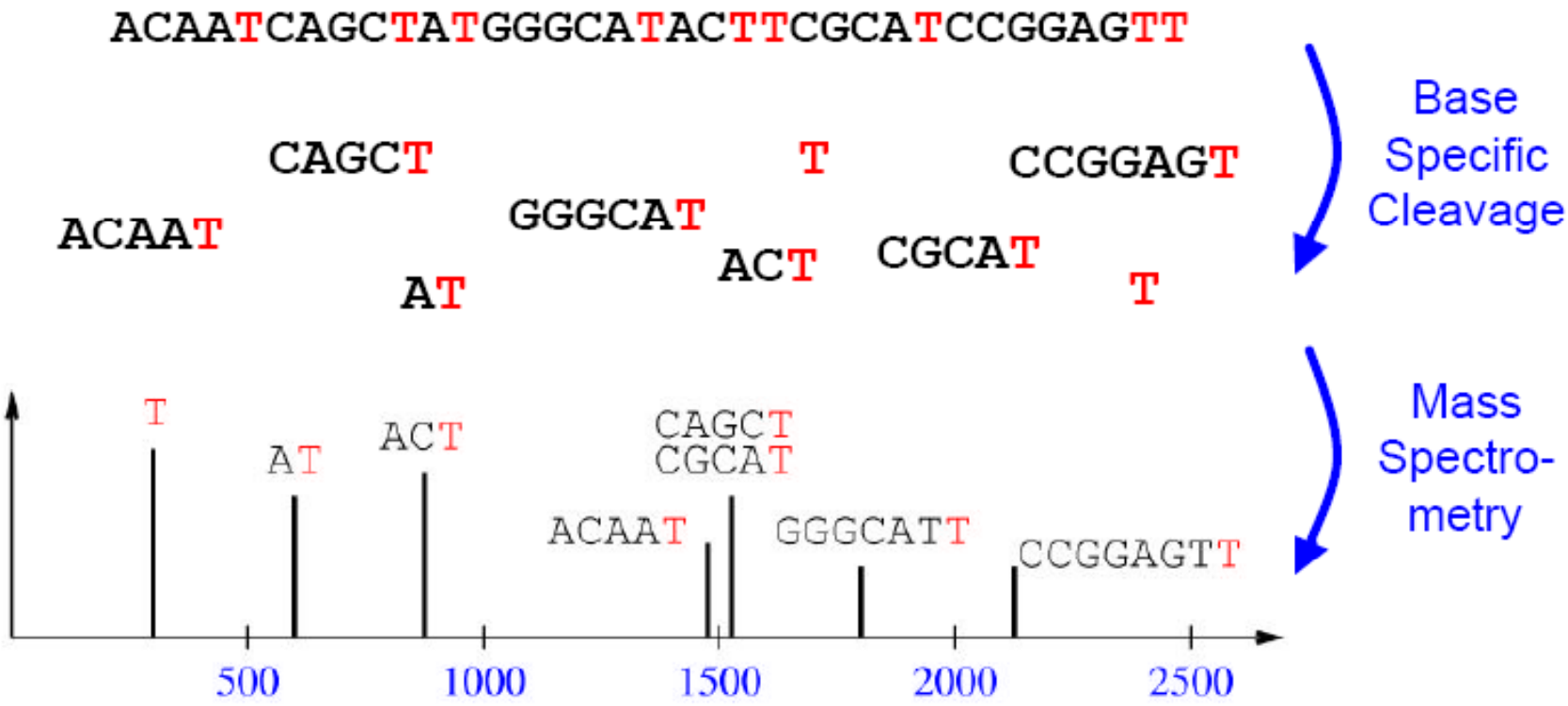
- large parts of an organism's genome are constant across all individuals of a population
- at certain positions, two or more alternative bases can be observed
- play an important role in areas such as disease predisposition or drug side effect predisposition

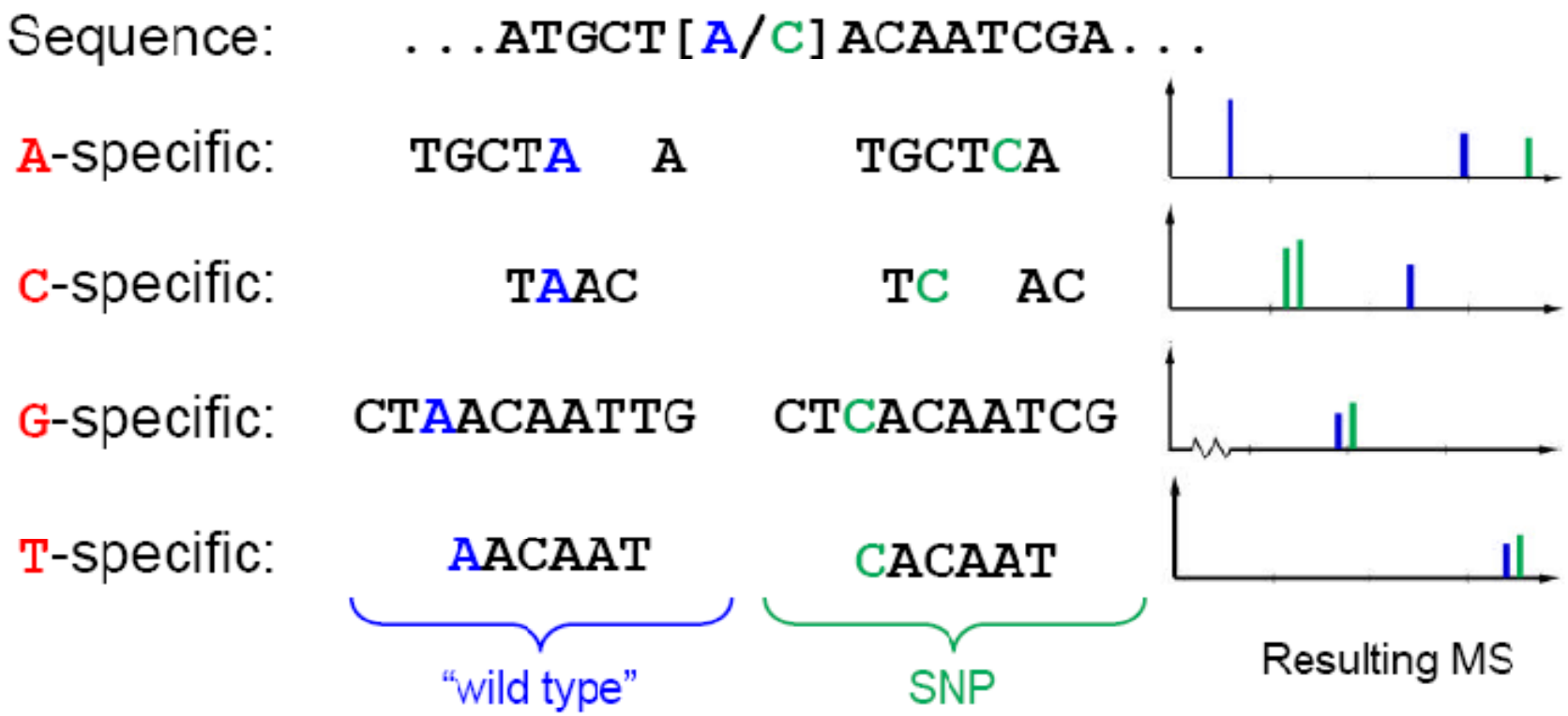
RNase-A: U and C

Base specific cleavage of DNA or RNA can be achieved by the use of RNAses, UDG, and others.



Base specific cleavage and MALDI-TOF mass spectrometry can be simulated *in silico* for a given reference sequence.



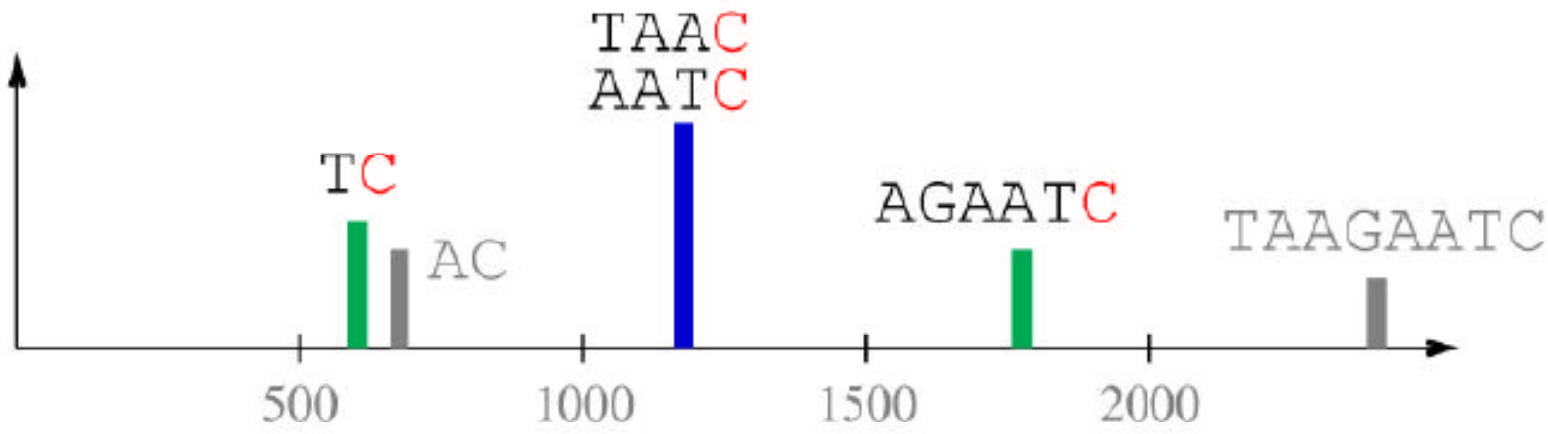


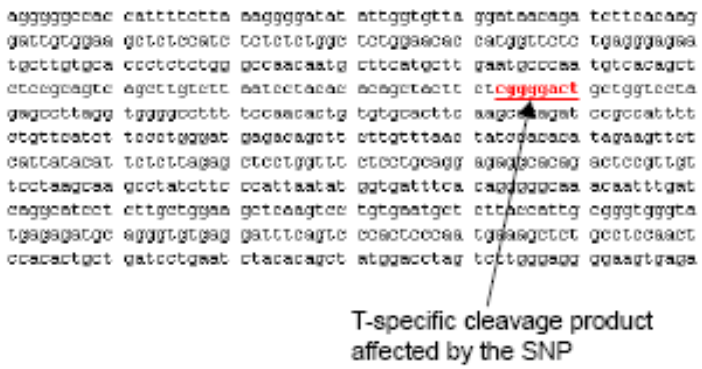
- Trivial approach: Simulate spectra for all possible sequence variations, compare with measured mass spectra
- time consuming, especially for "close" SNPs

If two sequence variations are close, they can change the MS in “complicated” ways:

5′...ACCATGCT[A/C]A[C/G]AATCGAG...3′

C-specific cleavage:





Comparison of reference and affected fragments

5'-...CACAGCTACTTCTC[G/A]GGGACTGC...3'

3'-...GTGTCGATGAAGAG[C/T]CCCTGACG..5'

	Reference G-allele		Affected A-allele
T Forward	CGGGGACT 2613 Da	*Mass Shift -16 Da*	CAGGGACT 2597 Da
C Forward	GGGGAC 2034 Da	*Mass Shift -16 Da*	AGGGAC 2018 Da
T Reverse	CCCCGAGAAGT 3505 Da	*Introduction of cleavage site*	CCCT 1192 Da GAGAAGT 2348 Da
C Reverse	C 324 Da GAGAAGTAGC 3326 Da	*Removal of cleavage site*	TGAGAAGTAGC 3630 Da

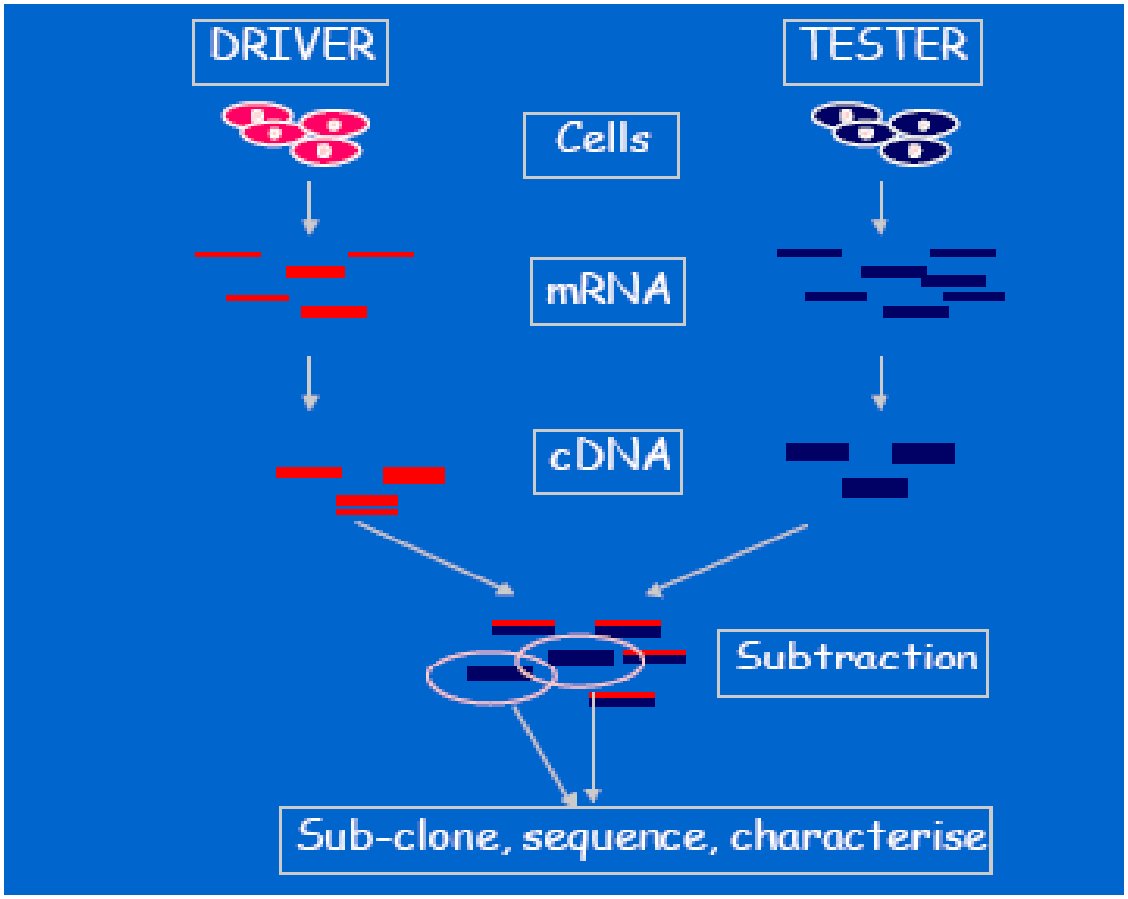
Figure 3. Fragments Affected by SNP

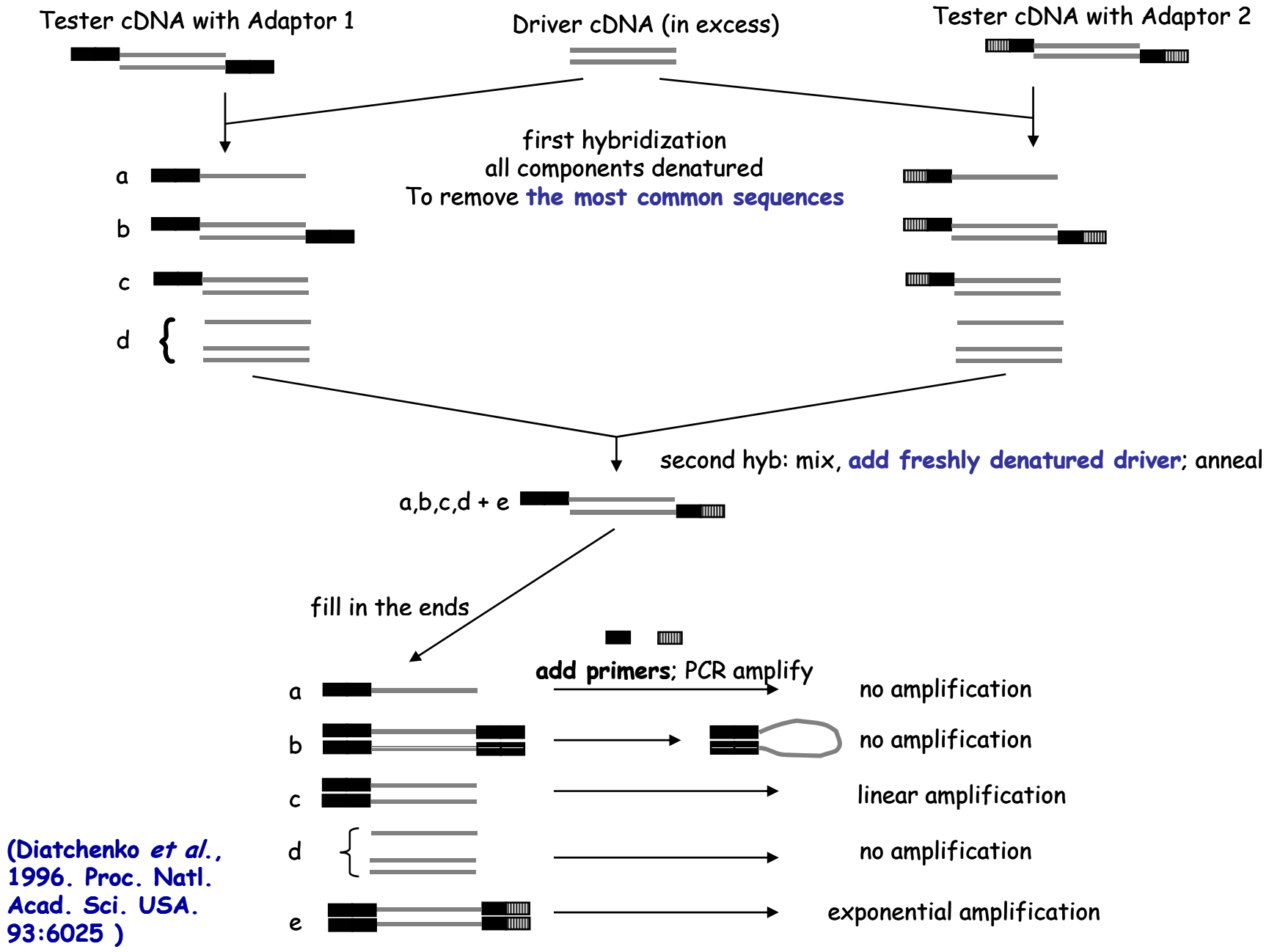
- A sequence change can have **multiple affects** on the mass spectra
- It can result in a mass shift, **introduction** of a cleavage site or **removal** of a cleavage site
- The forward reactions **indicate** the presence of a SNP through mass shift
- The reverse reactions **pinpoint the location of the SNP** in the amplicon reference sequence

Only One Final Word of Wisdom...

- "...although the computer is a wonderful helpmate for **the sequence searcher** and comparer, **biochemists** and **molecular biologists** must guard against **the blind acceptance** of any algorithmic output; given the choice, **think like a biologist and not a statistician**"
  - Russell F. Doolittle, 1990

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

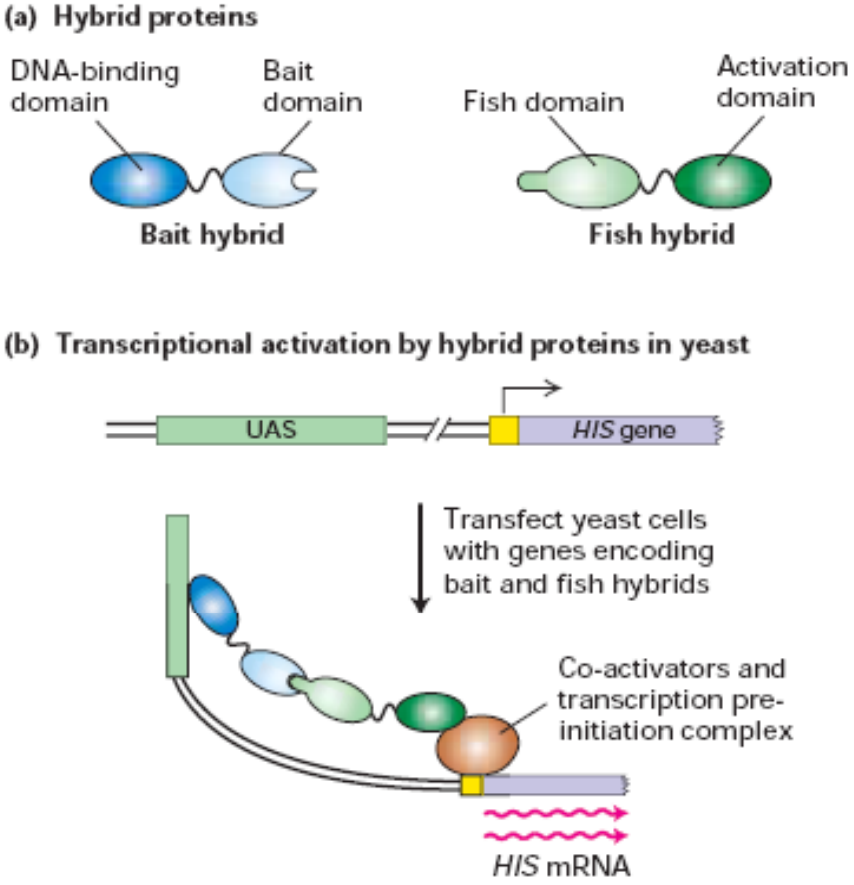
- Diatchenko *et al.* 1996 (PNAS **93:6025)**
  - Could detect as little as **0.001% target**
- Critical factor is **relative concentration** of target in **tester** and **driver** populations
- Effective enrichment when
  - Target present at **>= 0.01%**
  - Concentration ratio **>= 5-fold**

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**

Yeast Two-Hybrid System (1)

- Protein-protein interaction
- A yeast vector for expressing a **DNA-binding domain**
  - Flexible linker region **without the associated activation domain**, *e.g.*, the deleted *GAL4* containing amino acids 1-692
- A cDNA sequence encoding a protein or protein domain of interest = **bait domain** is **fused in frame** to **the flexible linker region** so that the vector will express **a hybrid protein composed** of the **DNA-binding domain**, **linker region**, and **bait domain**



◀ EXPERIMENTAL FIGURE 11-39 The yeast two-hybrid system provides a way of screening a cDNA library for clones encoding proteins that interact with a specific protein of interest. This is a common technique for screening a cDNA library for clones encoding proteins that interact with a specific protein of interest. (a) Two vectors are constructed containing genes that encode hybrid (chimeric) proteins. In one vector (*left*), coding sequence for the DNA-binding domain of a transcription factor is fused to the sequences for a known protein, referred to as the "bait" domain (light blue). The second vector (*right*) expresses an activation domain fused to a "fish" domain (green) that interacts with the bait domain. (b) If yeast cells are transformed with vectors expressing both hybrids, the bait and fish portions of the chimeric proteins interact to produce a functional transcriptional activator. In this example, the activator promotes transcription of a *HIS* gene. One end of this protein complex binds to the upstream activating sequence (UAS) of the *HIS3* gene; the other end, consisting of the activation domain, stimulates assembly of the transcription preinitiation complex (orange) at the promoter (yellow). (c) To screen a cDNA library for

Molecular Cell Biology,
Lodish 5th Ed.

Yeast Two-Hybrid System (2)

- ***A cDNA library*** is cloned into **multiple copies** of a second yeast vector that encodes a strong activation domain & flexible linker, to produce **a vector library expressing multiple hybrid proteins**, each containing a different fish domain

- The bait vector & library of fish vectors are then transfected into **engineered yeast cells** in which **the only copy of a gene** required for histidine synthesis (HIS) is **under control of a UAS** with binding sites for the DNA-binding domain of the hybrid bait protein

- Transformed cells that express the **bait hybrid & interacting fish hybrid** will be able **to activate transcription of the HIS gene**

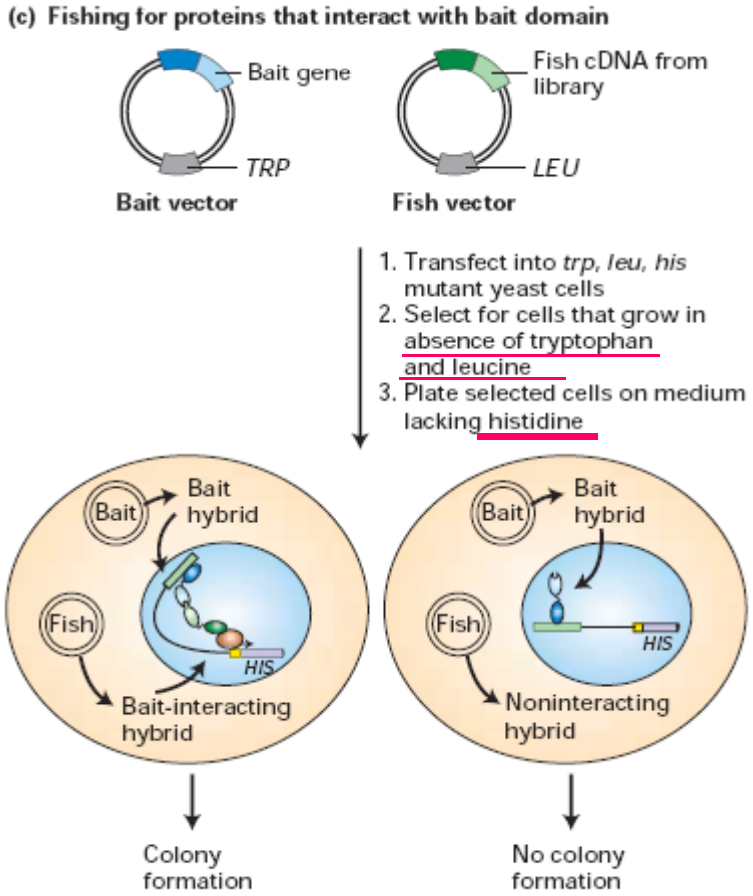
- The flexibility in **the spacing between** the DNA-binding & activation domains of eukaryotic activators makes this system work

Yeast Two-Hybrid System (3)

- A two-step selection process is used
- The bait vector also expresses **a wild-type TRP gene**, and the hybrid vector expresses **a wild-type LEU gene**
- Transfected cells are first grown in a medium that lack of tryptophan & leucine but contain histindine
  - Only cells that have taken up **the bait vector & one of the fish plasmids** will survive in this medium
- The cells that survive then are plated on a medium that **lacks histidine**

Yeast Two-Hybrid System (4)

- Those cells expressing a fish hybrid that does not bind to the bait hybrid cannot transcribe the HIS gene & consequently will not form a colony on medium lacking **histidine**
- The few cells that express **a bait-binding fish hybrid** will grow & form colonies in the absence of histindine
- Recovery of the **fish vectors** from these colonies yields **cDNA encoding protein domains that interact with the bait domain**



(orange) at the promoter (yellow). (c) To screen a cDNA library for clones encoding proteins that interact with a particular bait protein of interest, the library is cloned into the vector encoding the activation domain so that hybrid proteins are expressed. The bait vector and fish vectors contain wild-type selectable genes (e.g., a *TRP* or *LEU* gene). The only transformed cells that survive the indicated selection scheme are those that express the bait hybrid and a fish hybrid that interacts with it. See the text for discussion. [See S. Fields and O. Song, 1989, *Nature* **340**:245.]

Coffee Break

- What do boxers and astronomers have in common?
- They both see stars!!!