Measuring Gene Expression Part 2

David Wishart

Bioinformatics 301

david.wishart@ualberta.ca

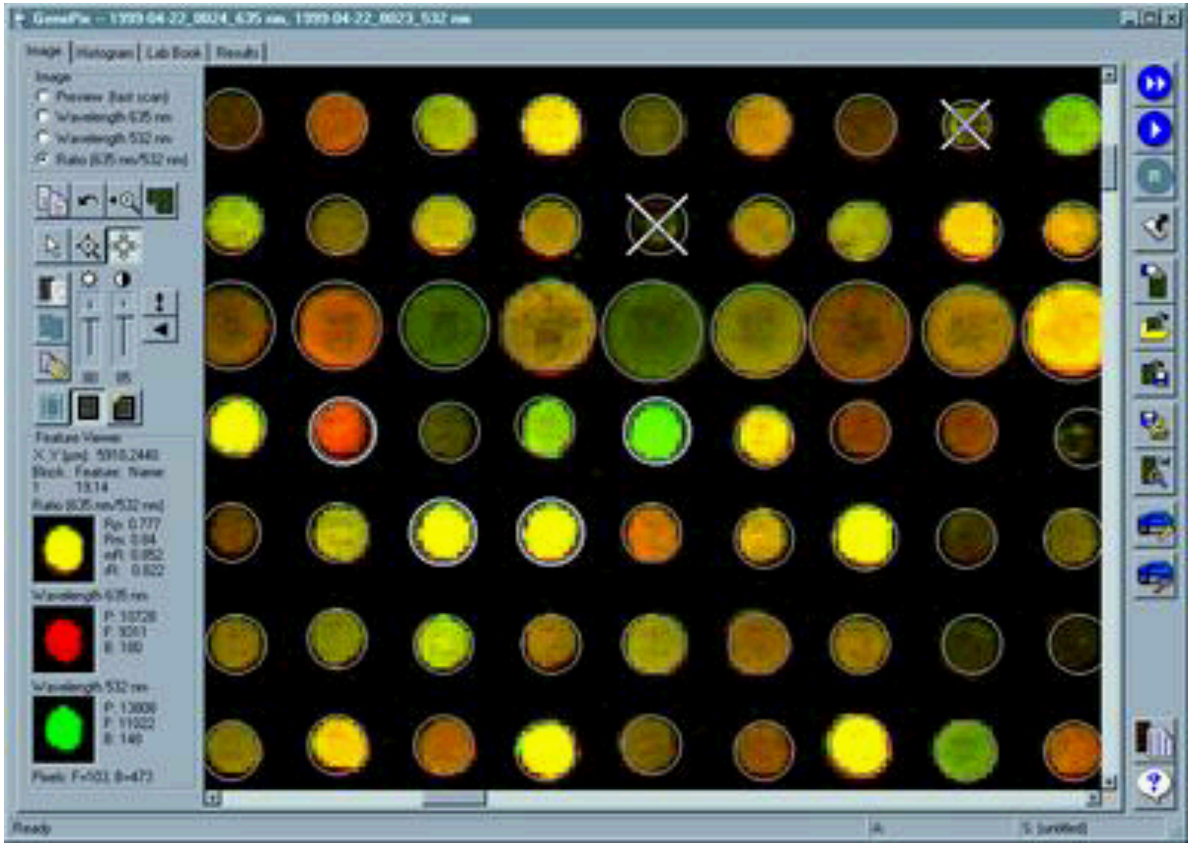
Objectives

- Review of detailed principles of microarrays (methods, data collection)
- Understand differences between spotted arrays versus Affy gene chips (advantages/disadvantages)
- Steps to doing microarrays and possible sources of error

Measuring Gene Expression*

- Differential Display
- Serial Analysis of Gene Expression (SAGE)
- RNA-Seq
- RT-PCR (real-time PCR)
- Northern/Southern Blotting
- DNA Microarrays or Gene Chips

Microarrays



GenePix – 1999-04-22_0024_635 nm, 1999-04-22_0023_532 nm

Image
○ Preview (fast scan)
○ Wavelength 635 nm
○ Wavelength 532 nm
● Ratio (635 nm/532 nm)

Feature Viewer
X,Y [µm]: 5910,2440
Block: Feature: Name
1 13,14
Ratio (635 nm/532 nm)
Rp: 0.777
Rm: 0.84
mR: 0.852
rR: 0.822
Wavelength 635 nm
P: 10728
F: 9261
B: 180
Wavelength 532 nm
P: 13808
F: 11022
B: 148
Pixels: F=503, B=473

Ready

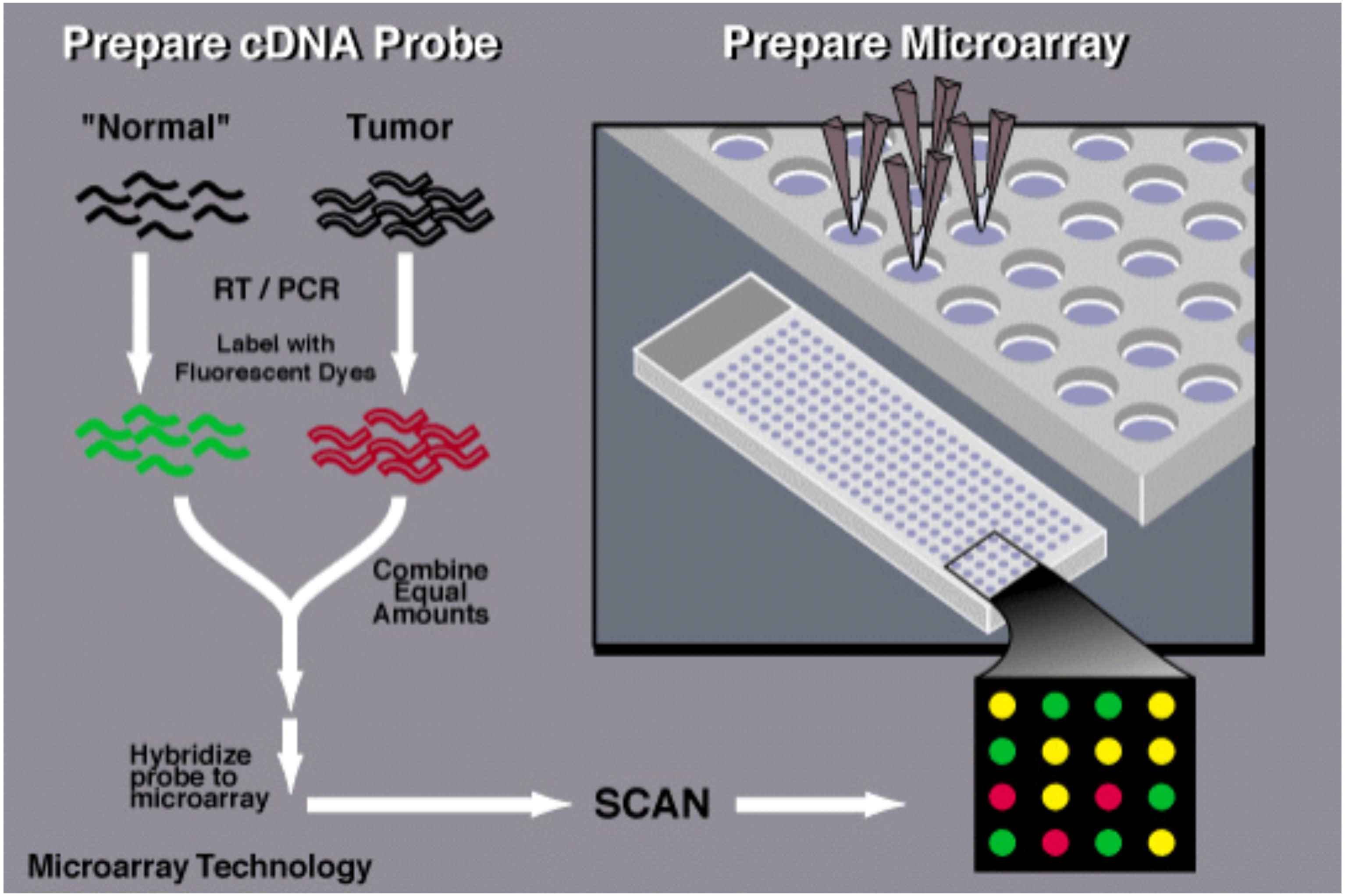
DNA Microarrays*

- Principle is to analyze gene (mRNA) or protein expression through large scale non-radioactive Northern (RNA) or Southern (DNA) hybridization analysis
- Essentially high throughput Northern Blotting method that uses Cy3 and Cy5 fluorescence for detection
- Allows expressional analysis of up to 20,000 genes simultaneously

Four Types of Microarrays*

- Photolithographically prepared short oligo (20-25 bp) arrays (1 colour)
- Spotted glass slide cDNA (500-1000 bp) arrays (2 colour)
- Spotted nylon cDNA (500-1000 bp) arrays (1 colour/radioactive)
- Spotted glass slide oligo (30-70 bp) arrays (1 or 2 colour)

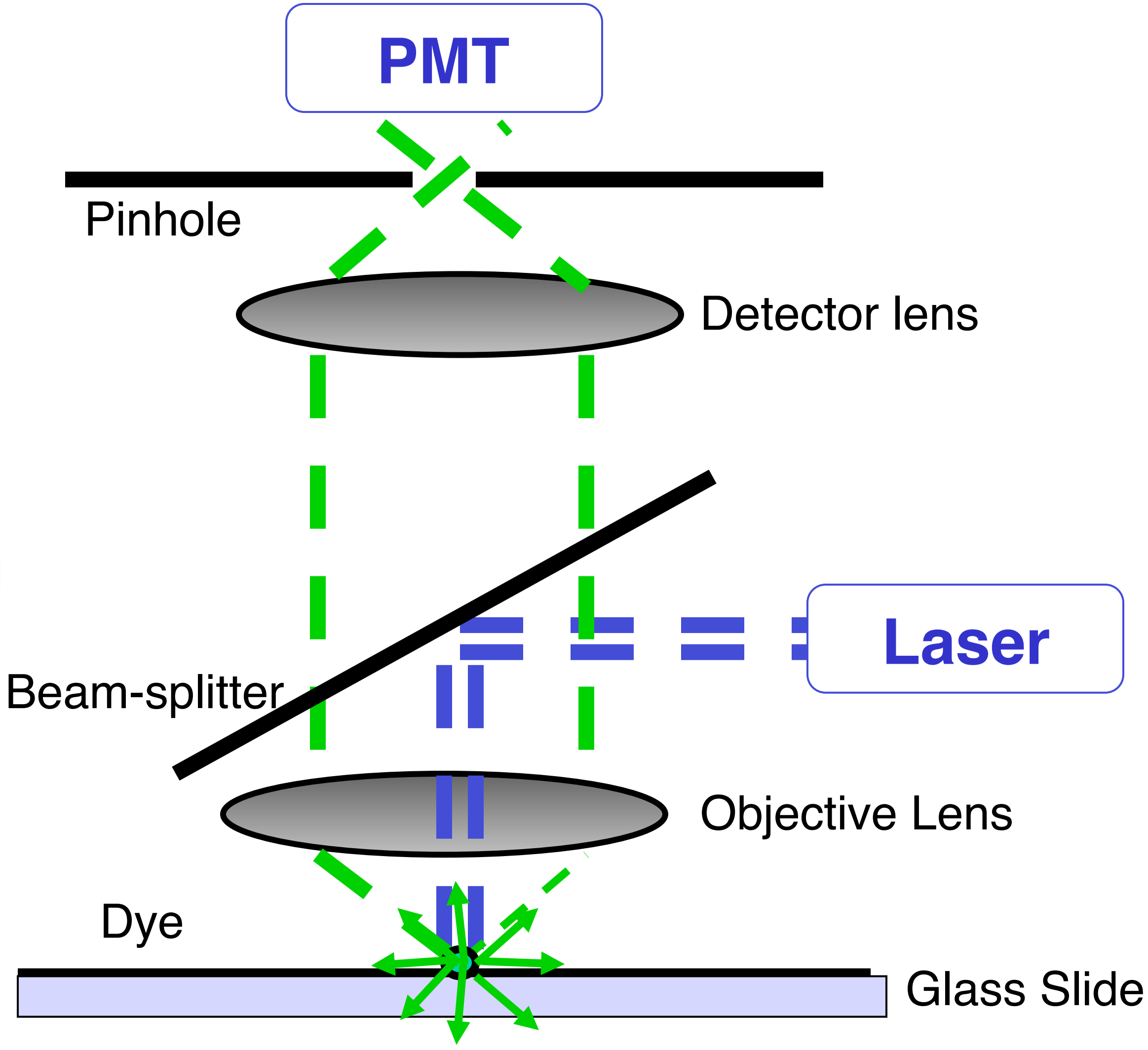
Principles of 2 Colour Microarrays*



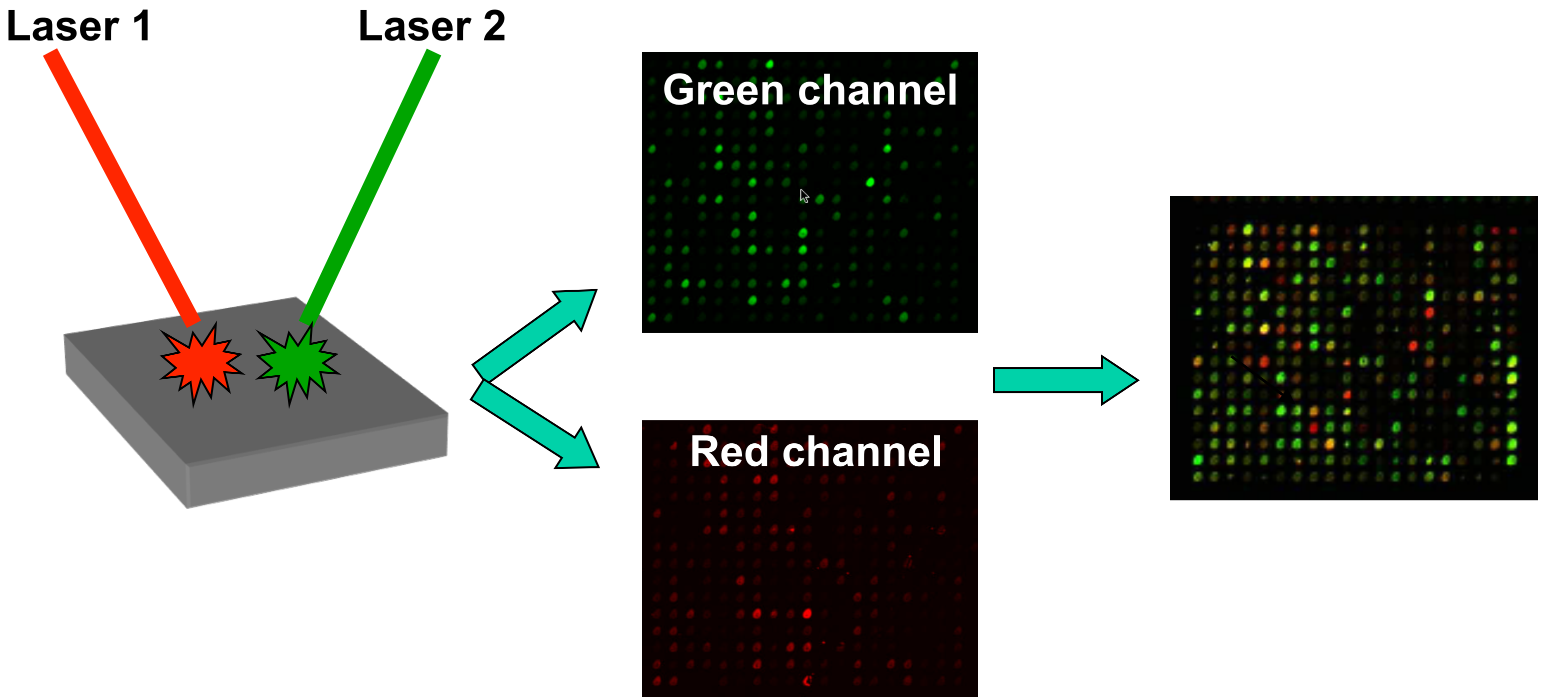
Microarray Definition of Probe and Target

- There are two acceptable and completely opposite definitions. We will use:
- Target = the DNA that is spotted on the array
- Probe = the DNA that is labeled with the fluorescent probe

Microarray Scanning



2-Colour Microarray Principles*



Laser 1

Laser 2

Green channel

Red channel

Scan and detect with confocal laser system

overlay images and normalize

Image process and analyze

Typical 2-Colour Data

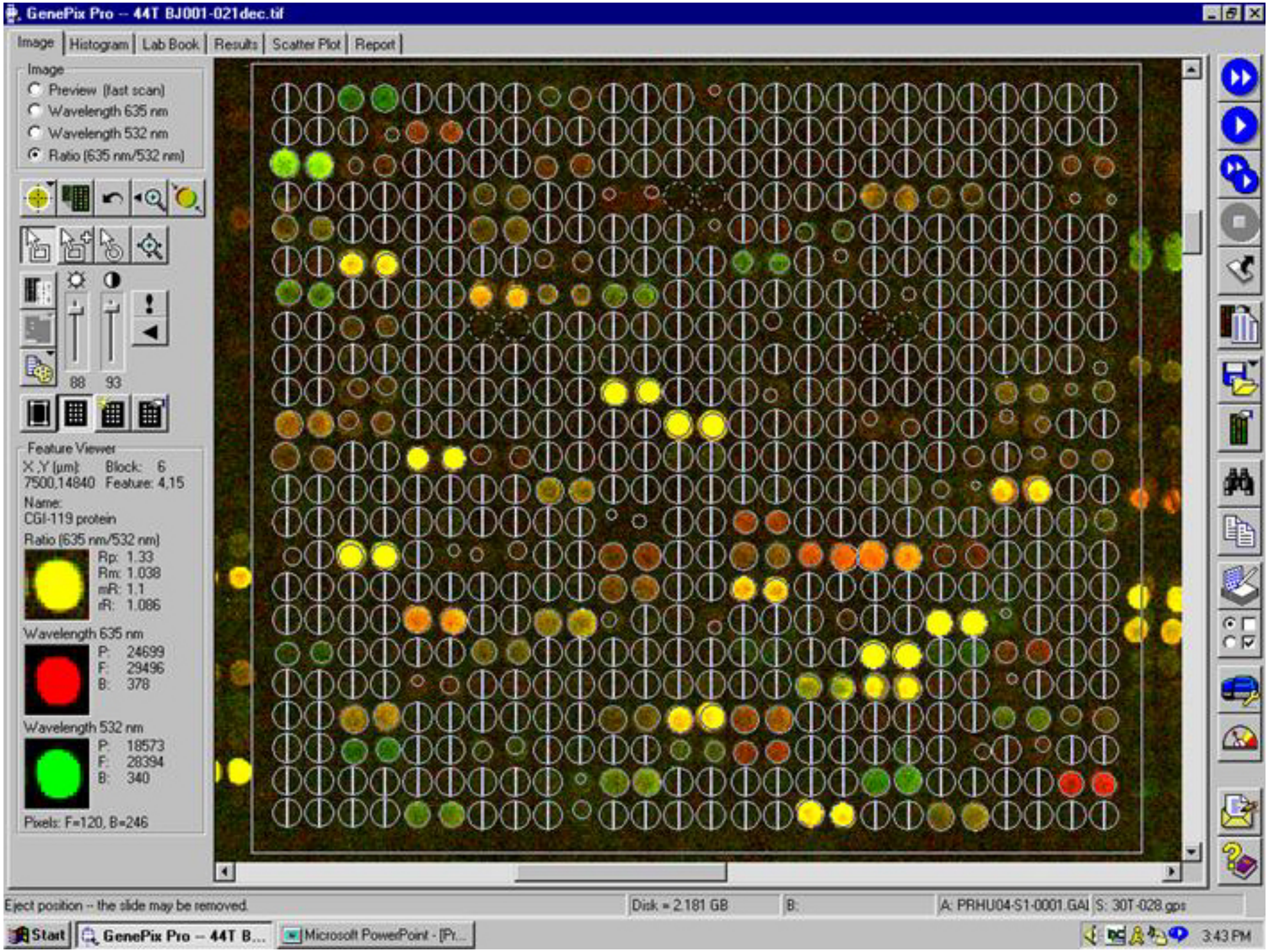


Image
○ Preview (fast scan)
○ Wavelength 635 nm
○ Wavelength 532 nm
● Ratio (635 nm/532 nm)

88 93

Feature Viewer
X,Y (µm): 7500,14840
Block: 6
Feature: 4,15
Name: CGI-119 protein

Ratio (635 nm/532 nm)
Rp: 1.33
Rm: 1.038
mR: 1.1
rR: 1.086

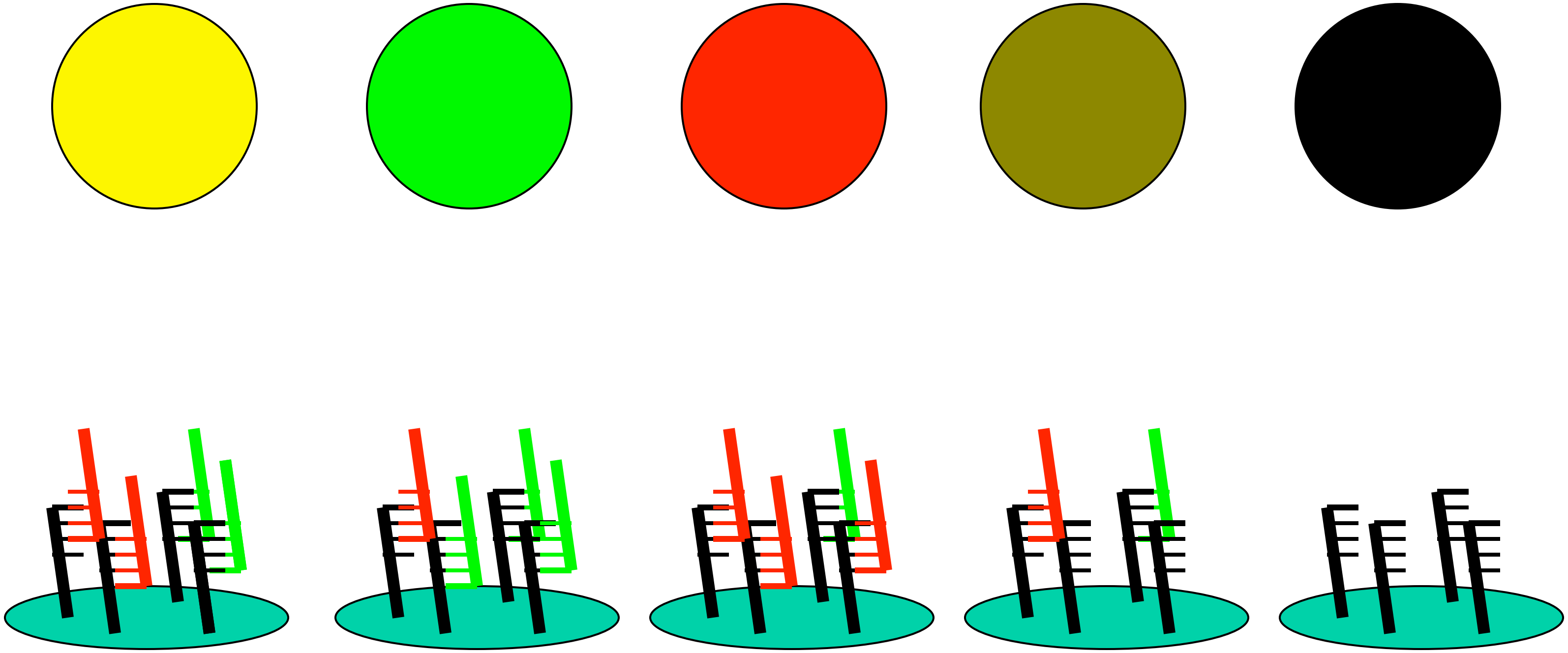
Wavelength 635 nm
P: 24699
F: 29496
B: 378

Wavelength 532 nm
P: 18573
F: 28394
B: 340

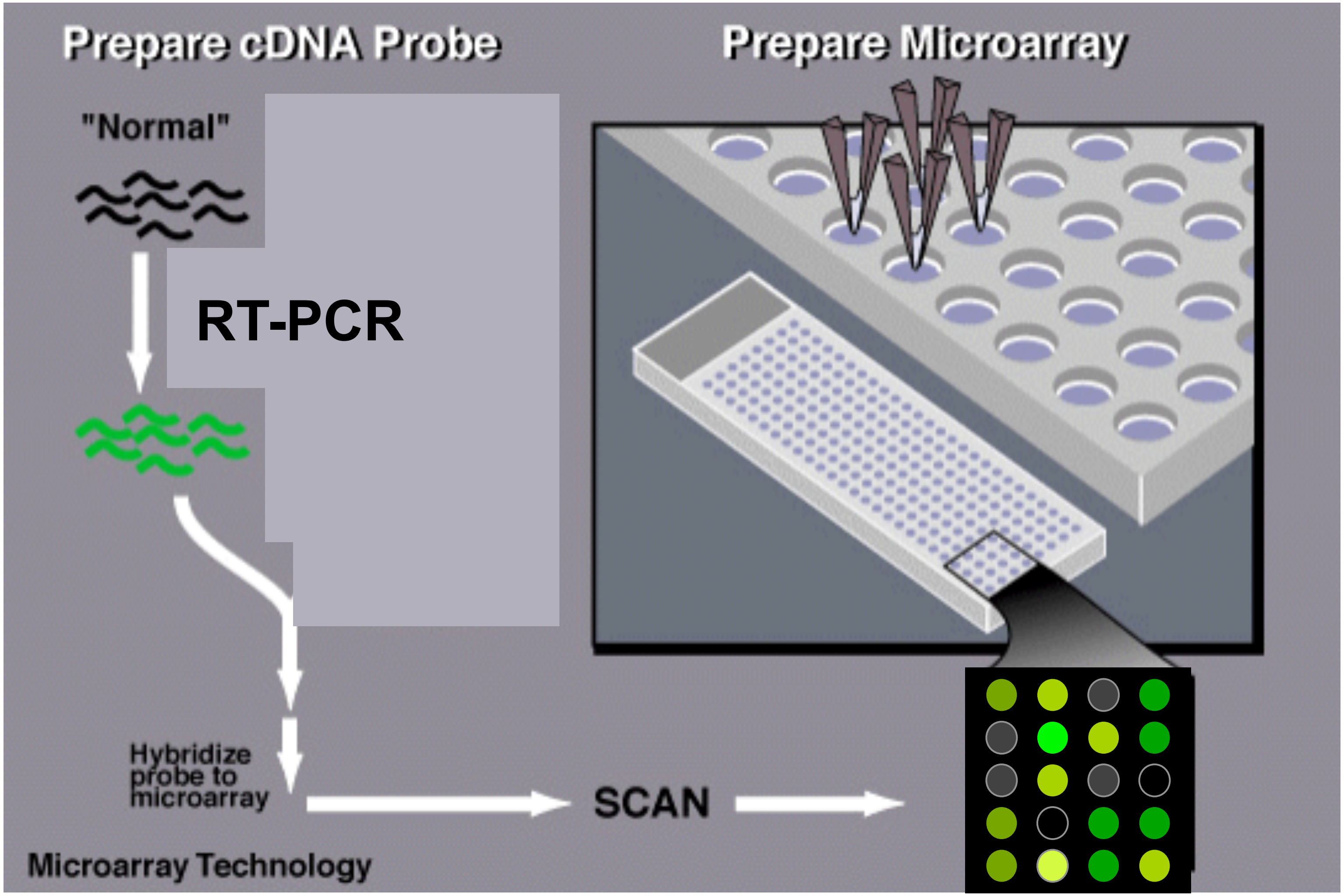
Pixels: F=120, B=246

Eject position -- the slide may be removed.
Disk = 2.181 GB
B:
A: PRHU04-S1-0001.GAL
S: 30T-028.gps

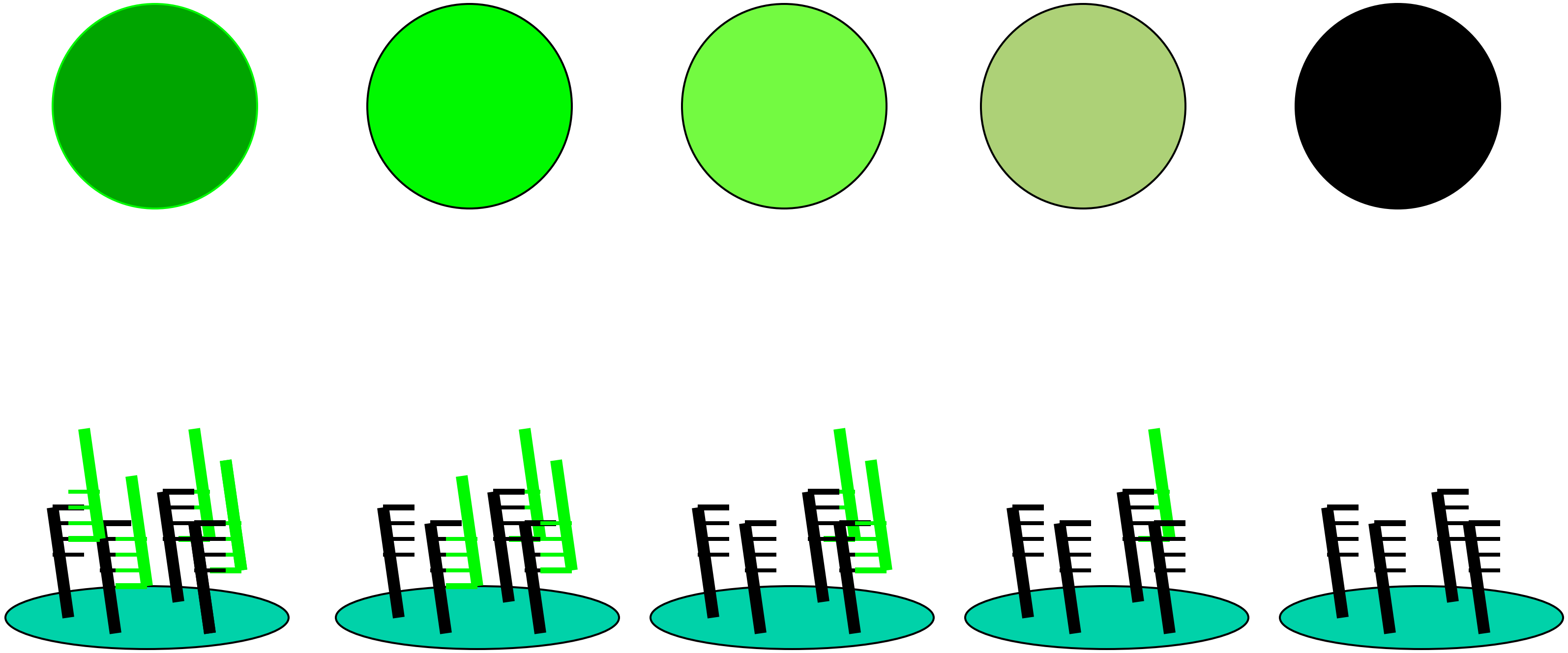
Microarrays & Spot Colour*



Principles of 1 Colour Microarrays



Microarrays & Spot Colour*



Two Colour vs. One Colour

- Two-colour hybridization eliminates artifacts due to variation in:
  - quantity of DNA spotted
  - stringency of hybridization
  - local concentration of label
- However,
  - both samples *must* label with equivalent efficiency
  - Information is lost for genes not expressed in the reference or control sample

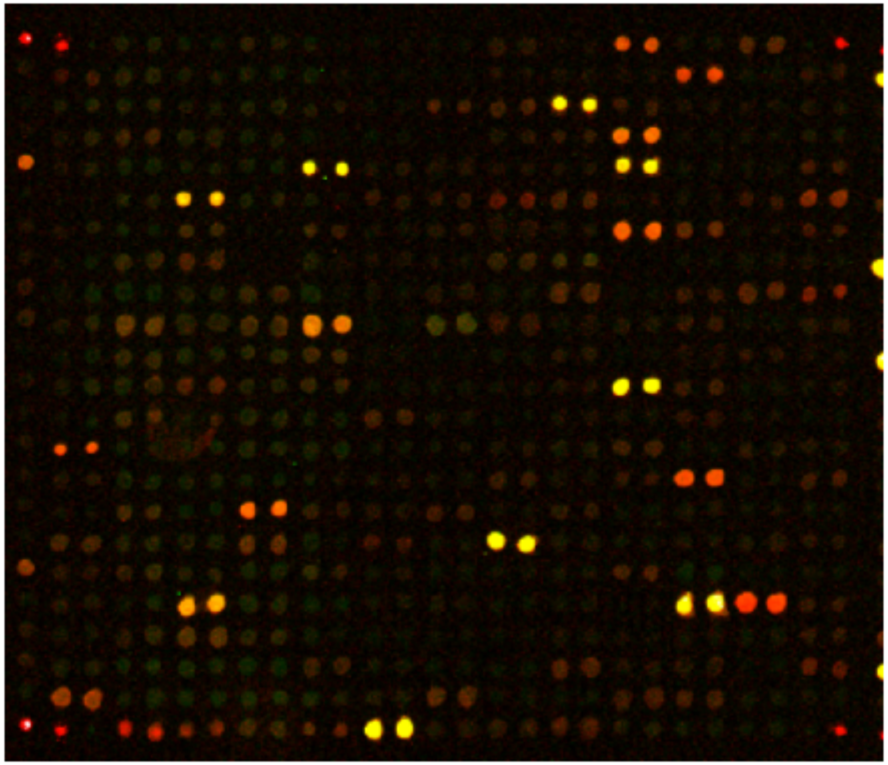
Two Colour vs. One Colour

- One-colour hybridization may have artifacts due to variation in:
  - quantity of DNA spotted
  - stringency of hybridization
  - local concentration of label
- However good quality control (QC) means,
  - fewer artifacts
  - less manipulation, lower cost
  - reduced loss of information (due to reference sample transcript content)

Specific Arrays of Interest

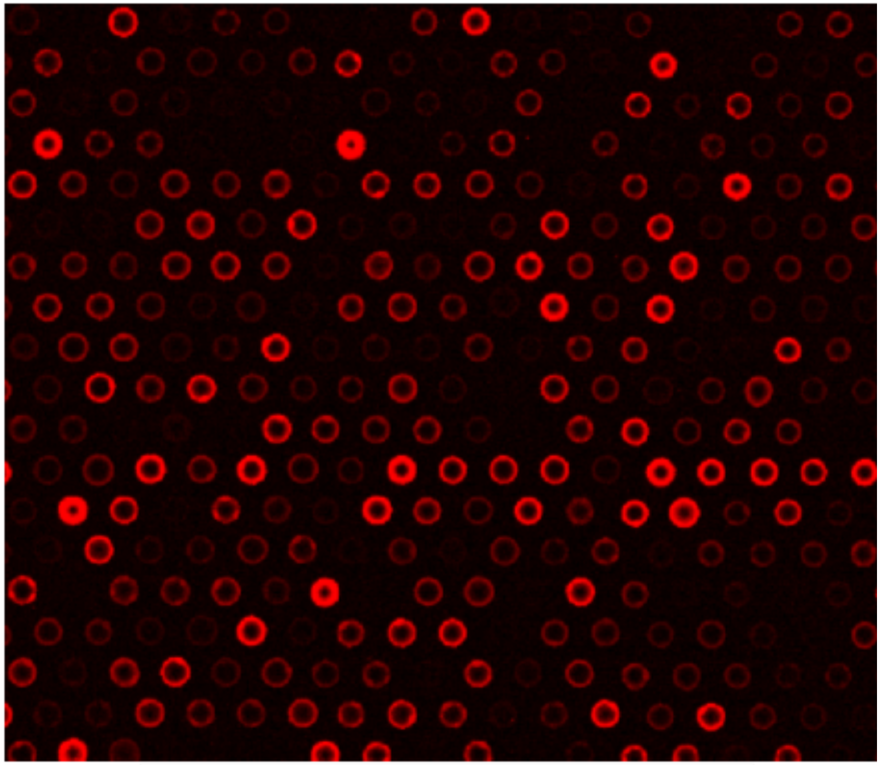
- Home-made Spotted Oligo Arrays
  - Made using glass slides, Operon oligos and robotic spotting equipment
- Applied Microarrays CodeLink Arrays
  - Made using specially treated slides, QC’d oligos and robotic spotting equipment
- Affymetrix Gene Chips
  - Made using photolithographically produced systems with multi-copy oligos

Array Images*



Oligo Microarray

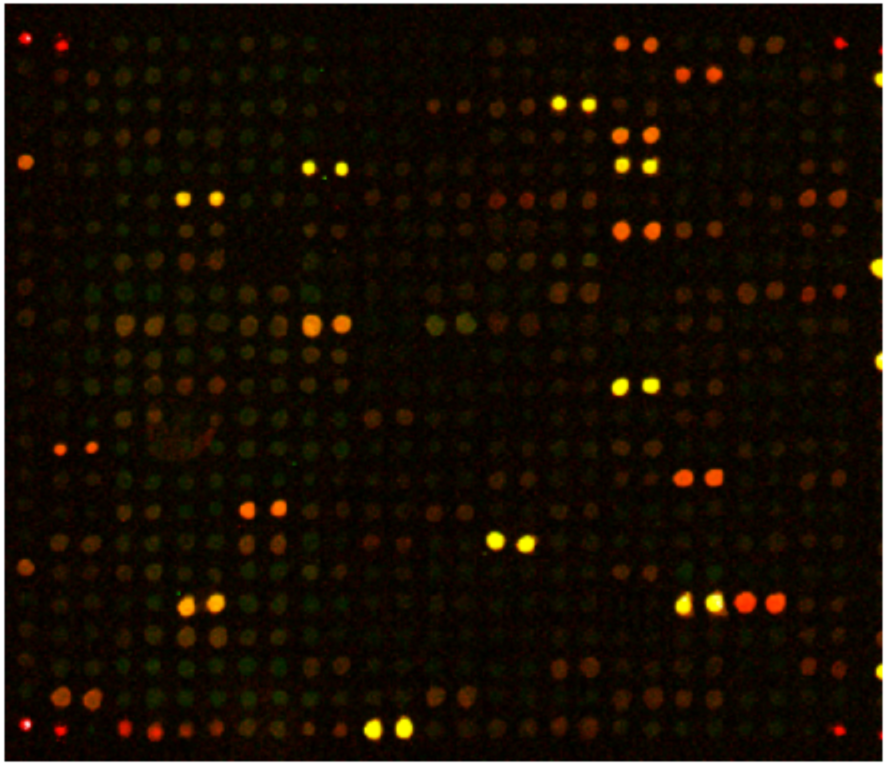
2 colour



Applied Microarrays

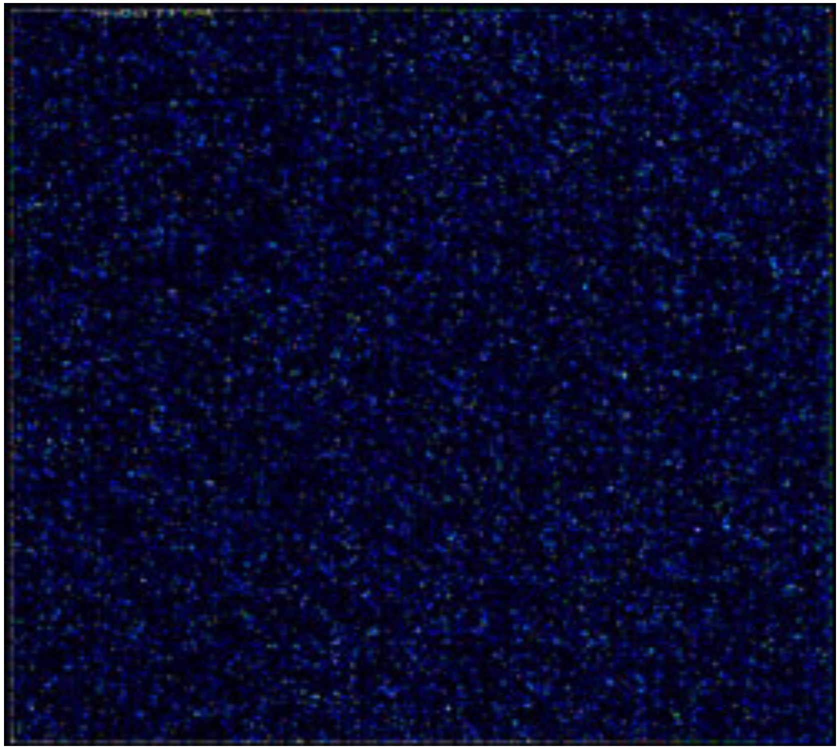
1 colour

Array Images*



Oligo Microarray

2 colour



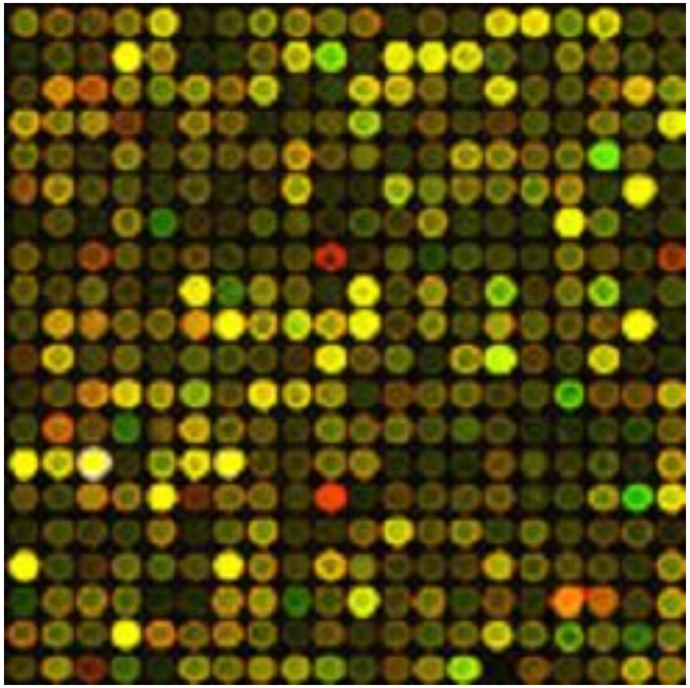
Affymetrix Gene Chip

1 colour

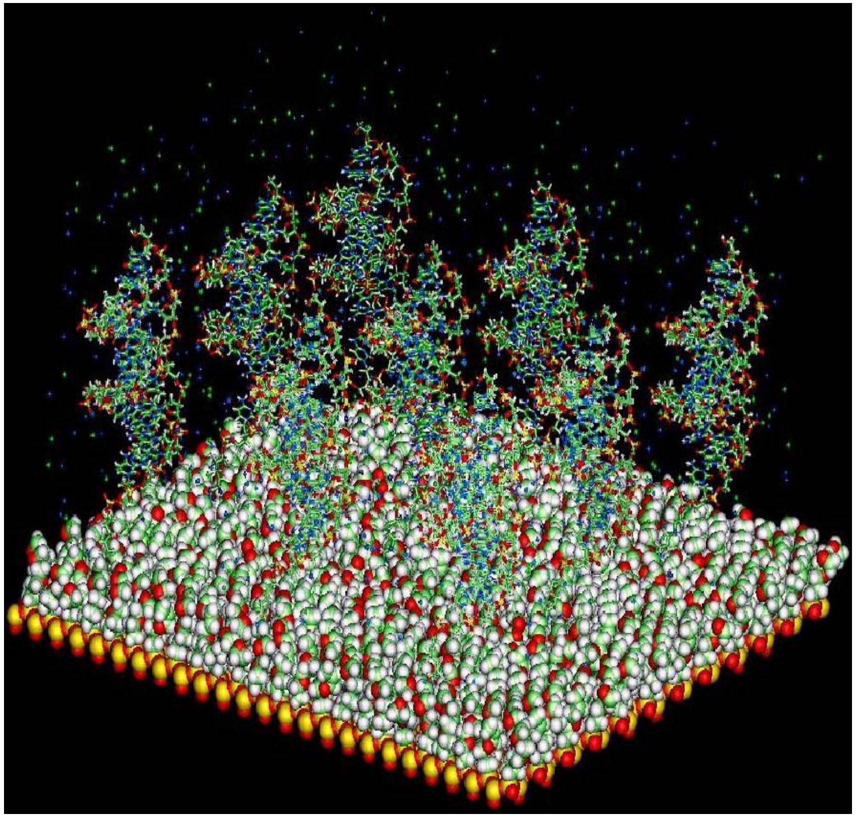
Home-made Spotted Arrays

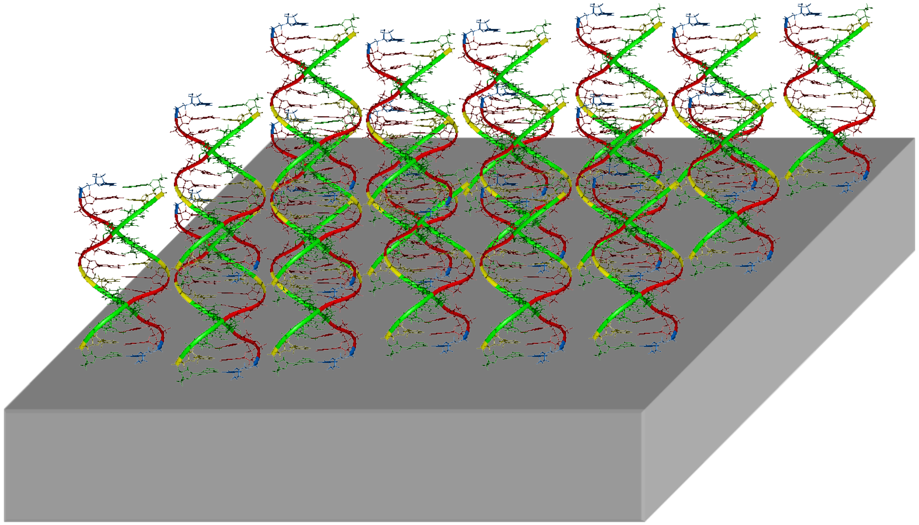
Spotted Microarrays*

- Target spots are >100μm and are usually deposited on glass
- Targets can be:
  - oligos (usually >40mers)
  - PCR fragments from cDNA/EST or genomic templates (rarely done)
- Not reused; 2-colour hybridizations

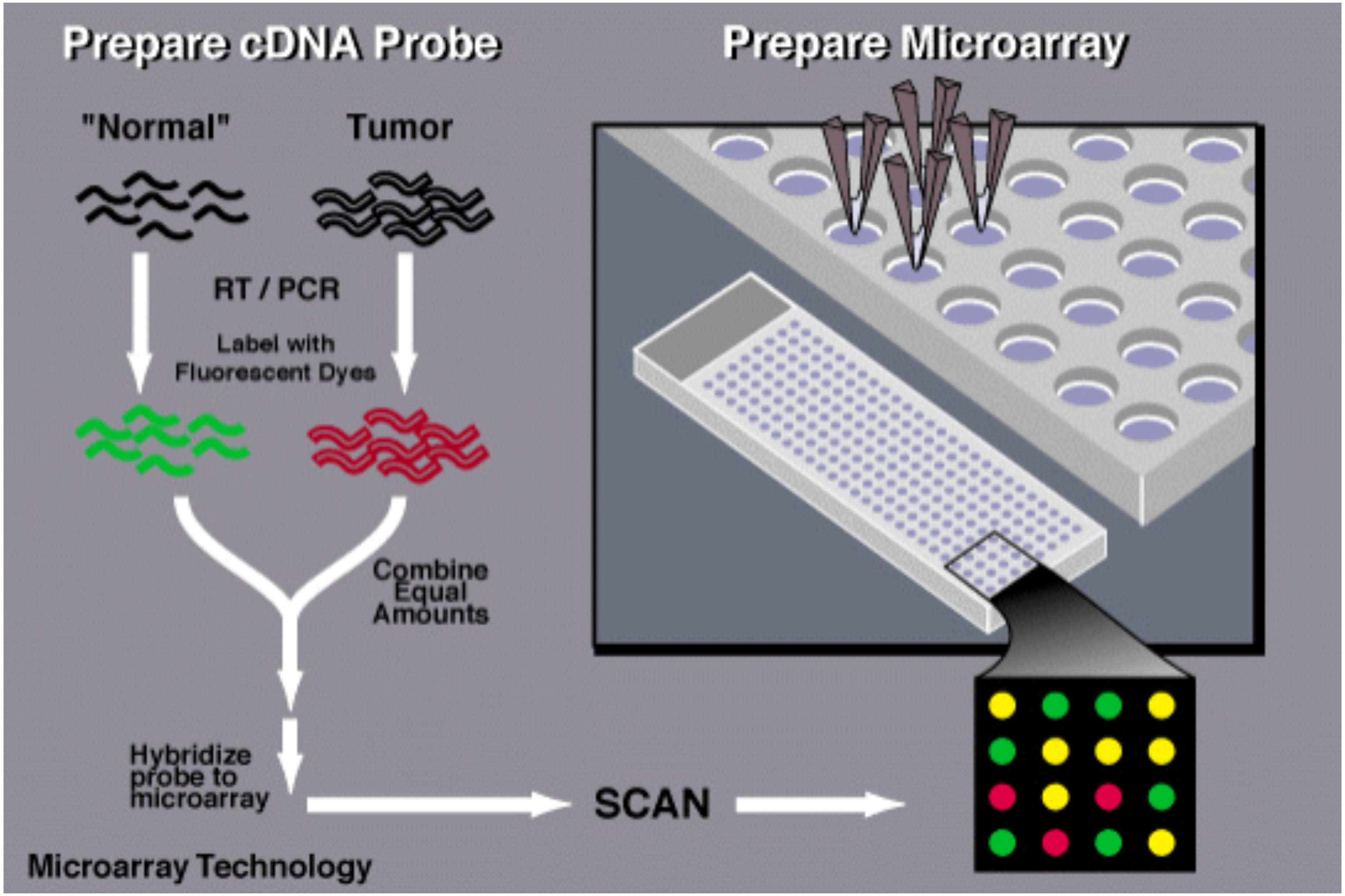


Standard Spotted Array

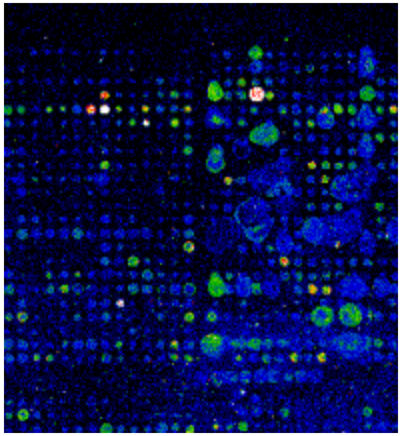




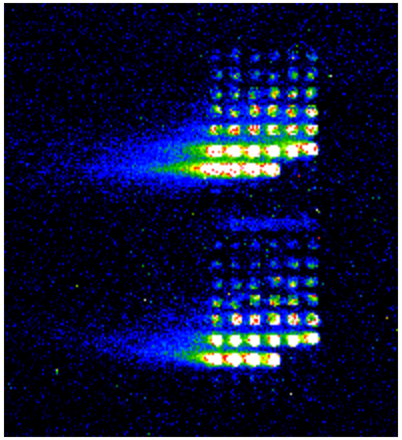
Home-made Microarrays



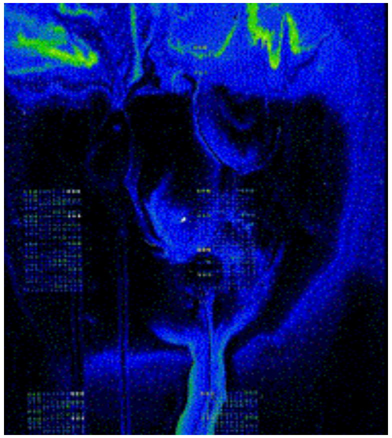
Common Home-made Microarray Errors*



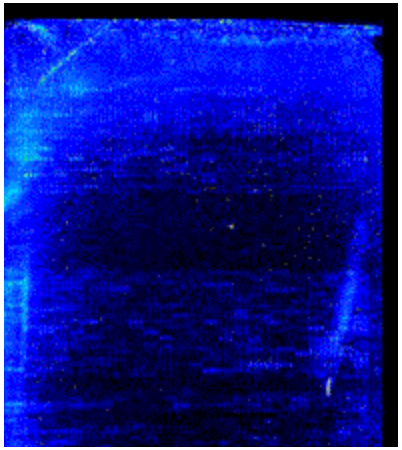
Irregular Spot



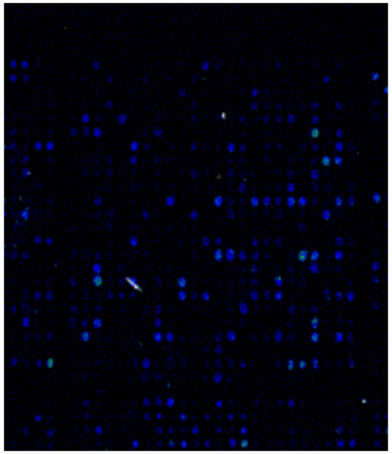
Comet Tail



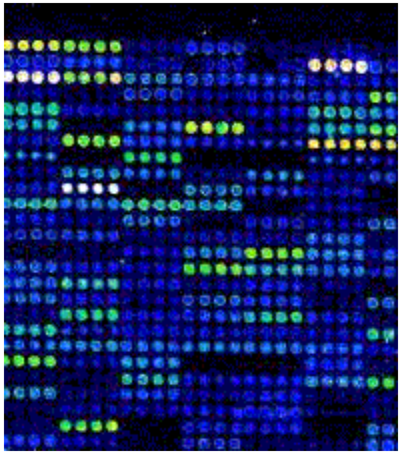
Streaking



Hi Background



Low Intensity

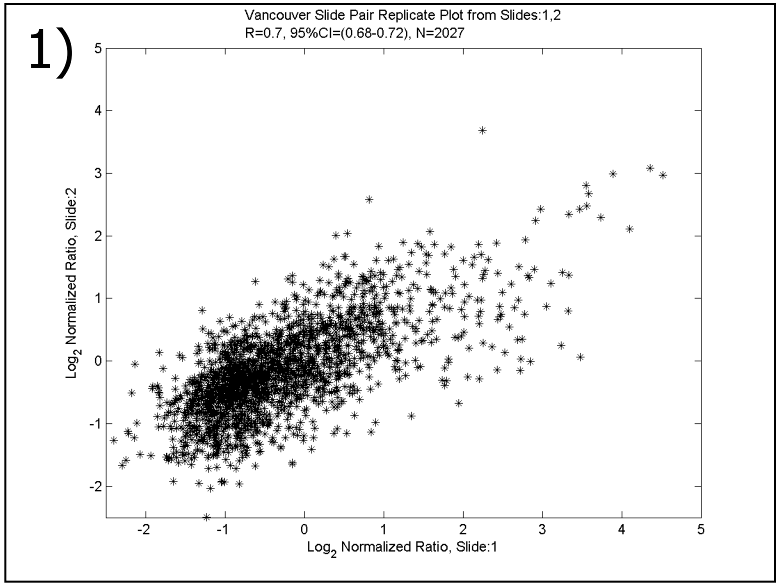


A Good Array

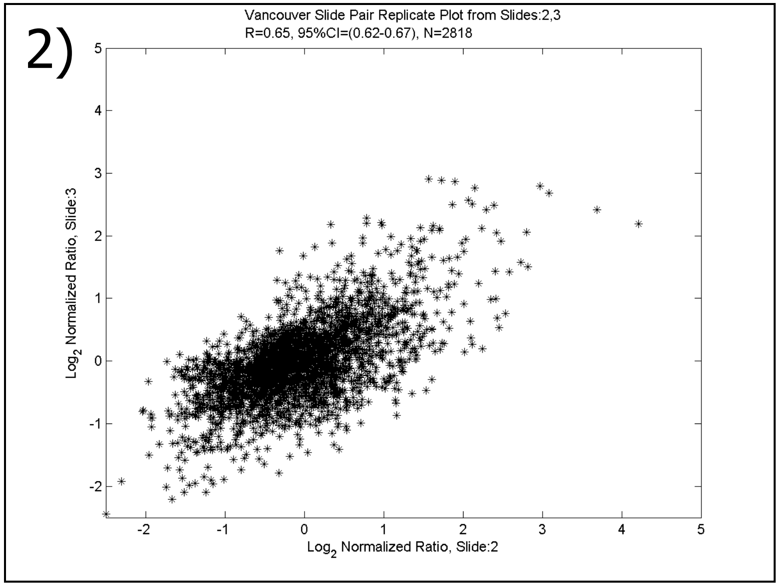
Testing Reproducibility

- Breast tumor tissue biopsy
- mRNA prepared using standard methods
- Control sample made from pooled mRNA from several cell types
- 3 RNA samples prepared from 1 tissue source – arrayed onto two sets of home-made chips from different suppliers
- Conducted pairwise comparison of intensity correlations & no. of spots

Home-made Arrays

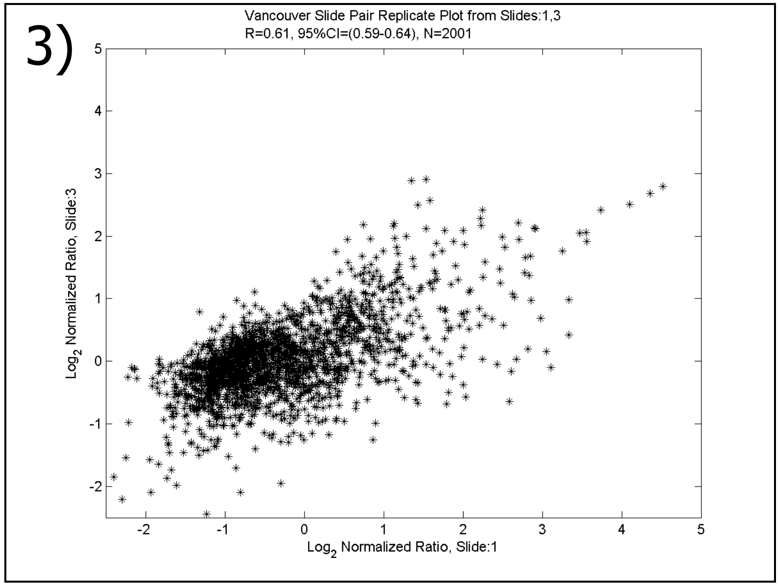

1)

Vancouver Slide Pair Replicate Plot from Slides:1,2
R=0.7, 95%CI=(0.68-0.72), N=2027


2)

Vancouver Slide Pair Replicate Plot from Slides:2,3
R=0.65, 95%CI=(0.62-0.67), N=2818

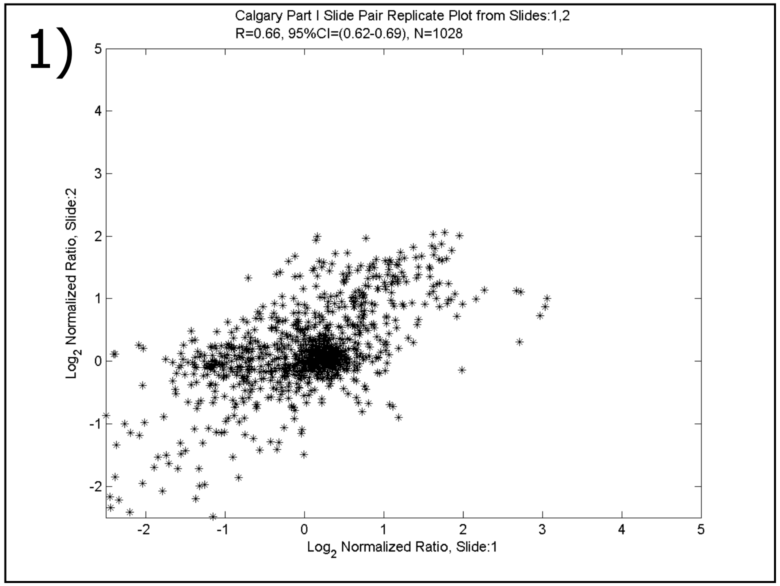
Oligo Microarray 1


3)

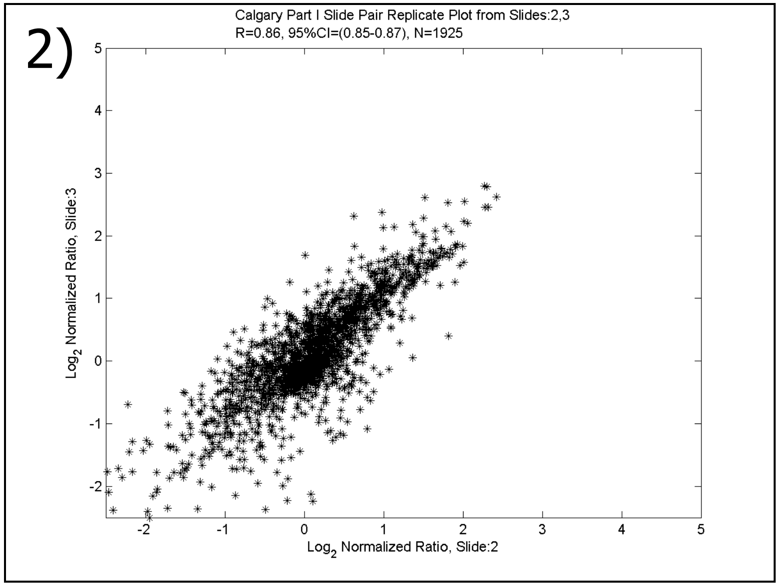
Vancouver Slide Pair Replicate Plot from Slides:1,3
R=0.61, 95%CI=(0.59-0.64), N=2001

1)	R=0.7	95%CI=(0.68-0.72)	N=2027
2)	R=0.65	95%CI=(0.62-0.67)	N=2818
3)	R=0.61	95%CI=(0.59-0.64)	N=2001

Home-made Arrays

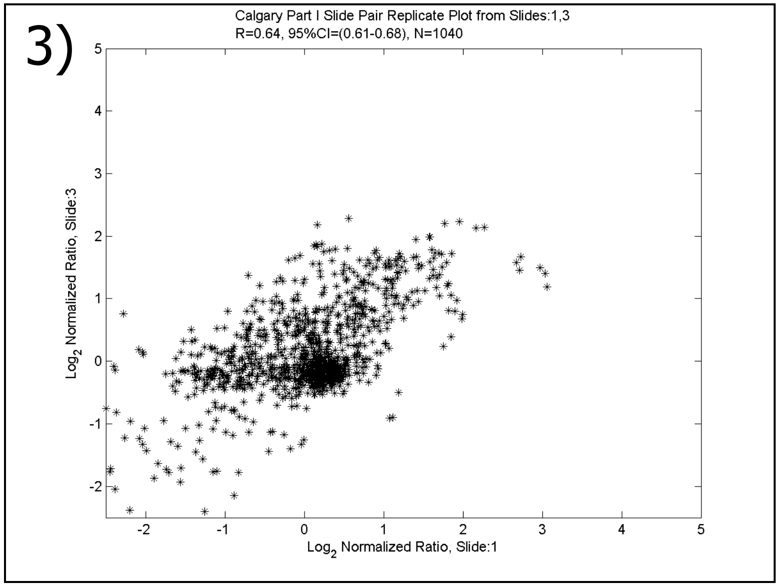

1)

Calgary Part I Slide Pair Replicate Plot from Slides:1,2
R=0.66, 95%CI=(0.62-0.69), N=1028


2)

Calgary Part I Slide Pair Replicate Plot from Slides:2,3
R=0.86, 95%CI=(0.85-0.87), N=1925

Oligo Microarray 2


3)

Calgary Part I Slide Pair Replicate Plot from Slides:1,3
R=0.64, 95%CI=(0.61-0.68), N=1040

1)	R=0.66	95%CI=(0.62-0.69)	N=1028
2)	R=0.86	95%CI=(0.85-0.87)	N=1925
3)	R=0.64	95%CI=(0.61-0.68)	N=1040

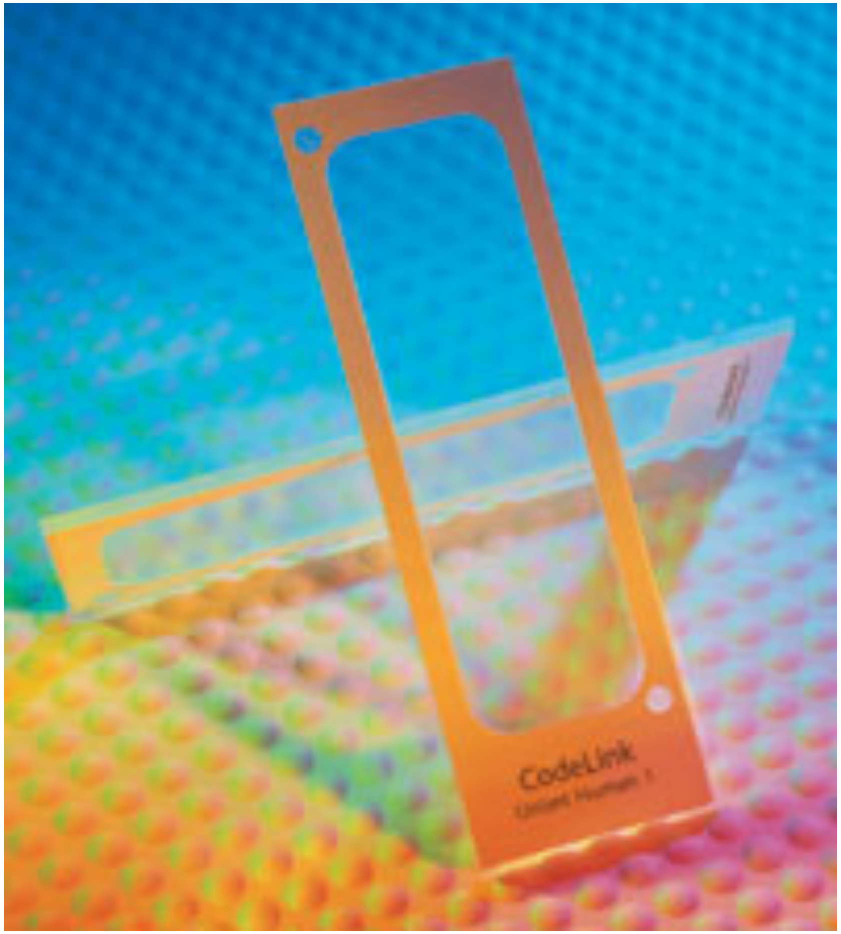
Advantages to Home-made Systems*

- Cheapest method to produce arrays ($100 to $300/slide)
- Allows lab full control over design and printing of arrays (customizable)
- Allows quick adaptation to new technologies, new target sets
- Allows more control over analysis

Disadvantages to Home-made Systems*

- Quality and quality-control of oligo target set is highly variable
- Quality of spotting and spot geometry is highly variable
- Technology is very advanced, difficult and expensive to maintain (robotics)
- Reproducibility is poor

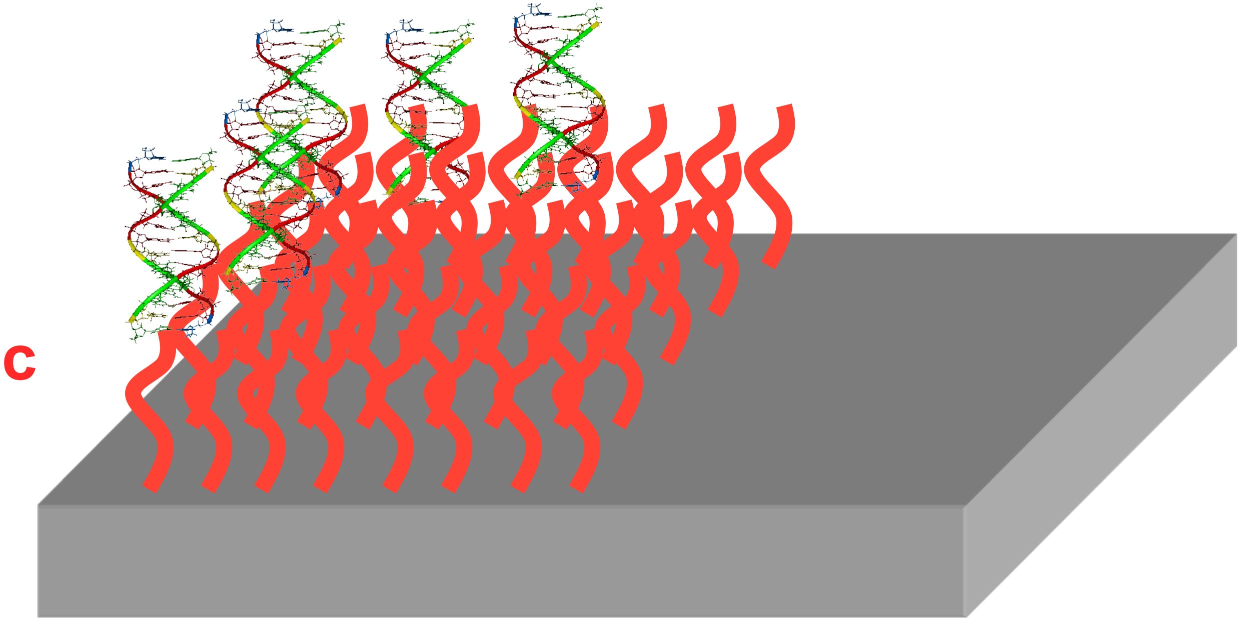
Applied Microarrays CodeLink Arrays



Applied Microarrays CodeLink Arrays

- Applied Microarrays synthesizes its 30-nucleotide oligos offline, tests them by mass spectrometry, deposits them on specially coated (polyacrylamide) array, and then assays them for quality control
- Uses a special Flex Chamber™—a disposable hybridization chamber already attached to the slide to improve hybridization consistency

Applied Microarrays CodeLink Oligo Chip



DNA

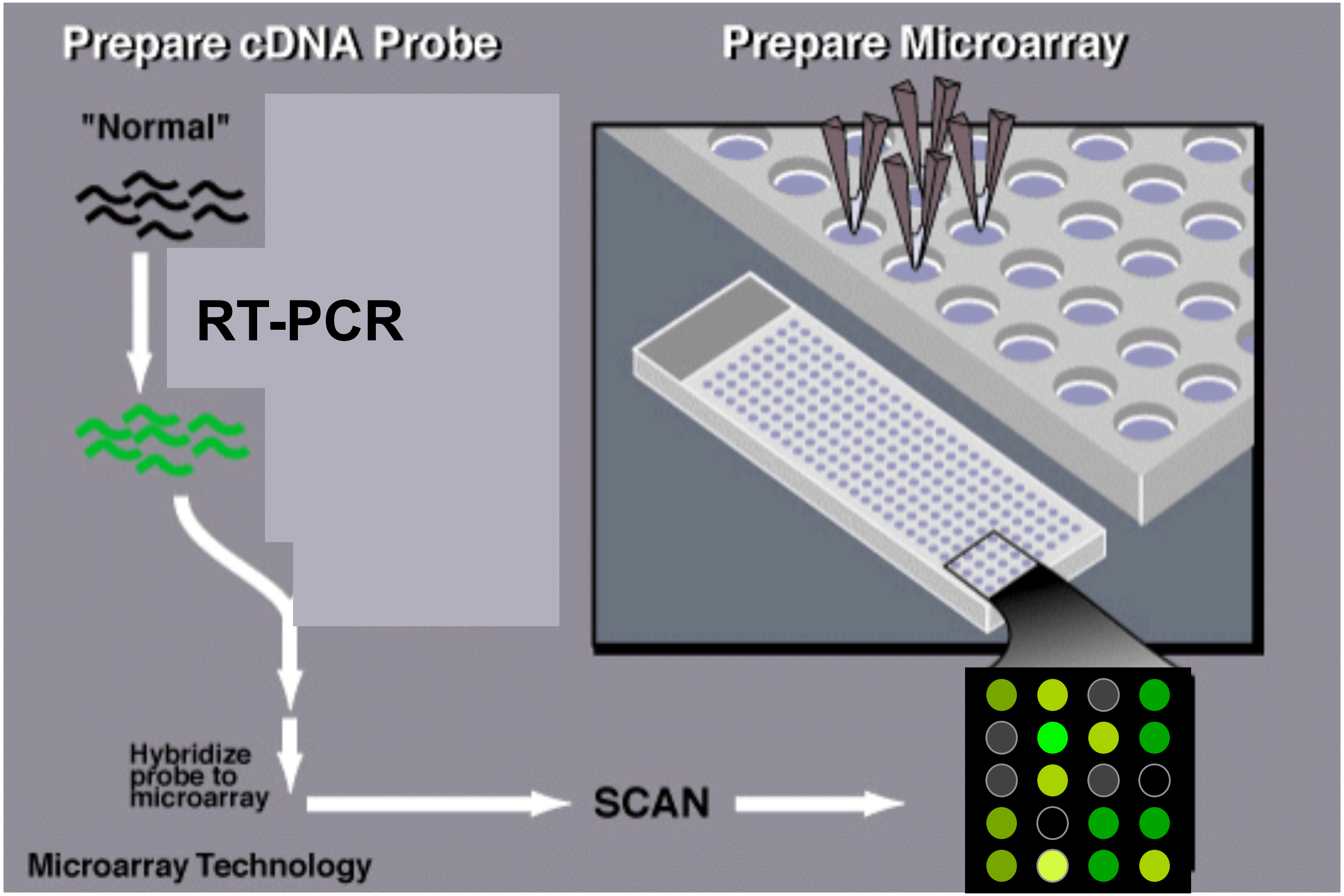
Hydrophilic polymer

Glass

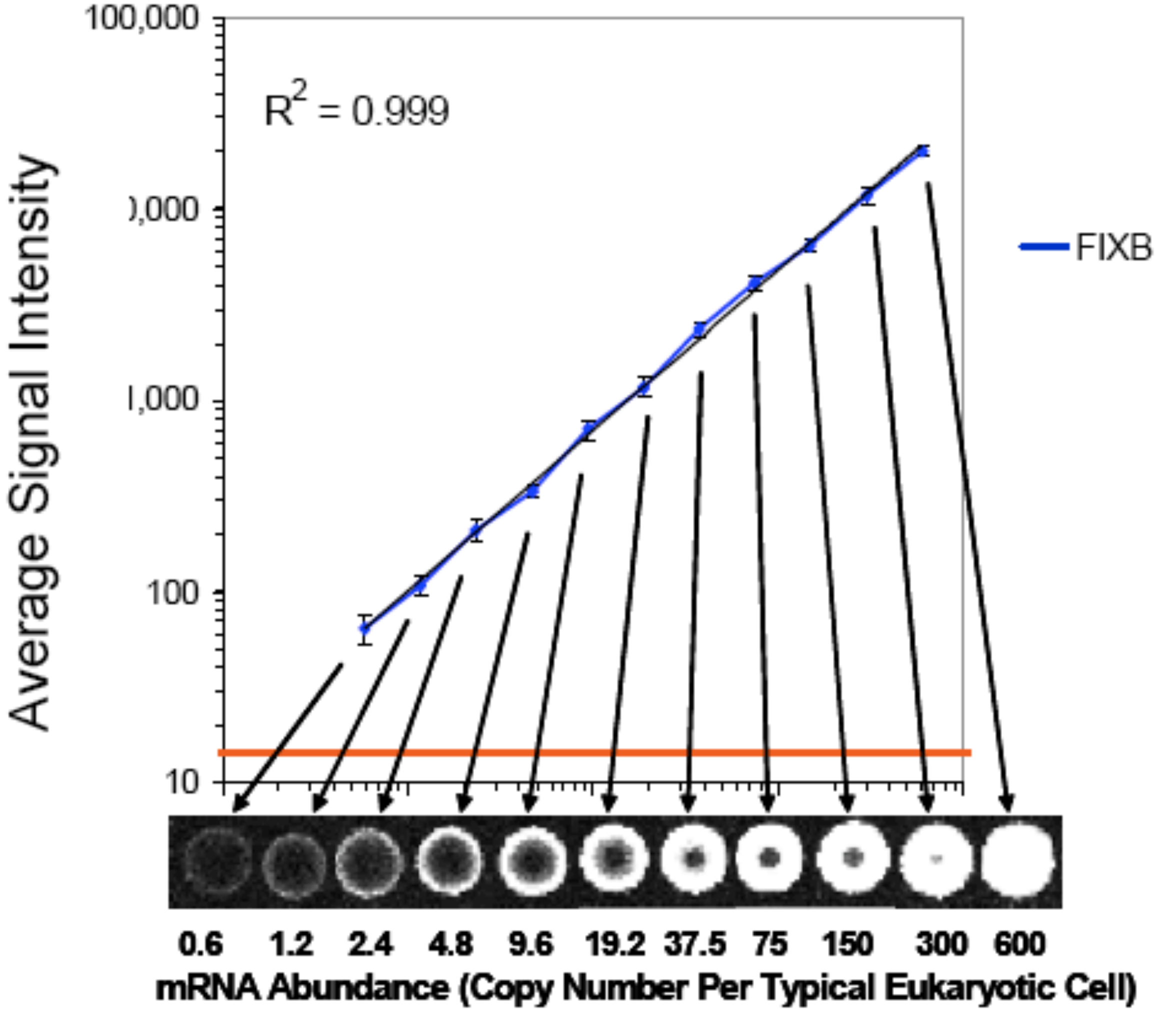
CodeLink Special Coating

- Most glass substrates are quite hydrophobic
- This hydrophobicity affects the local binding and surface chemistry of most glass-slide chips making most of the attached DNA oligo inaccessible
- Coating the slide with a hydrophilic polymer allows the cDNA to pair up with the substrate oligos much better

Applied Microarrays Array



Morphology Does Not Affect Dynamic Range
CodeLink Bioarrays Can Achieve Linearity Across 3 Logs*



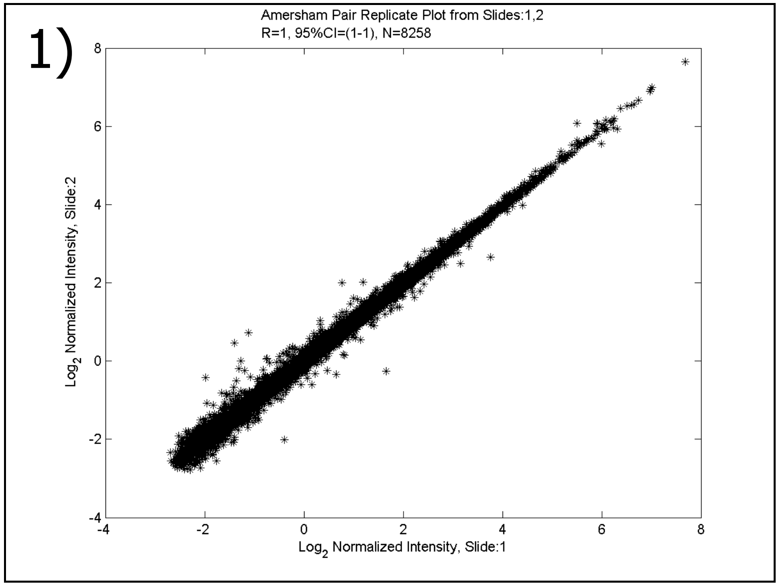
- The red line indicates the signal level for non-spiked target.
- Error bars represent one standard deviation for each mean (n=18) signal

*Data obtained from cRNA dilution series.

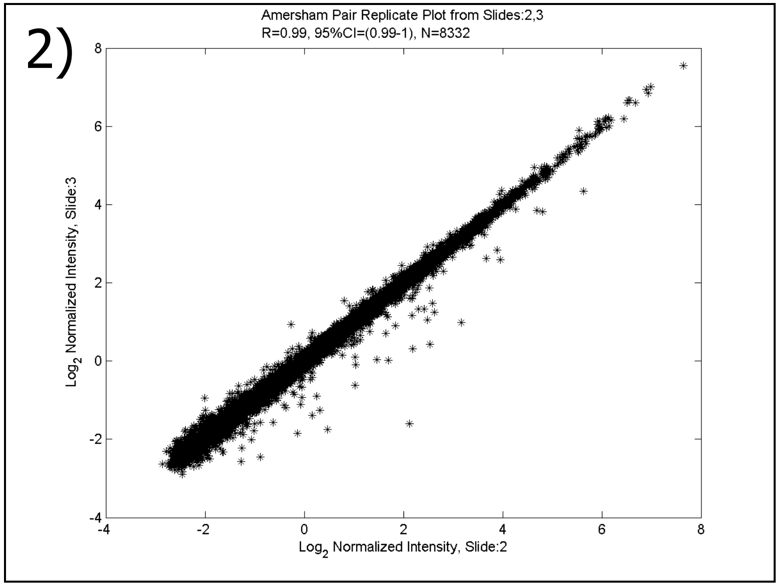
Testing Reproducibility

- Breast tumor tissue biopsy
- mRNA prepared using standard methods
- 3 RNA samples prepared from 1 tissue source – arrayed onto 3 different sets of CodeLink chips
- Conducted pairwise comparison of intensity correlations, intensity ratio correlations & number of “passed” spots

Intensity, Pairwise Comparisons



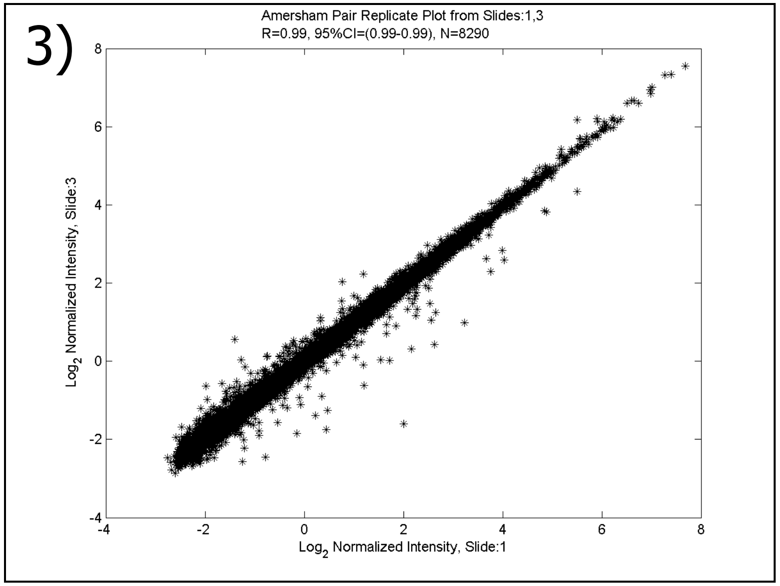
Amersham Pair Replicate Plot from Slides:1,2
R=1, 95%CI=(1-1), N=8258



Amersham Pair Replicate Plot from Slides:2,3
R=0.99, 95%CI=(0.99-1), N=8332

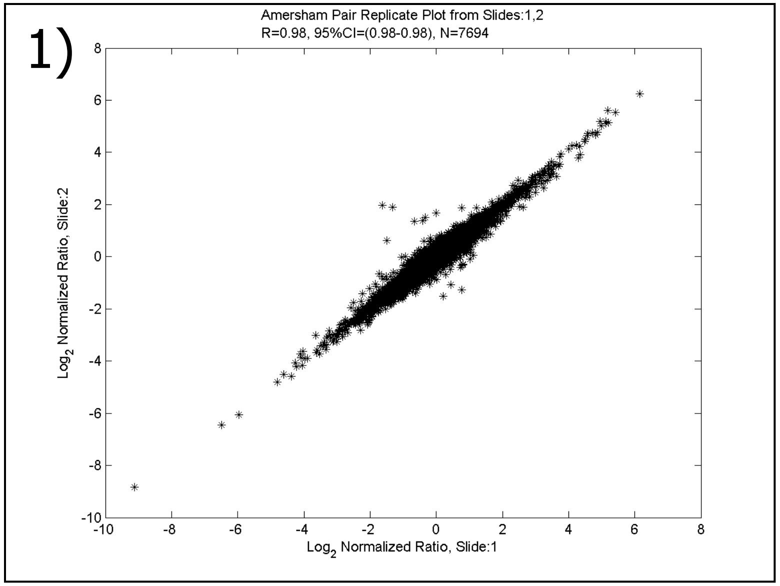
Applied Microarrays Slides

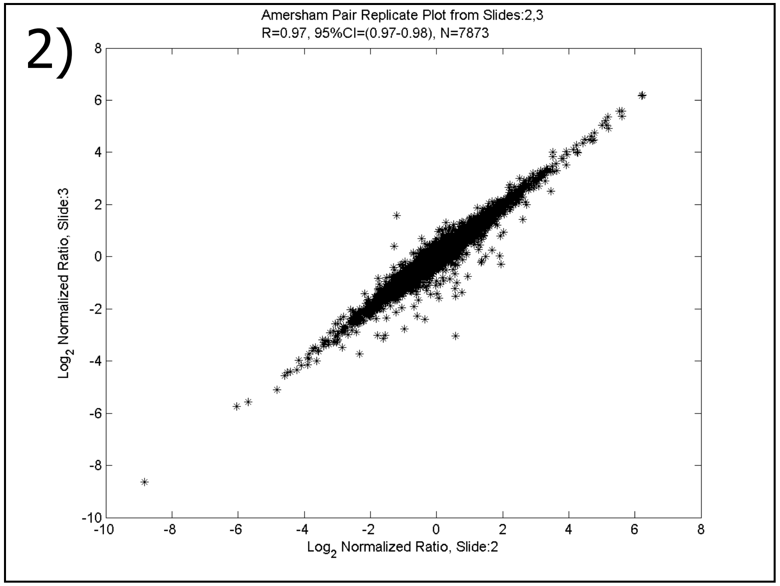
1)	R=1	95%CI=(1-1)	N=8258
2)	R=0.99	95%CI=(0.99-1)	N=8332
3)	R=0.99	95%CI=(0.99-0.99)	N=8290



Amersham Pair Replicate Plot from Slides:1,3
R=0.99, 95%CI=(0.99-0.99), N=8290

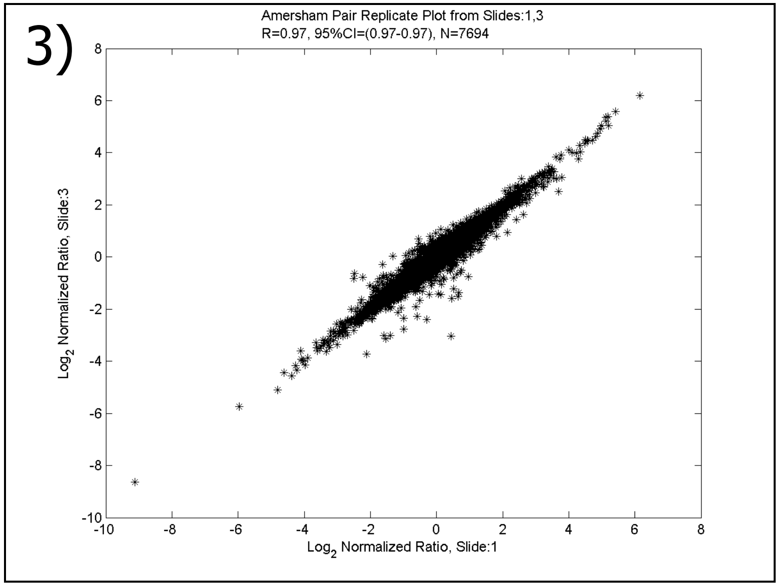
Ratio, Pairwise Comparisons





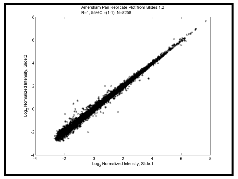
Applied Microarrays Slides

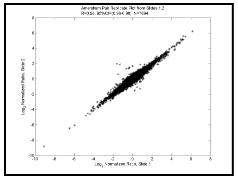
1) R=0.98 95%CI=(0.98-0.98) N=7694
2) R=0.97 95%CI=(0.97-0.98) N=7873
3) R=0.97 95%CI=(0.97-0.97) N=7694

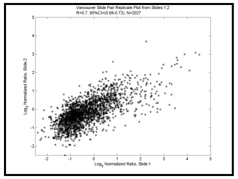


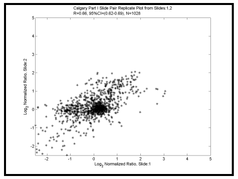
General Comparison

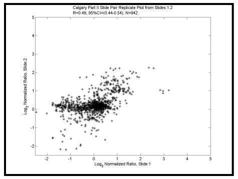
Appl Micro Intensity	1)	R=1	95%CI=(1-1)	N=8258
	2)	R=0.99	95%CI=(0.99-1)	N=8332
	3)	R=0.99	95%CI=(0.99-0.99)	N=8290
Appl Micro Ratio	1)	R=0.98	95%CI=(0.98-0.98)	N=7694
	2)	R=0.97	95%CI=(0.97-0.98)	N=7873
	3)	R=0.97	95%CI=(0.97-0.97)	N=7694
Vancouver	1)	R=0.7	95%CI=(0.68-0.72)	N=2027
	2)	R=0.65	95%CI=(0.62-0.67)	N=2818
	3)	R=0.61	95%CI=(0.59-0.64)	N=2001
Calgary I	1)	R=0.66	95%CI=(0.62-0.69)	N=1028
	2)	R=0.86	95%CI=(0.85-0.87)	N=1925
	3)	R=0.64	95%CI=(0.61-0.68)	N=1040
Calgary II	1)	R=0.49	95%CI=(0.44-0.54)	N=942
	2)	R=0.81	95%CI=(0.8-0.83)	N=1700
	3)	R=0.57	95%CI=(0.52-0.61)	N=973











Comparative Accuracy

	RT-PCR	Spotted Array	CodeLink
GENES	Expression Pattern TaqMan	Expression Pattern Operon	Expression Pattern Applied Micr
hENT1	+	-	+
hENT2	+	-	+
hCNT1	-	-	-
hCNT2	-	+	-
dck	+	-	+
ER	+	-	+

CodeLink Advantages*

- Exceptional reproducibility because of:
  - careful target design
  - QC of oligo preparations and spotting
  - high proportion of oligo binding to cDNA substrate due to hydrophilic coating
  - well controlled/uniform hybridization
- Allows users to continue using same scanners/software as in spotted arrays

CodeLink Disadvantages*

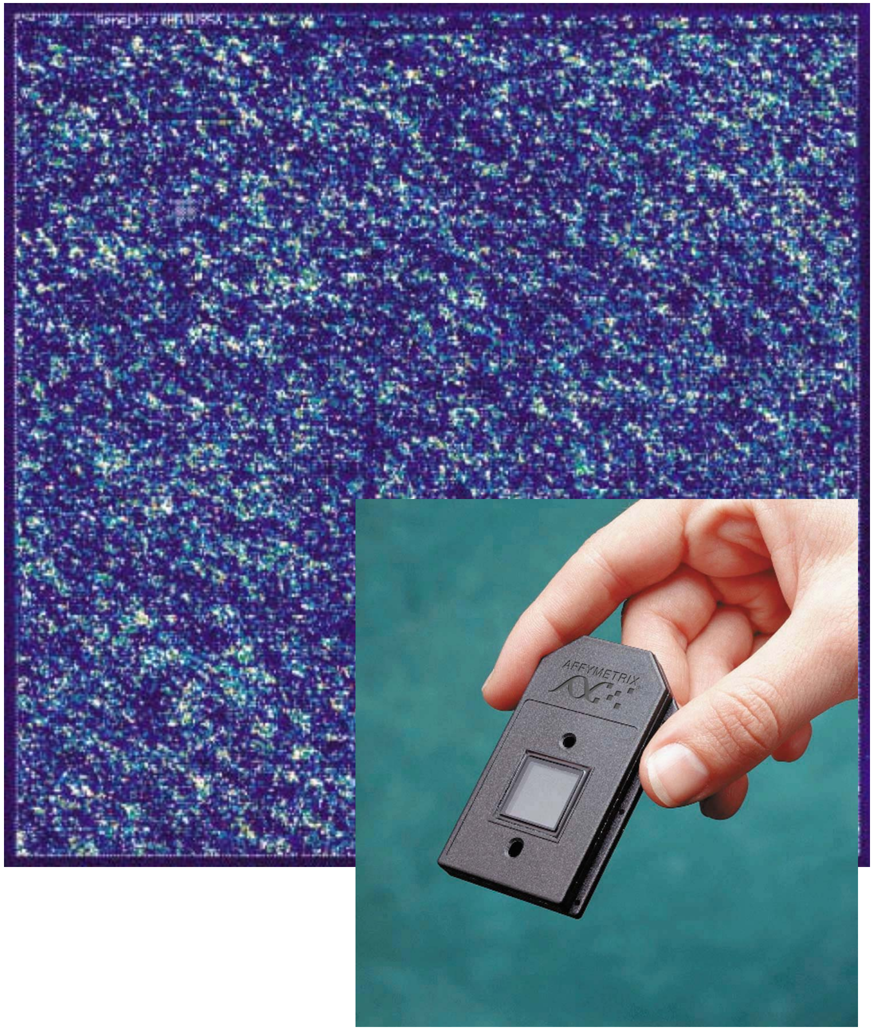
- Lack of flexibility or customizability (users depend on Applied Microarrays to provide & design chips)
- Dependent on proprietary kits and reagents
- More expensive than spotted arrays ($700/chip)

Cost per Sample in Triplicate

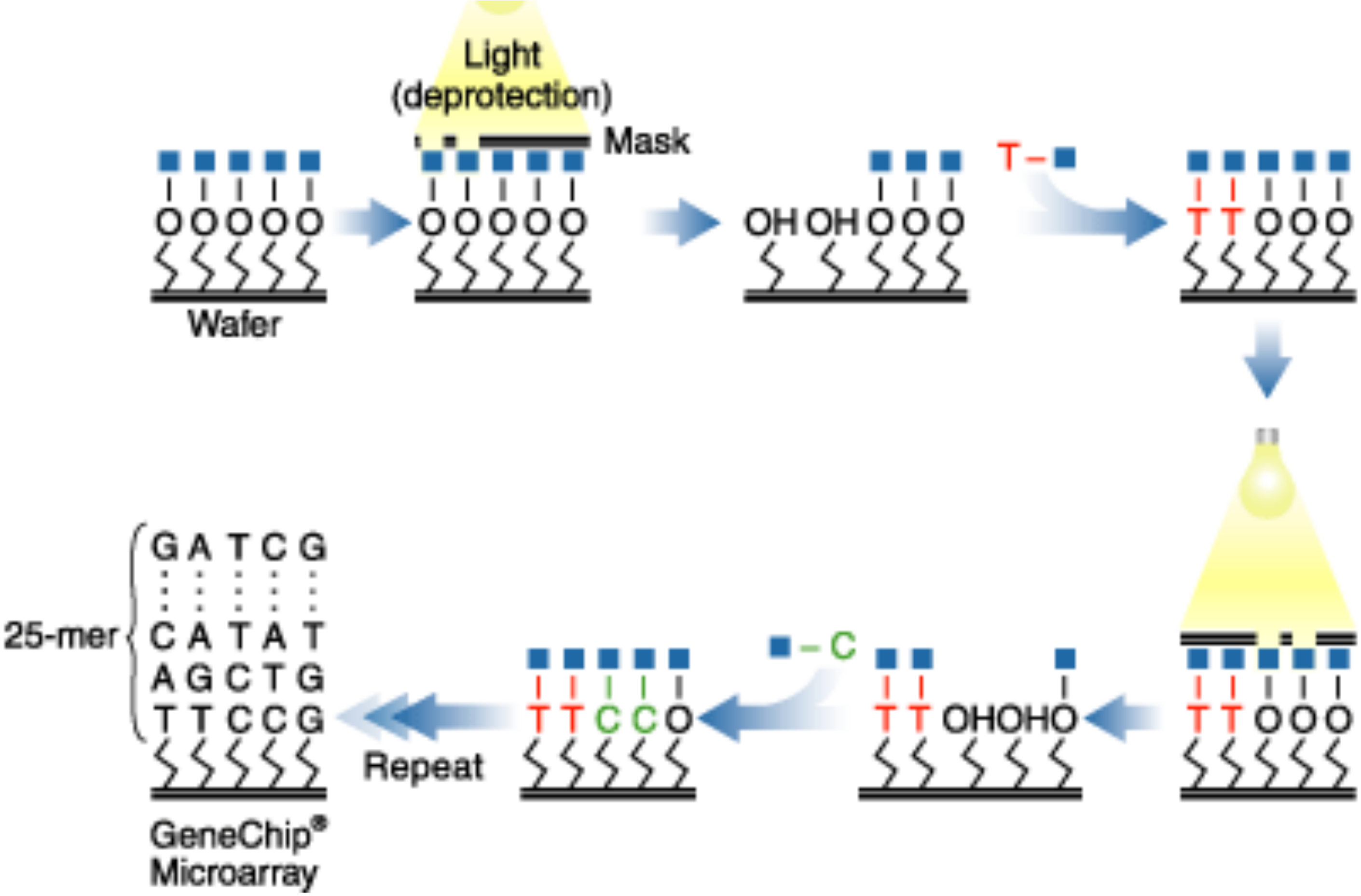
- Applied Microarrays Slides (single channel)
  - $2000
- Vancouver Spotted Arrays (two colour)
  - $800
- Calgary Spotted Arrays (two colour)
  - $1100

Affymetrix Gene Chips*

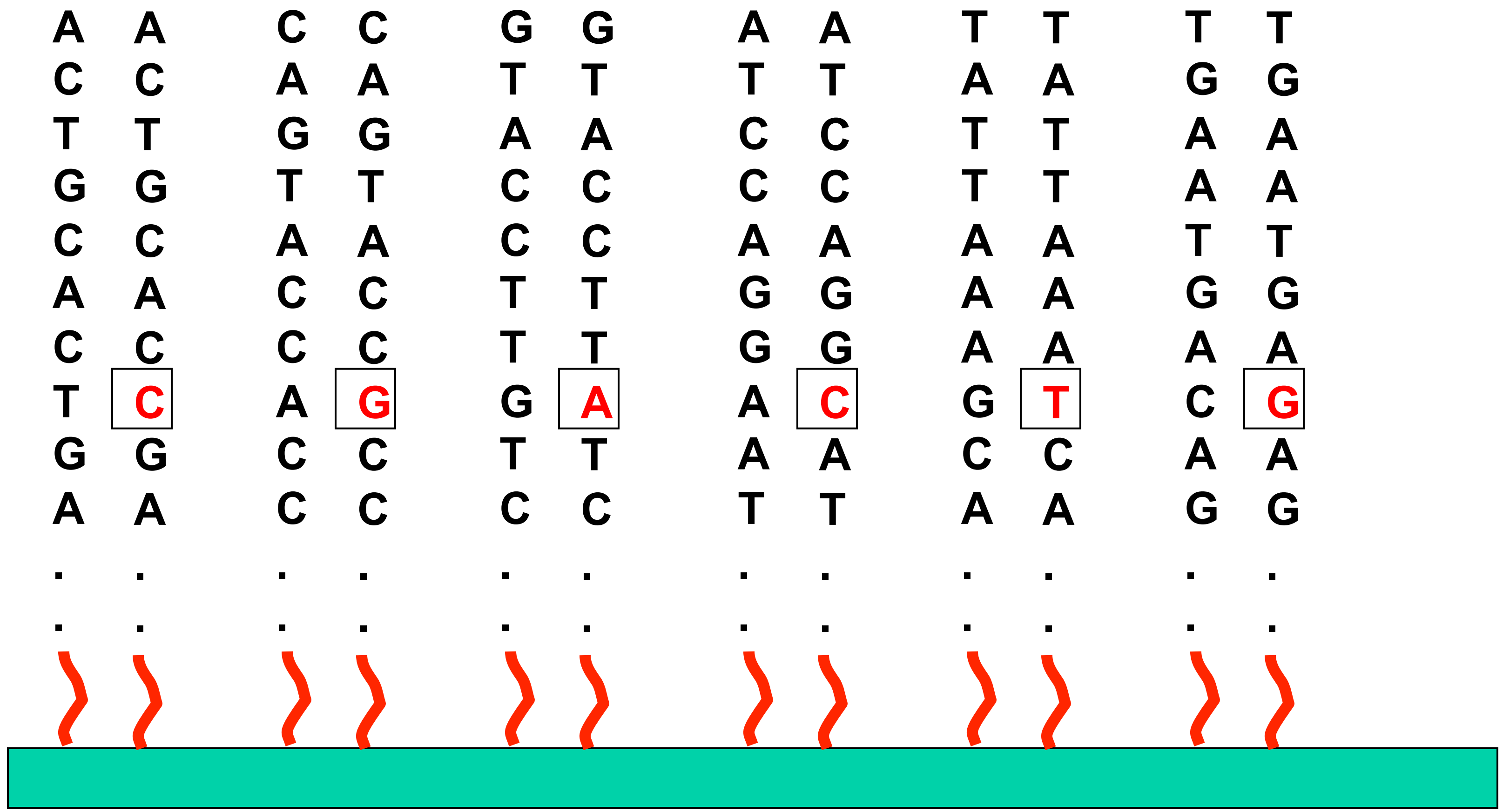
- Chips are 1.7 cm^2
- 400,000 oligo set pairs
- Probe "spots" are 20μ x 20μ
- Each target is 25 bases long
- 11-20 "match" targets and 11-20 "mismatch" targets per gene



Affymetrix Gene Chip*



Affy Chip*

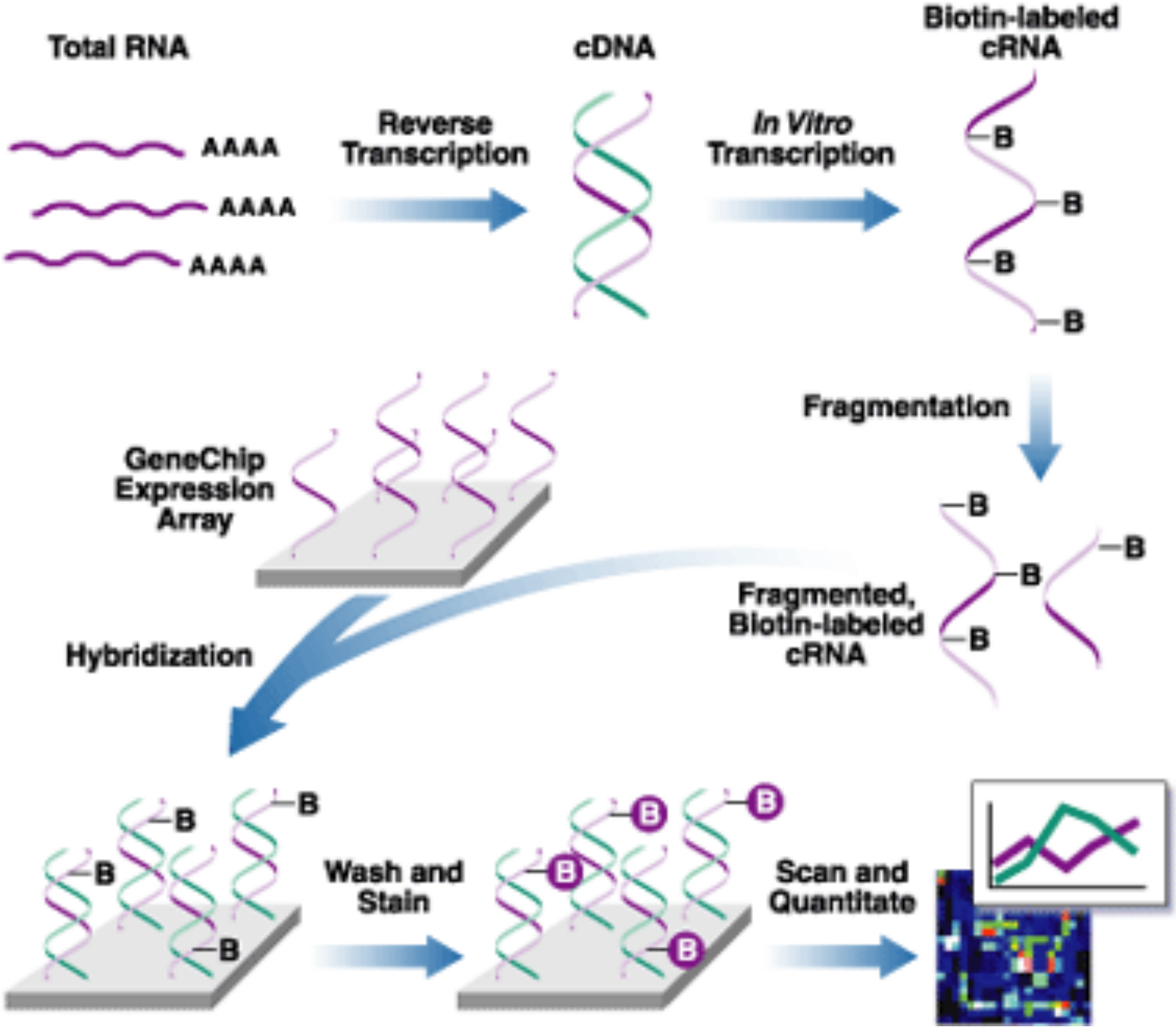


Match target 1	Mis-Match target 1									Match target 14	Mis-Match target 14
A	A	C	C	G	G	A	A	T	T	T	T
C	C	A	A	T	T	T	T	A	A	G	G
T	T	G	G	A	A	C	C	T	T	A	A
G	G	T	T	C	C	C	C	T	T	A	A
C	C	A	A	C	C	A	A	A	A	T	T
A	A	C	C	T	T	G	G	A	A	G	G
C	C	C	C	T	T	G	G	A	A	A	A
T	C	A	G	G	A	A	C	G	T	C	G
G	G	C	C	T	T	A	A	C	C	A	A
A	A	C	C	C	C	T	T	A	A	G	G
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

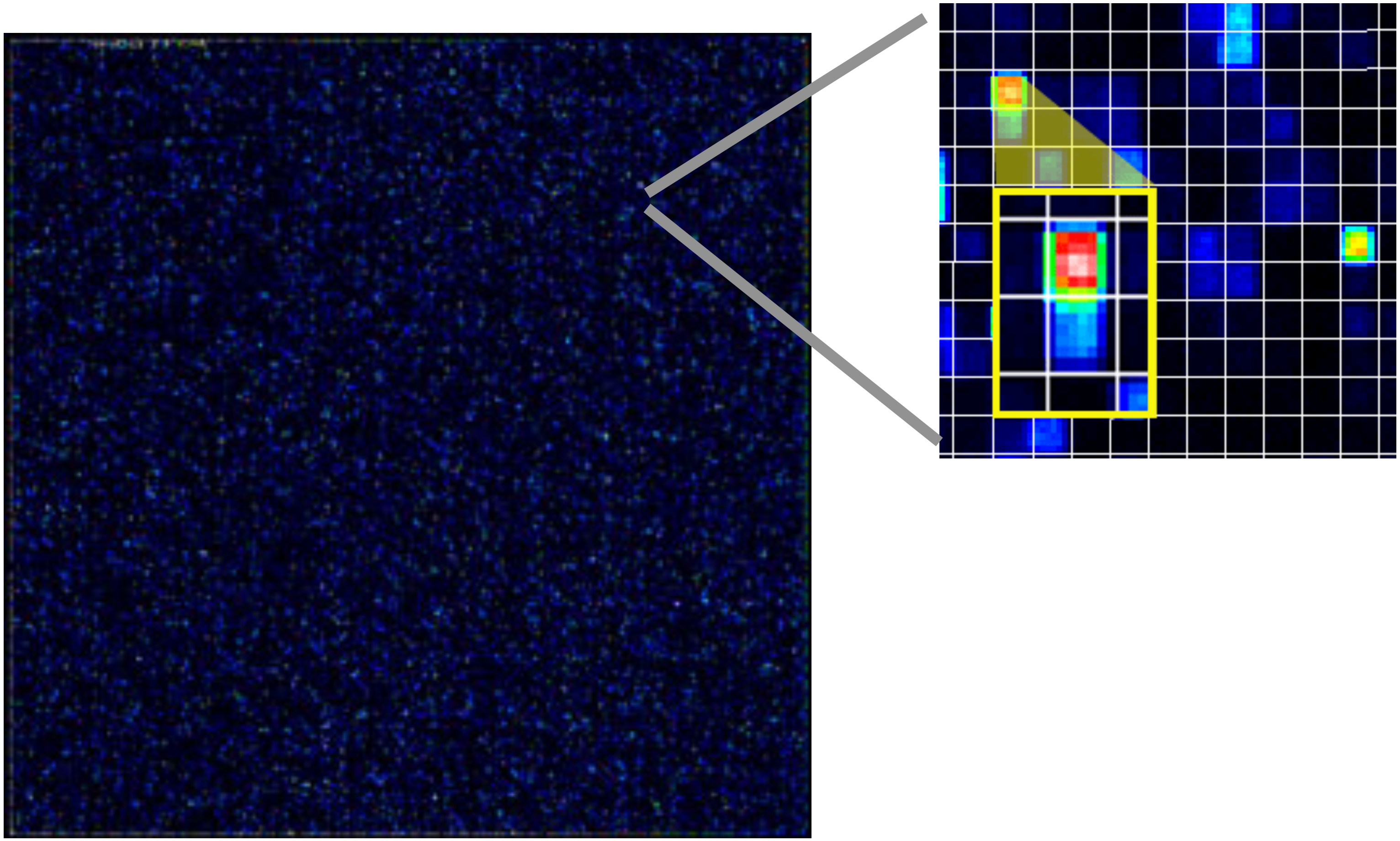
Affy Chip*

- 11-20 targets for each gene/EST
- Each target is 25 bases long
- 1 has exact match, the other is mismatched in the middle base
- Match (M) and mismatch (MM) pairs are placed next to each other
- Expression levels calculated using intensity difference between M & MM for all target pairs

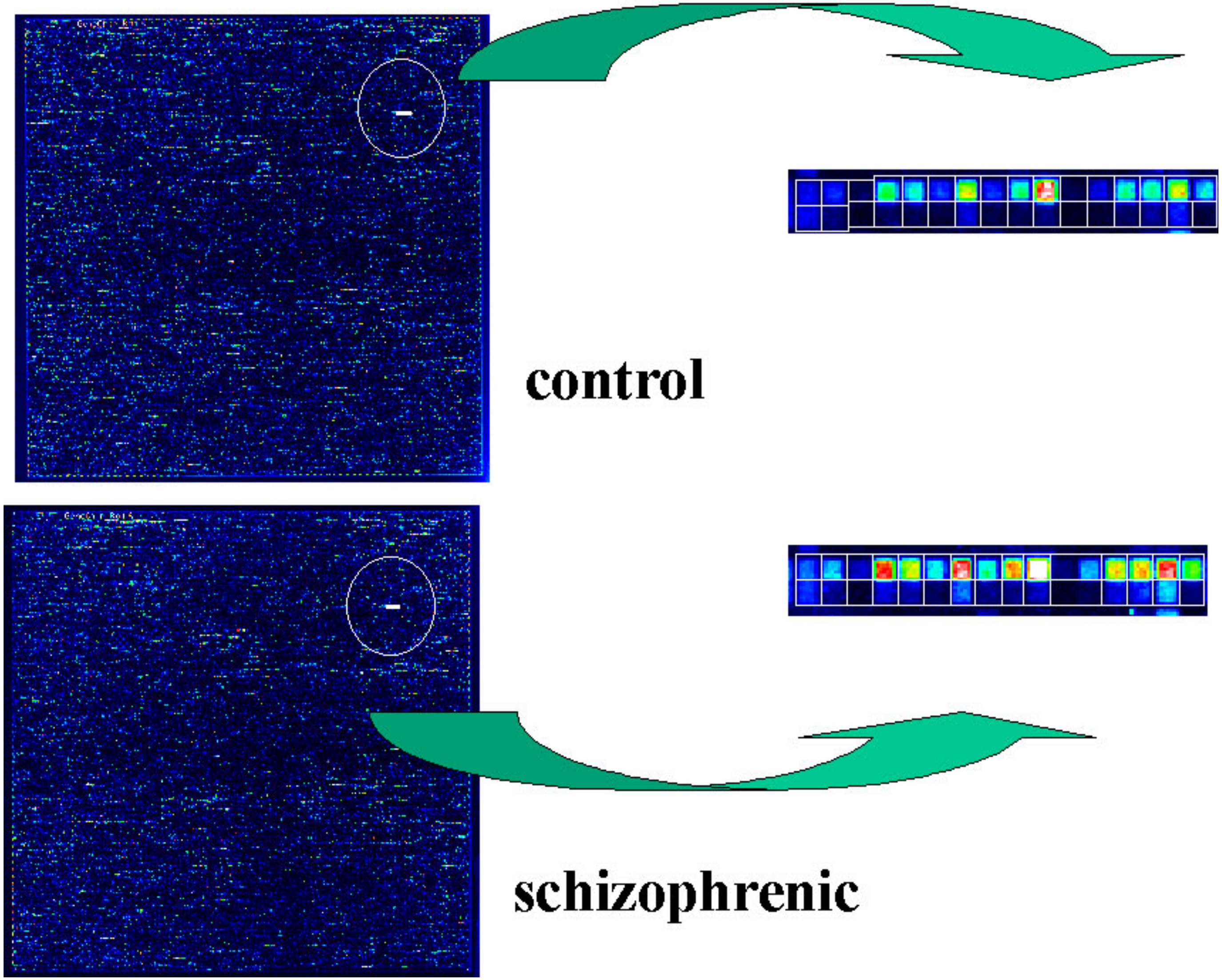
Affymetrix Hybridization*



Affy Chips



Affy Chips



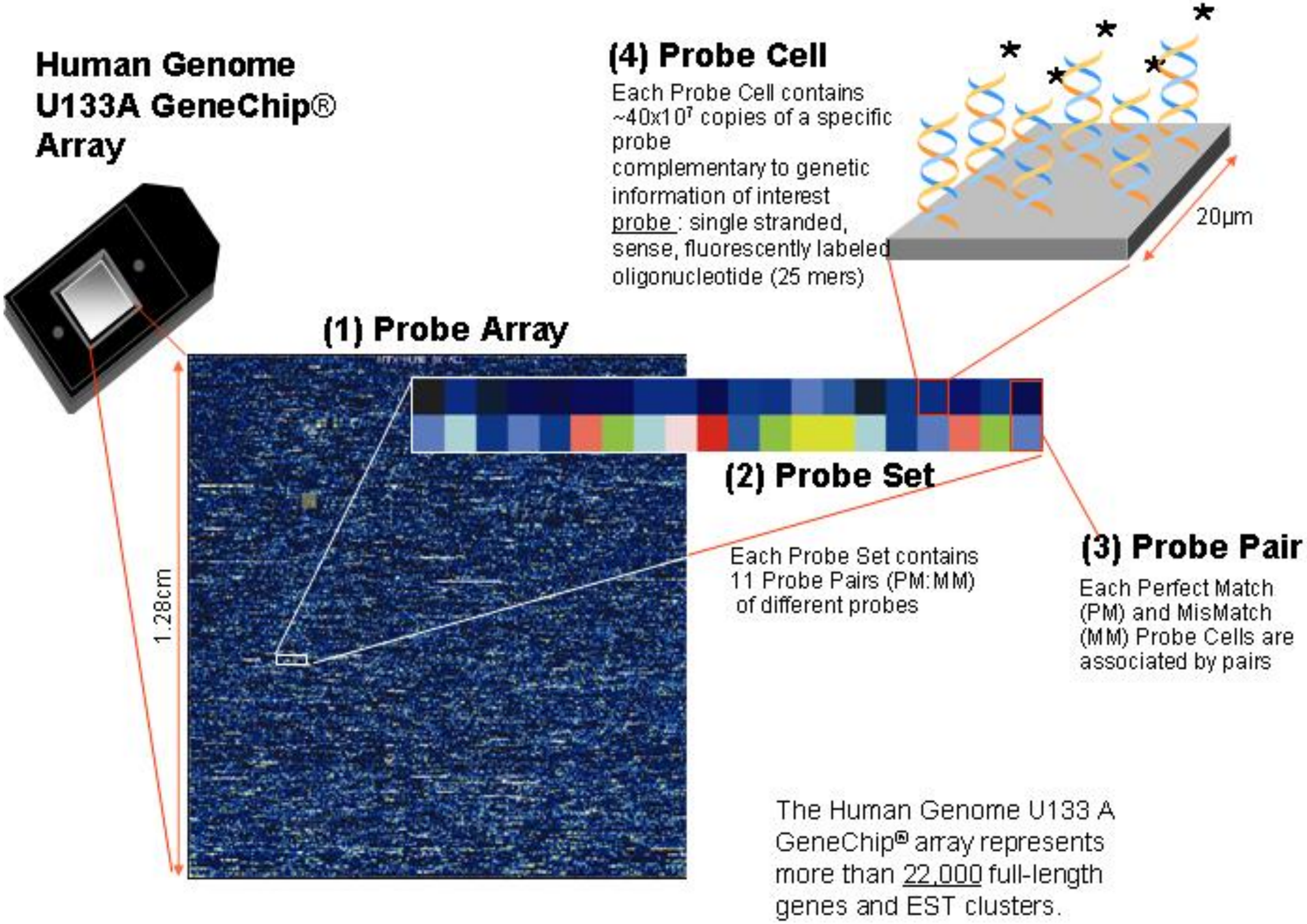
match
mismatch

control

match
mismatch

schizophrenic

Affy Chips



Human Genome U133A GeneChip® Array

(1) Probe Array

(2) Probe Set

Each Probe Set contains 11 Probe Pairs (PM:MM) of different probes

(3) Probe Pair

Each Perfect Match (PM) and MisMatch (MM) Probe Cells are associated by pairs

(4) Probe Cell

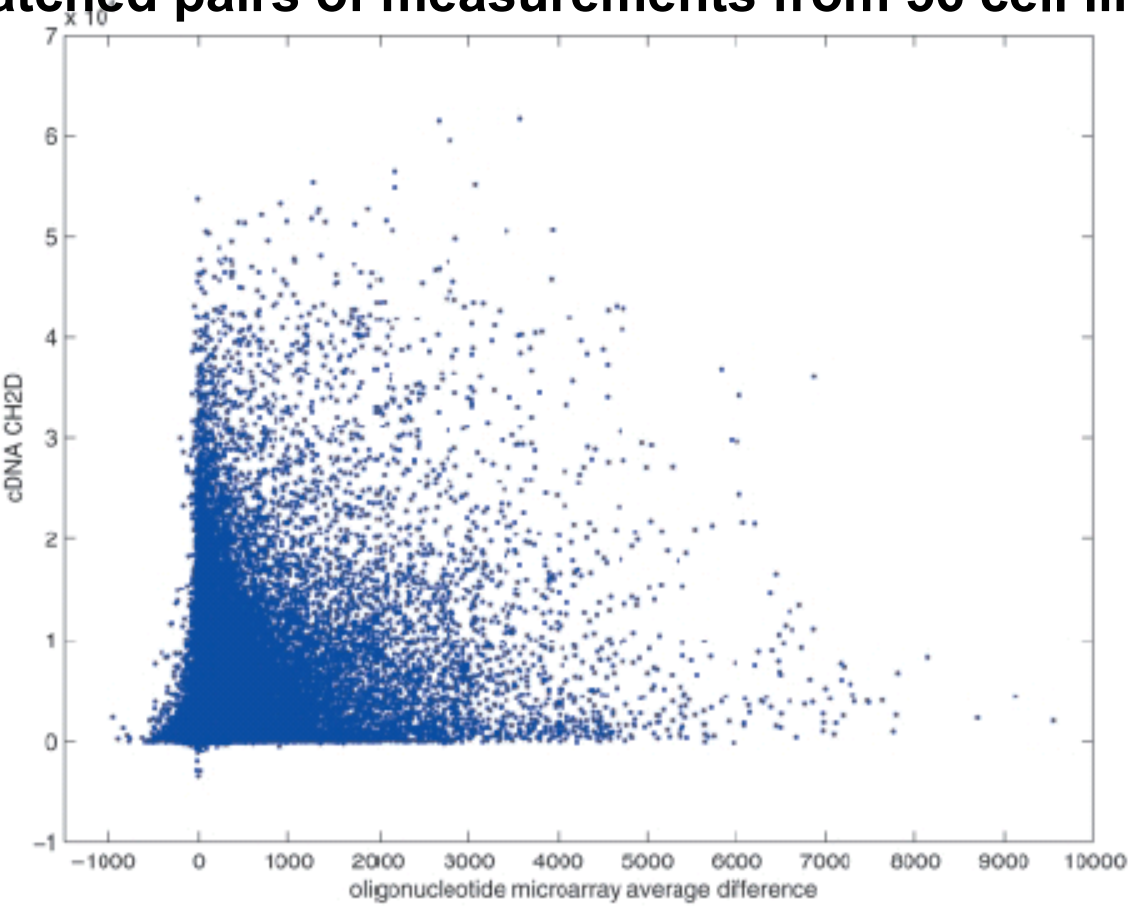
Each Probe Cell contains ~40x10^{7} copies of a specific probe
complementary to genetic information of interest
probe : single stranded, sense, fluorescently labeled oligonucleotide (25 mers)

The Human Genome U133 A GeneChip® array represents more than 22,000 full-length genes and EST clusters.

Kuo et al. (2002) Bioinformatics

Comparison of Affymetrix and Spotted cDNA Arrays

161 620 matched pairs of measurements from 56 cell lines



Spotted Array

Affymetrix

Affymetrix GeneChip Advantages*

- High precision because of:
  - careful target design
  - up to 20 targets per gene
  - up to 20 mismatch targets
- Very precise measurements
- Very high density (500,000 elements/ array)

Affymetrix GeneChips Disadvantages*

- Inflexible: each array requires custom photolithographic masks
- More expensive than spotted arrays ($600-$800 per chip)
- Proprietary technology
  - not all algorithms, information public
  - only one manufacturer of readers, etc.

General Comments*

- Spotted arrays are still wildly popular and widely used – a great learning tool for expression analysis
- Problems have been resolved but spotted arrays are generally less reliable than commercial systems
- Commercial systems (CodeLink and Affy) offer much greater reliability but are expensive & inflexible

Microarray Production*

- Target design and selection
- Printing

Slide making

- RNA extraction
- Labeling
- Hybridization and washing
- Scanning
- Data analysis

Experimental

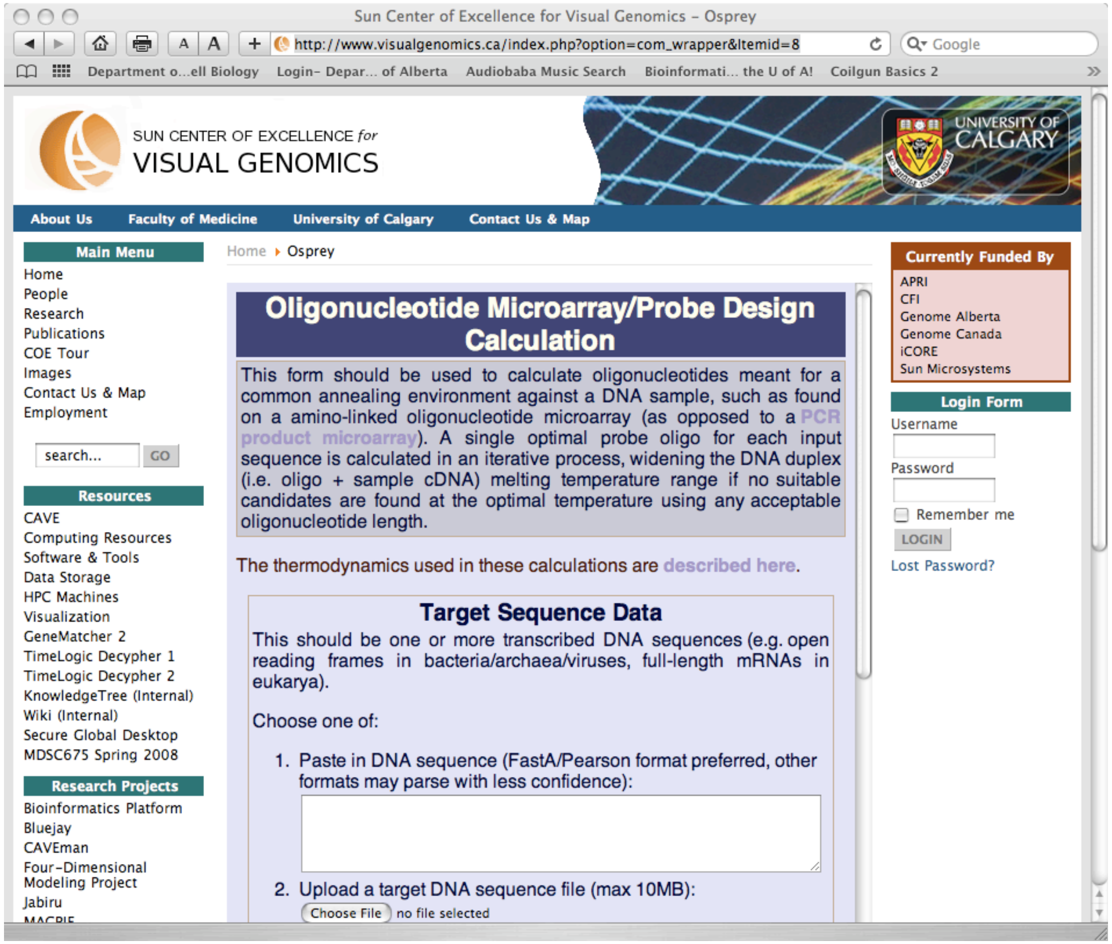
Target Design & Selection*

- Synthetic oligos 25-70 bases in length
- Choose sequences complementary to mRNA of interest
- Random base distribution and average GC content for organism
- Avoid long A+T or G+C rich regions
- Minimize internal secondary structure (hairpins or other loops)
- 1 M salt + 65 °C thermostability

Target Design & Selection*

- Design and select oligo sequences that are less than 75% identical to existing genes elsewhere in the genome (i.e. do a BLAST search)
- Sequences with >75% sequence identity to other sequences will cross-hybridize – leading to confounding results

Osprey - Software for Microarray Target Design



Oligonucleotide Microarray/Probe Design Calculation

This form should be used to calculate oligonucleotides meant for a common annealing environment against a DNA sample, such as found on a amino-linked oligonucleotide microarray (as opposed to a PCR product microarray). A single optimal probe oligo for each input sequence is calculated in an iterative process, widening the DNA duplex (i.e. oligo + sample cDNA) melting temperature range if no suitable candidates are found at the optimal temperature using any acceptable oligonucleotide length.

The thermodynamics used in these calculations are described here.

Target Sequence Data

This should be one or more transcribed DNA sequences (e.g. open reading frames in bacteria/archaea/viruses, full-length mRNAs in eukarya).

Choose one of:

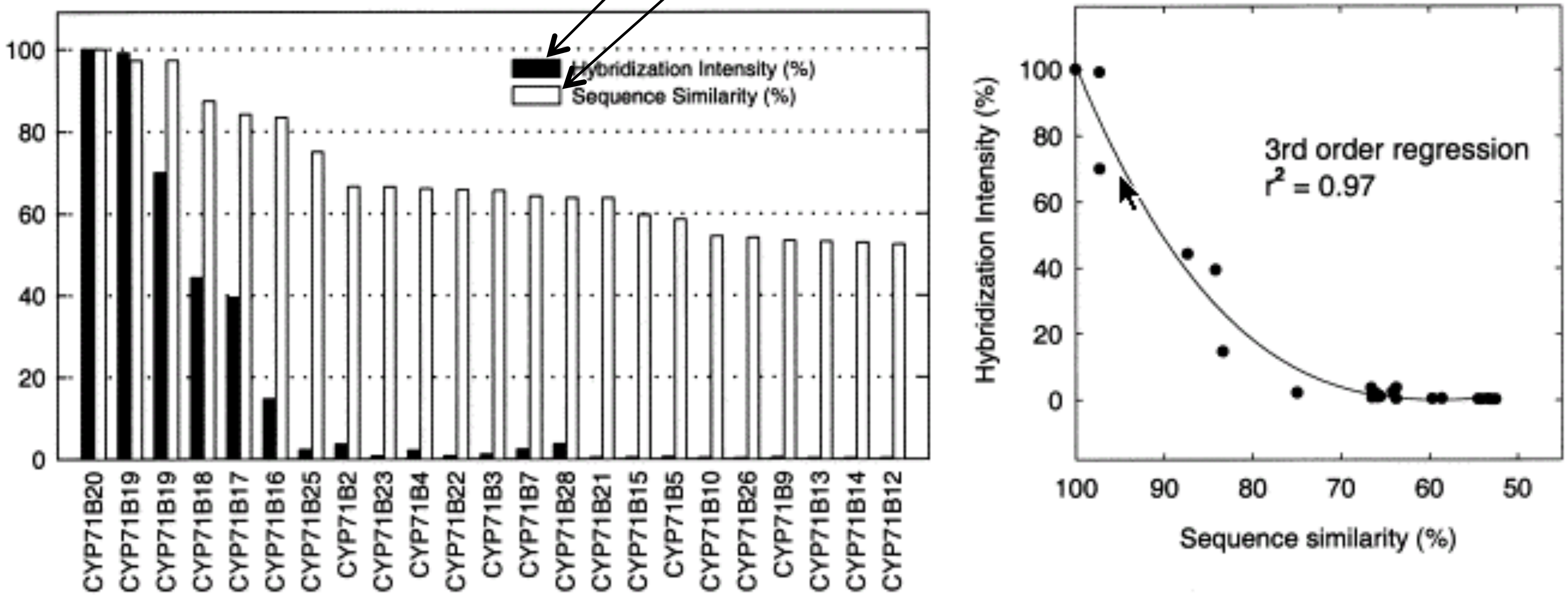
1. Paste in DNA sequence (FastA/Pearson format preferred, other formats may parse with less confidence):
2. Upload a target DNA sequence file (max 10MB):

http://www.visualgenomics.ca/index.php?option=com_wrapper&Itemid=8

Cross-hybridization

hybridization intensity

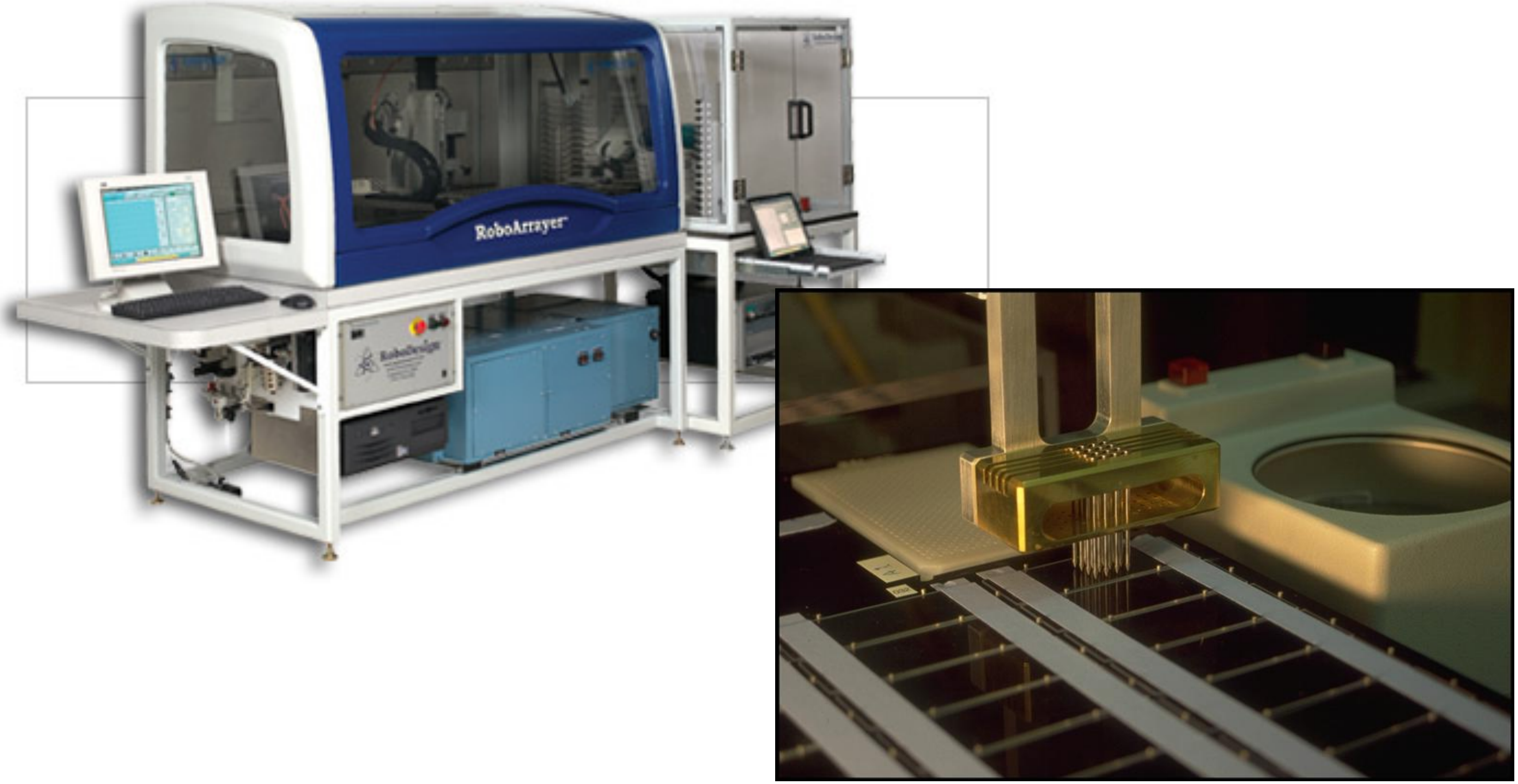
sequence similarity



Analysis of a cross-hybridization within the CYP450 superfamily

Xu et al. (2001) Gene

Microarray Printing



Microarray Printing

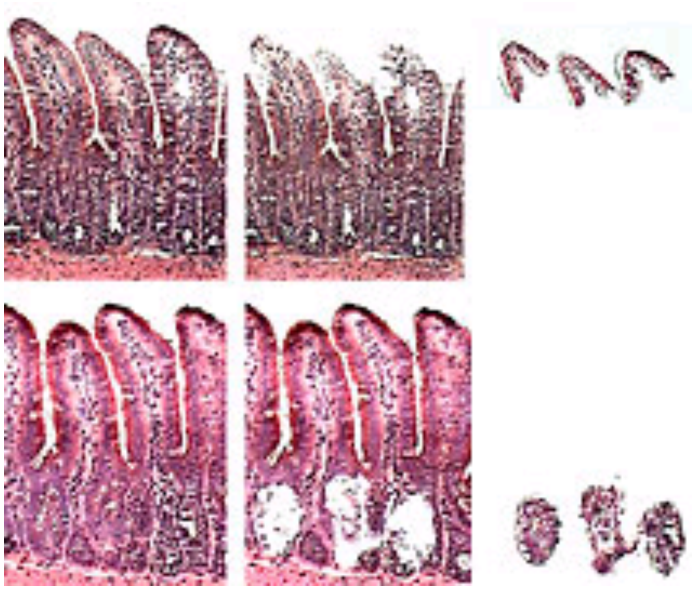
- Targets are deposited by robots using:
  - piezo-electric jets
  - microcapillaries
  - split or solid pins
- Coated glass is the most common substrate
  - aminosilane, poly-lysine, etc. give non-covalent linkages
  - covalent linkage is possible with modified oligos + aldehyde (etc.) coatings

RNA Extraction

- RNA is extremely unstable
- Probably the most problematic step in all microarray analysis
- RNA is extracted as “total RNA”
  - only 1-2% is mRNA
  - remainder is rRNA, tRNA, etc.
- RNA extracted from tissue is often very heterogeneous (many cells and cell types) – watch selectivity

Laser Capture Microdissection

- Cells of interest are visually selected and exposed to an IR laser, which adheres them to a transfer film

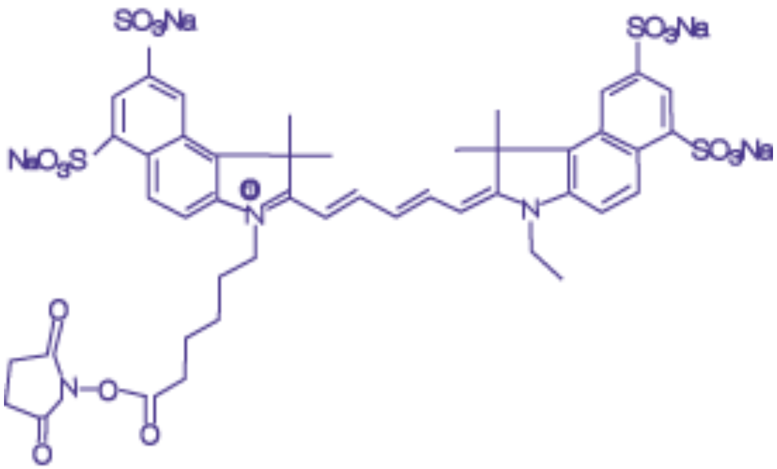


arcturus.com

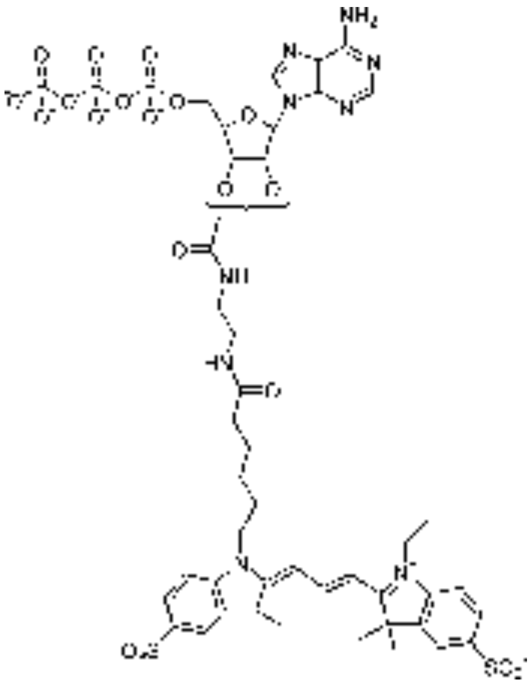
RNA Labeling*

- Common source of systematic error (freshness, contaminants)
- Direct labeling
  - fluorescent nucleotides are incorporated during reverse transcription (“first strand”)
- Indirect labeling
  - reactive nucleotides (aminoallyl-dUTP) are incorporated during RT; first strand product is mixed with reactive fluorescent dyes that bind to amino group

Direct Labeling*

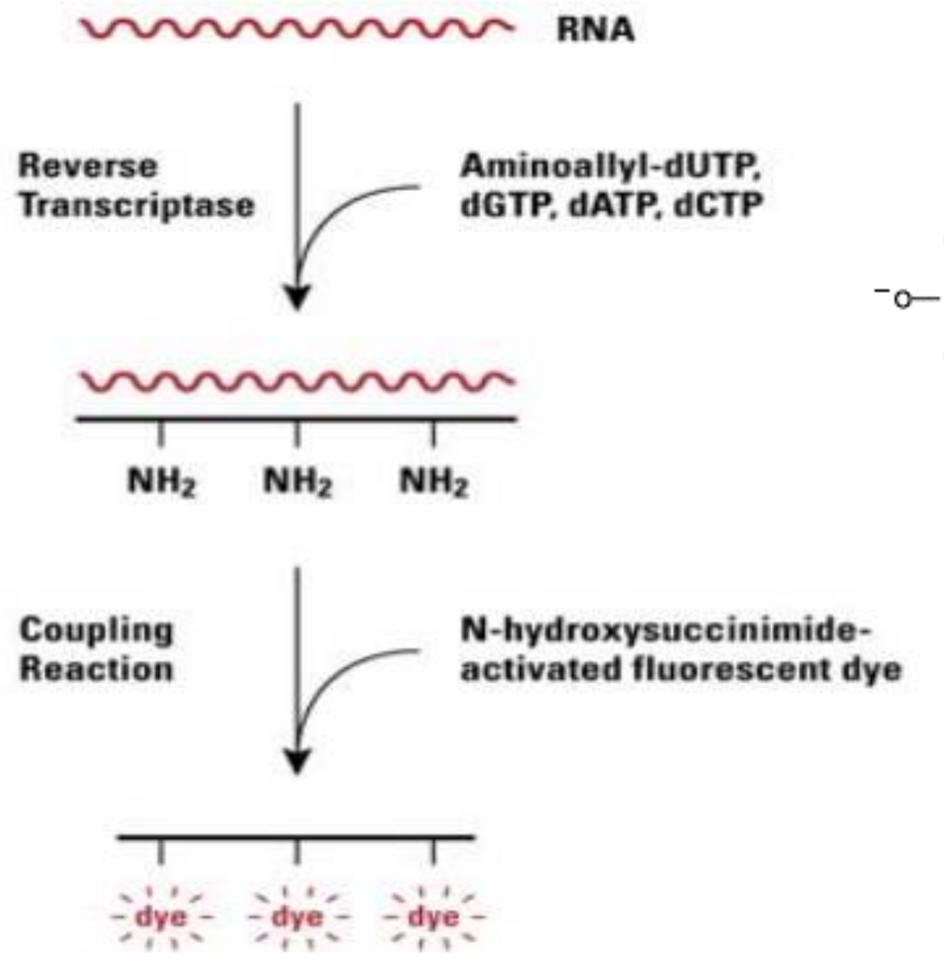


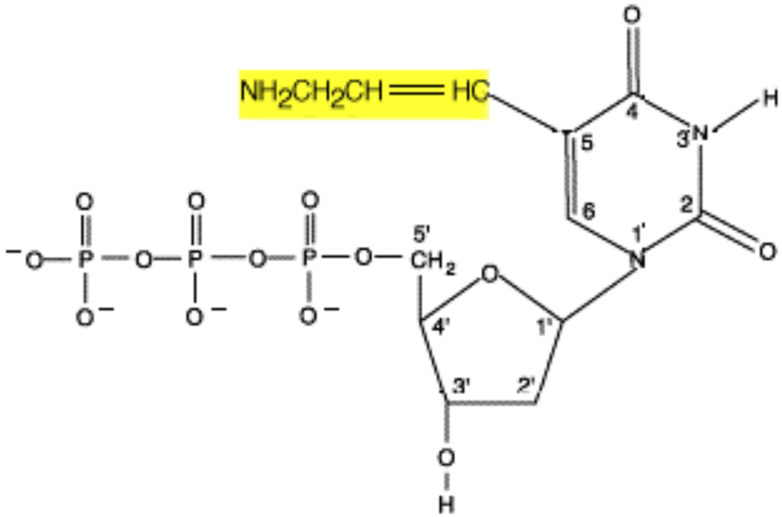
Cy5



Cy3-ATP

Indirect Labeling


RNA
Reverse Transcriptase
Aminoallyl-dUTP, dGTP, dATP, dCTP
NH2
NH2
NH2
Coupling Reaction
N-hydroxysuccinimide-activated fluorescent dye
dye
dye
dye



aminoallyl-dUTP

Hybridization

- Stringency of hybridization is affected by ions, detergents, formamide, temperature, time...
- Hybridization may be an important source of systematic error
- Automated hybridization systems exist; value is debatable

Lee et al. (2000) *PNAS*

How Many Replicates?

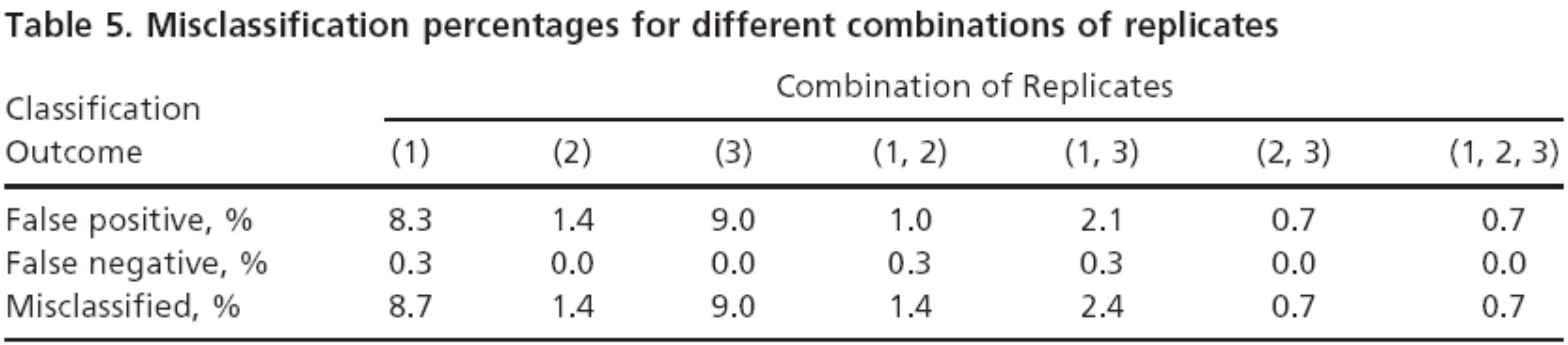


Table 5. Misclassification percentages for different combinations of replicates

Classification	Combination of Replicates						
Outcome	(1)	(2)	(3)	(1, 2)	(1, 3)	(2, 3)	(1, 2, 3)
False positive, %	8.3	1.4	9.0	1.0	2.1	0.7	0.7
False negative, %	0.3	0.0	0.0	0.3	0.3	0.0	0.0
Misclassified, %	8.7	1.4	9.0	1.4	2.4	0.7	0.7

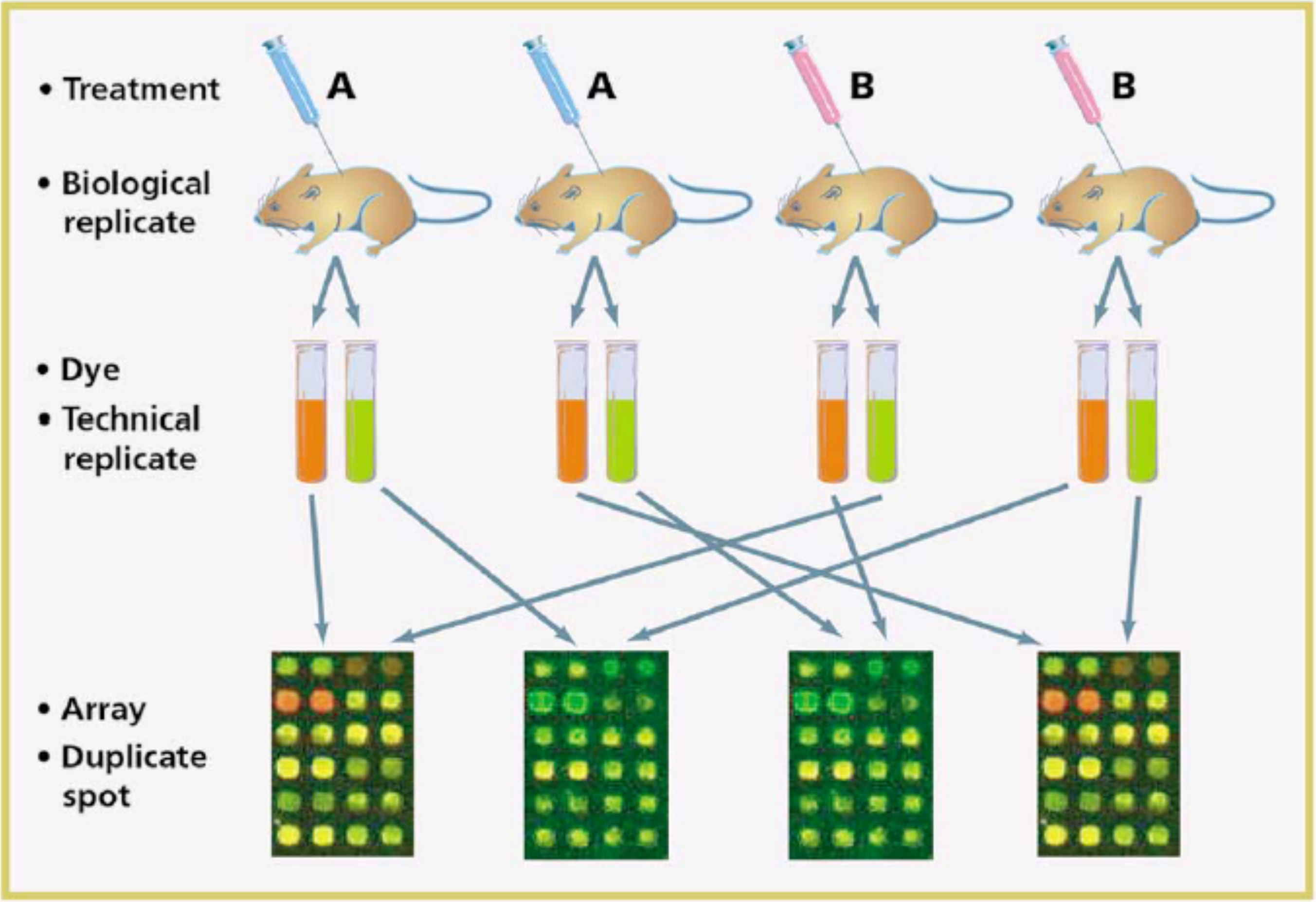
Singletons **Duplicates** **3X**

- **Substantial error when only one array analyzed, standard is to use 3 replicates**

What Types of Replicates?*

Biological replicates

Technical replicates



Bob Crimi

Biological replication is most important because it includes all of the potential sources for error

Microarray Production

- Target design and selection
- Printing
- RNA extraction
- Labeling
- Hybridization and washing
- Scanning
- Data analysis