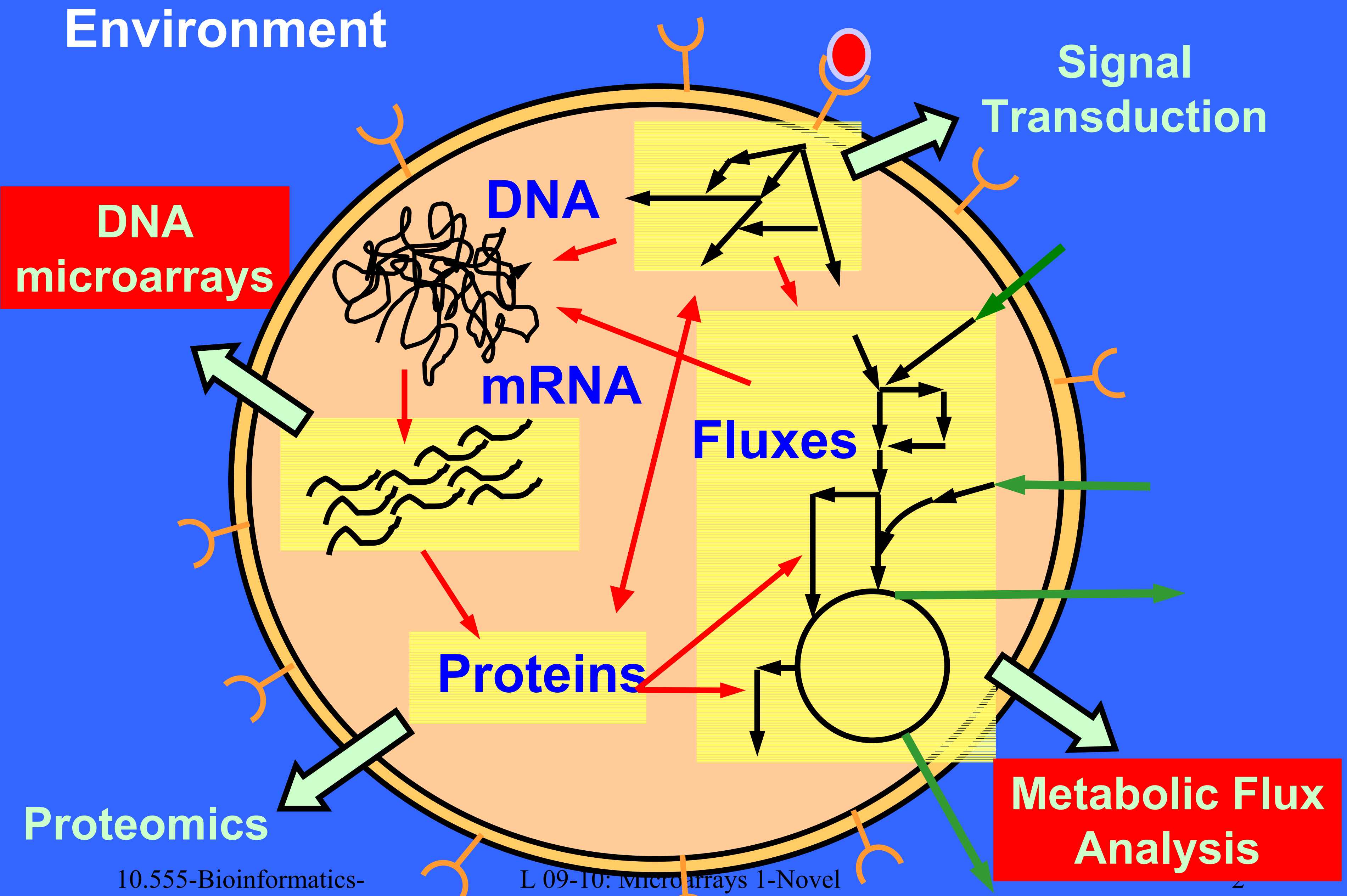
10.555
Bioinformatics: Principles, Methods and Applications

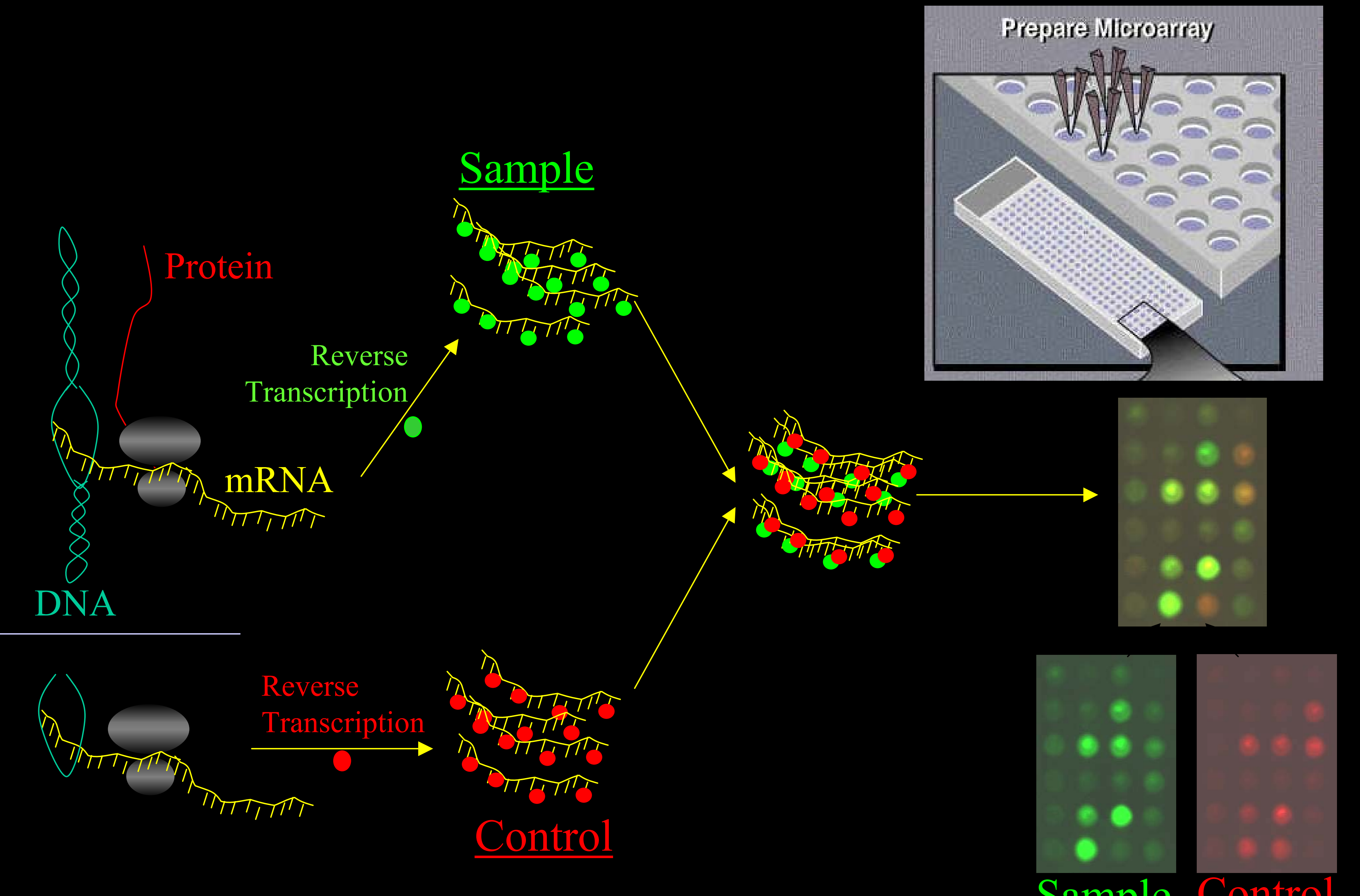
MIT, Spring term, 2003

Lecture 9,10:
DNA Microarrays: Novel Applications

Probing cellular function



DNA Micro-Array Methodology



Protein

DNA

mRNA

Reverse Transcription

Sample

Reverse Transcription

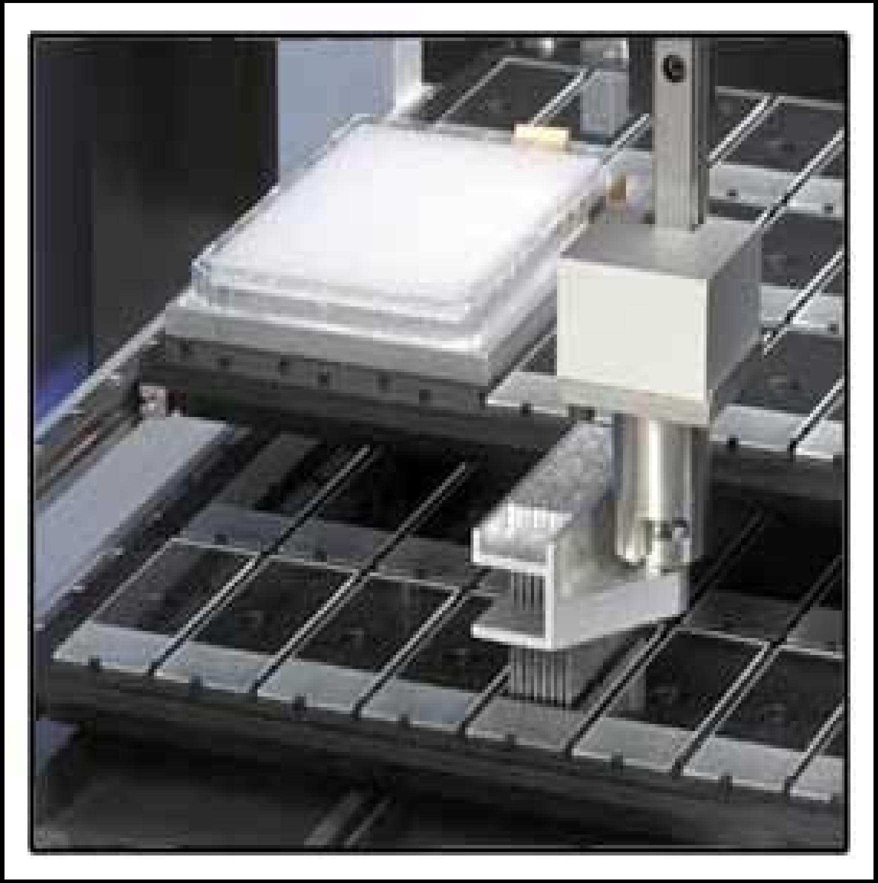
Control

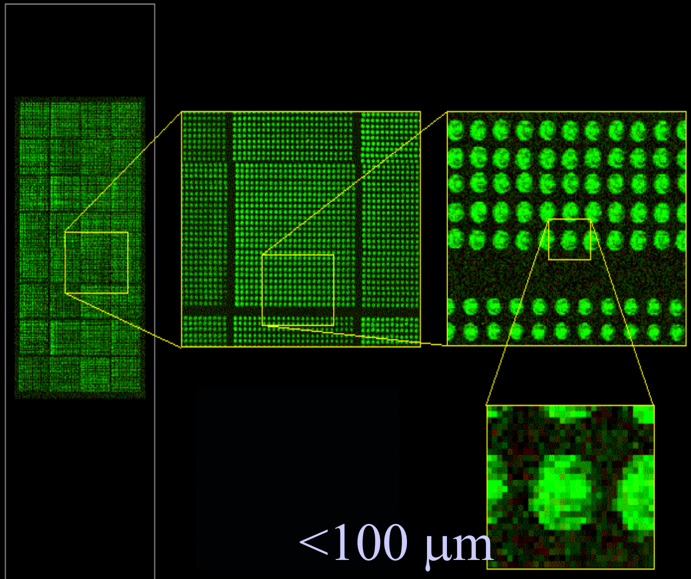
Prepare Microarray

Sample

Control

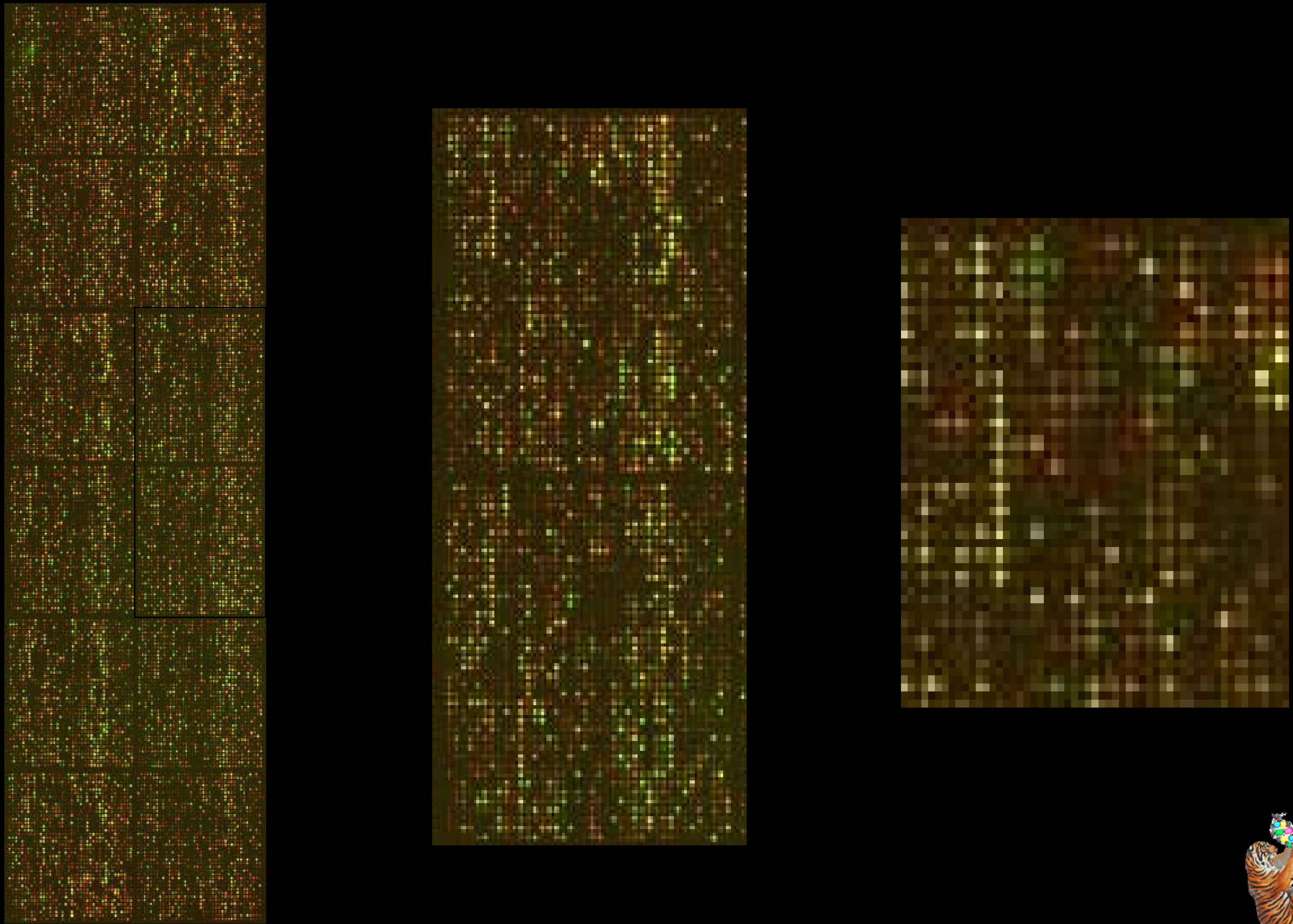
DNA Micro-Arraying Technology





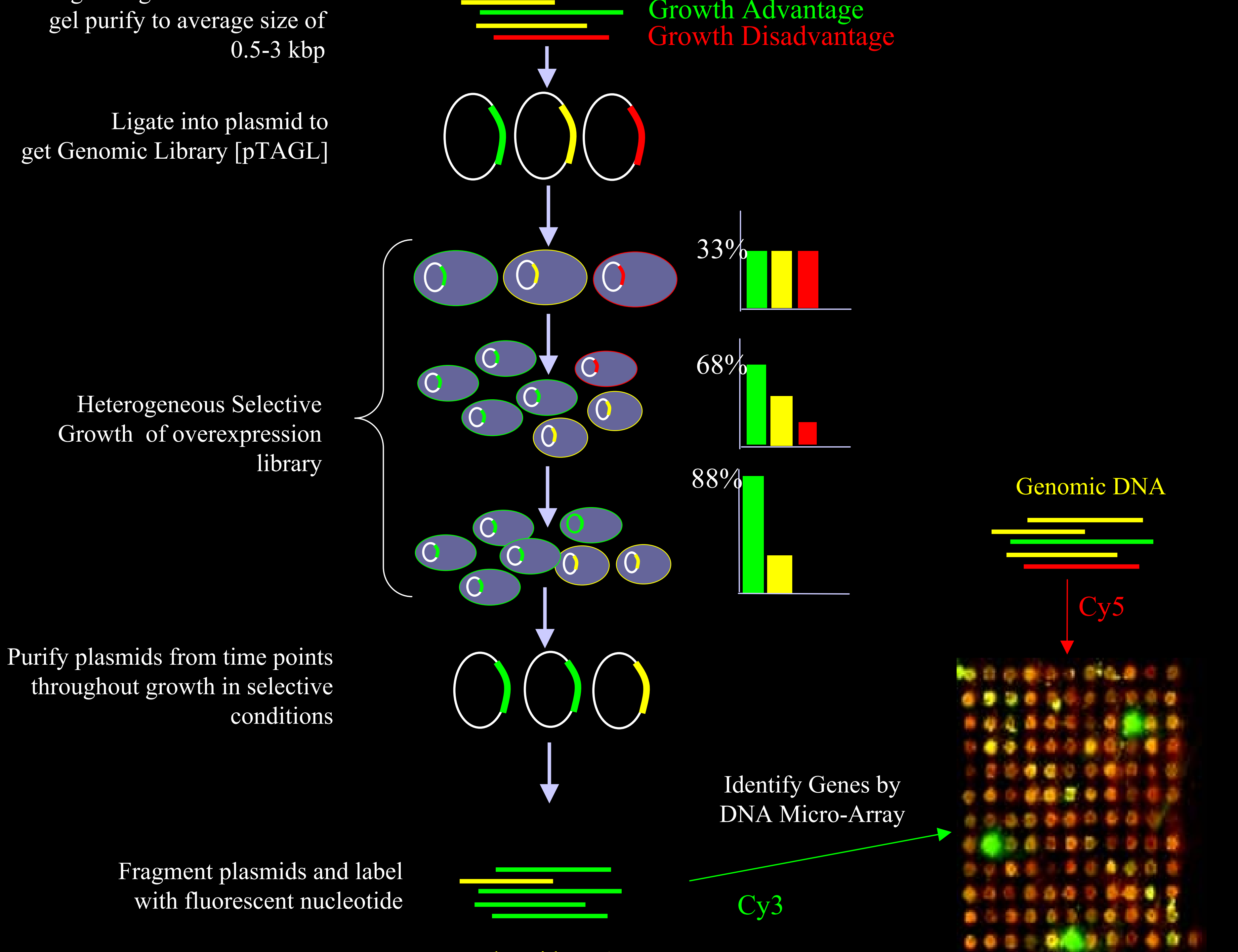
Mountains of Biological Data

19,200 Human Gene Array



Beyond transcriptional studies: New applications of microarrays

1. **Genome-wide screening for genes conferring specific cellular traits (Gill *et al.*, *PNAS*, May 2002)**



gel purify to average size of
0.5-3 kbp

Growth Advantage
Growth Disadvantage

Ligate into plasmid to
get Genomic Library [pTAGL]

Heterogeneous Selective
Growth of overexpression
library

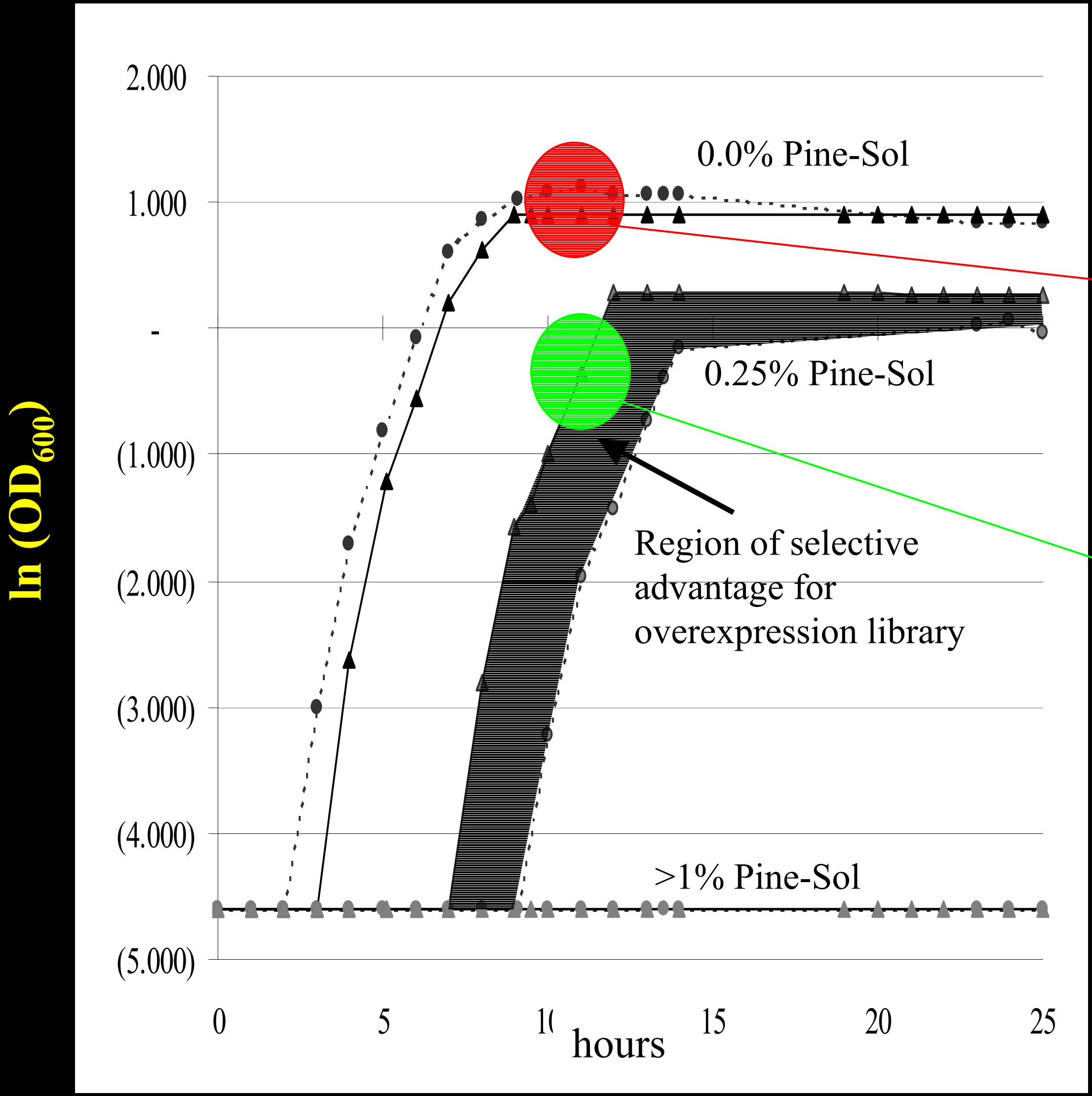
Purify plasmids from time points
throughout growth in selective
conditions

Fragment plasmids and label
with fluorescent nucleotide

Identify Genes by
DNA Micro-Array

Genomic DNA

Determining Selective Conditions



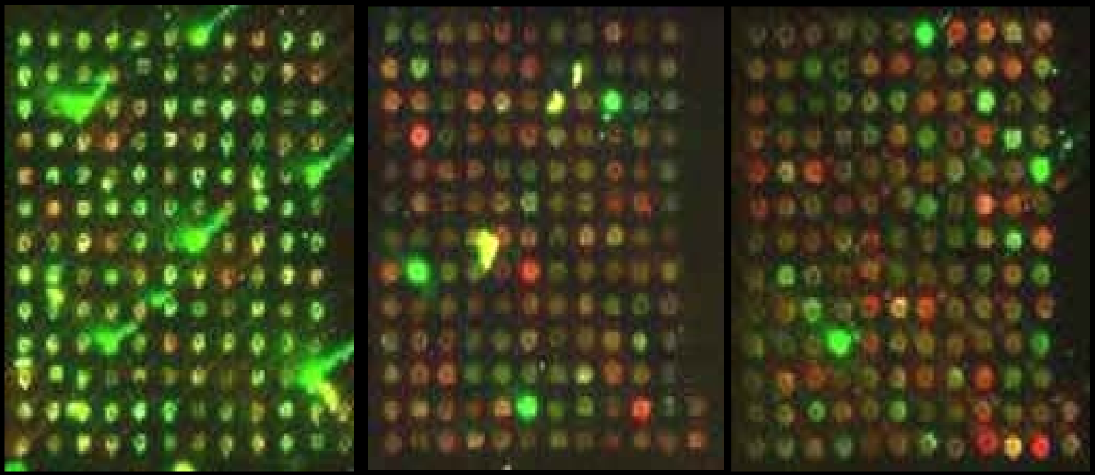
Which plasmids are enriched?

Which plasmids are enriched?

○ Wild-Type Cell

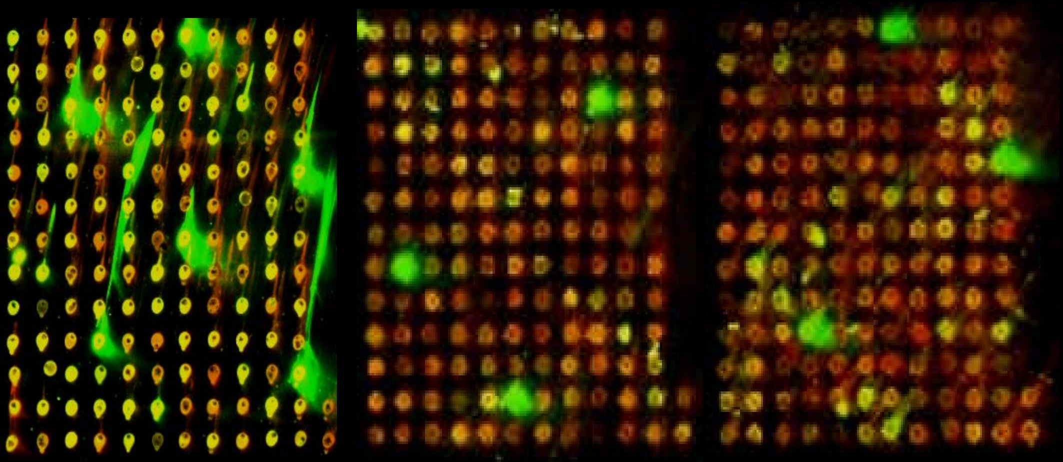
▲ Library Transformed Wild-Type Cell

0.4% Pine-Sol
mid-exponential phase

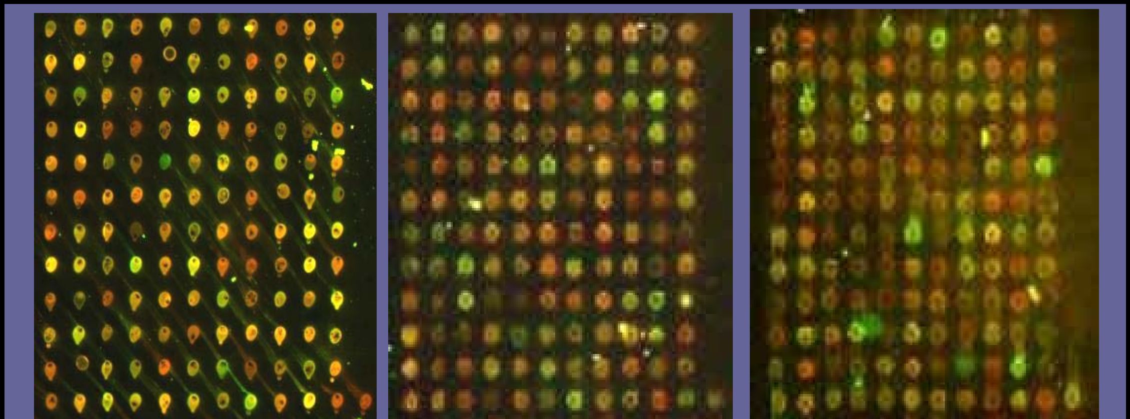


Same culture
+3 hours

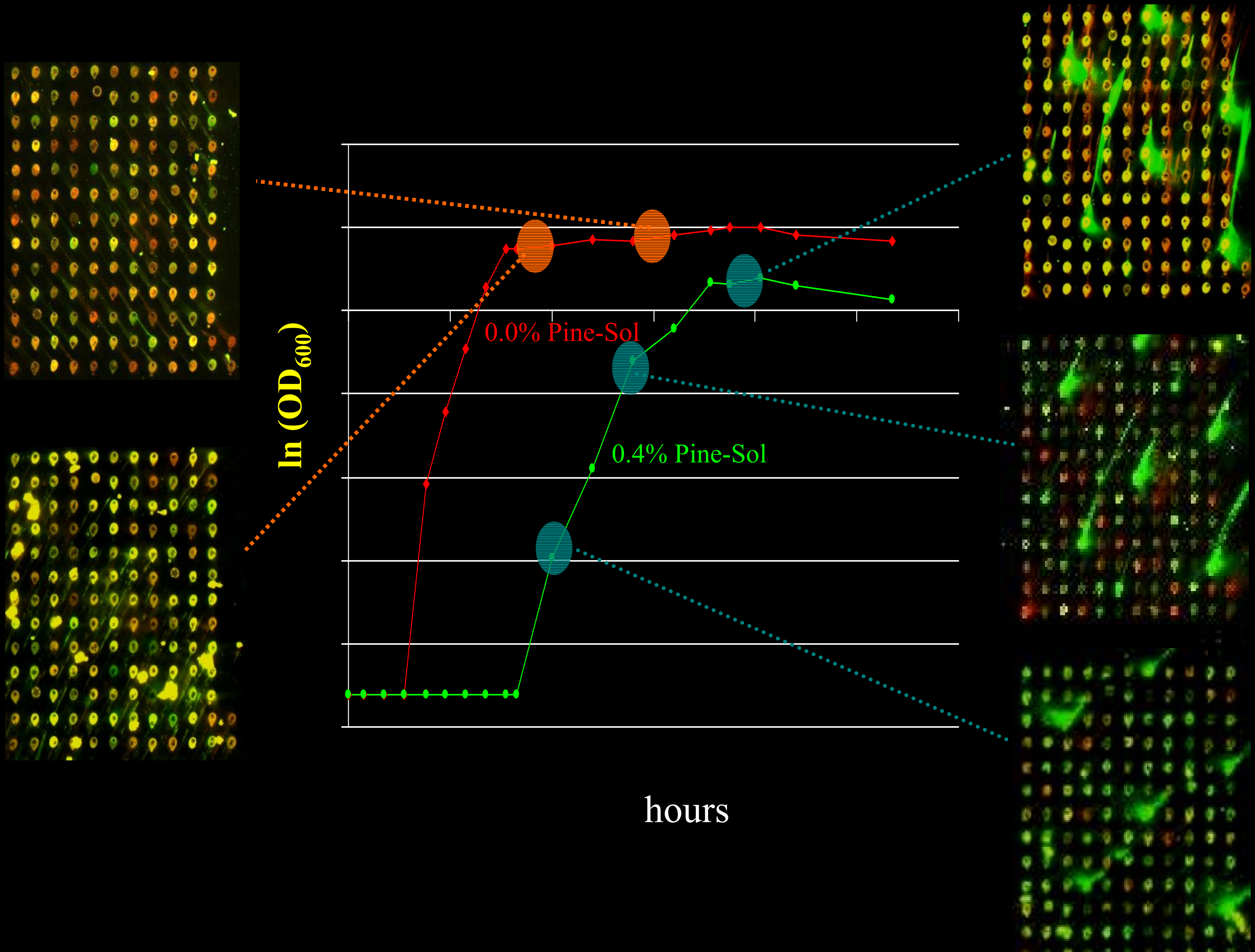
0.4% Pine-Sol
early stationary phase



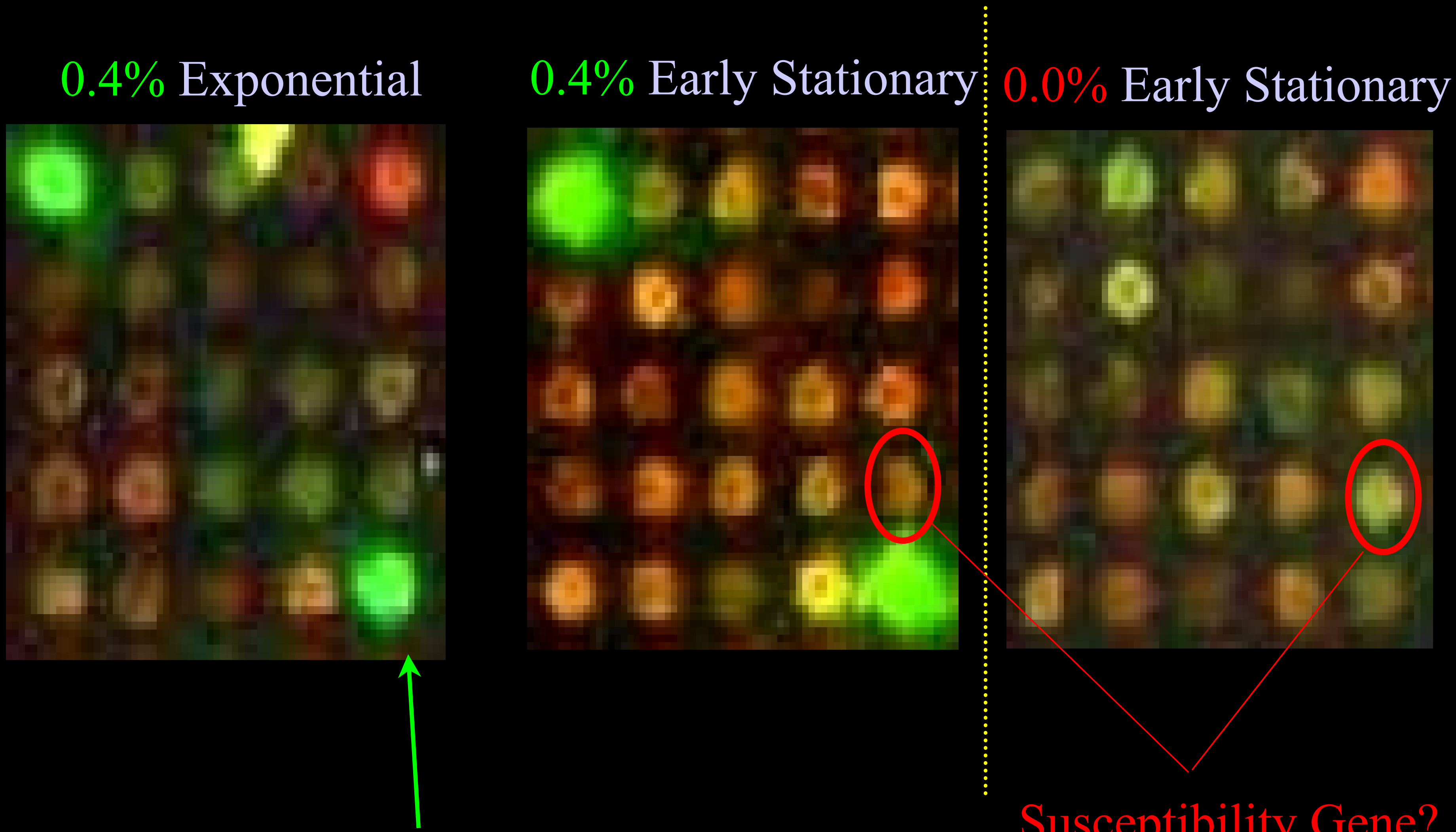
Control
early stationary phase



Dynamic Screening of Plasmid Populations



Discovering Antibiotic Resistance and Susceptibility Genes



0.4% Exponential

0.4% Early Stationary

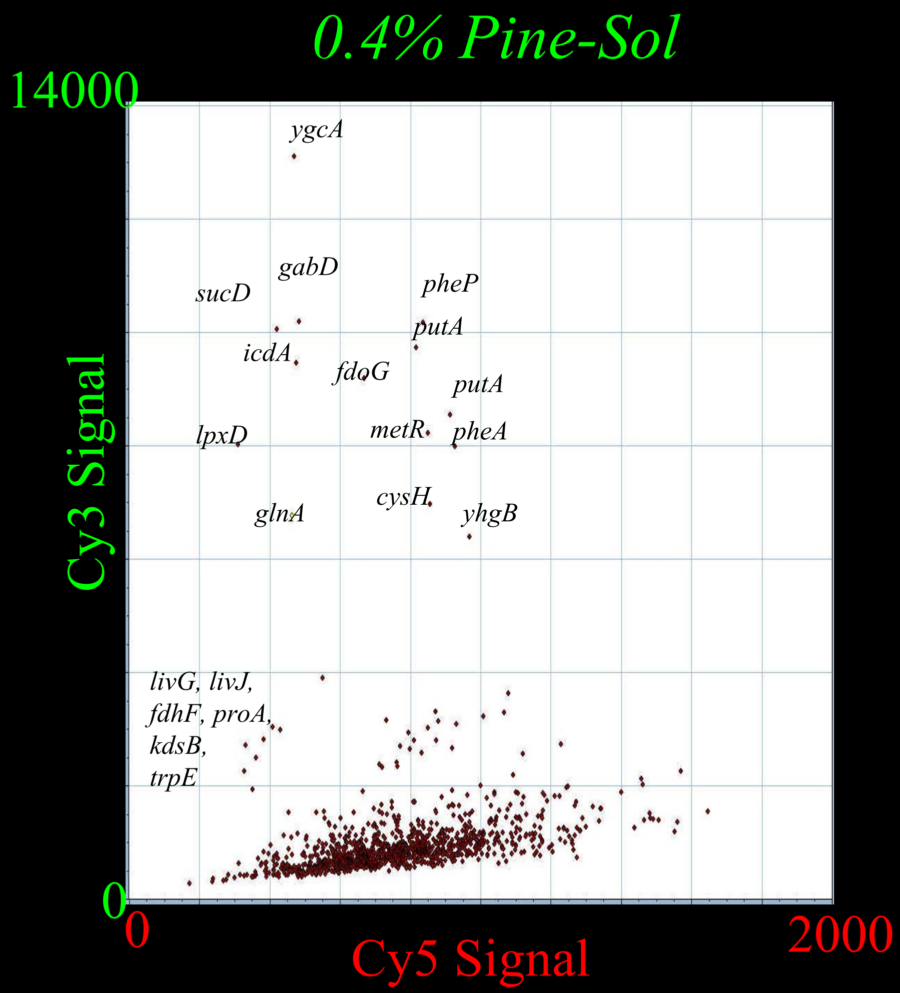
0.0% Early Stationary

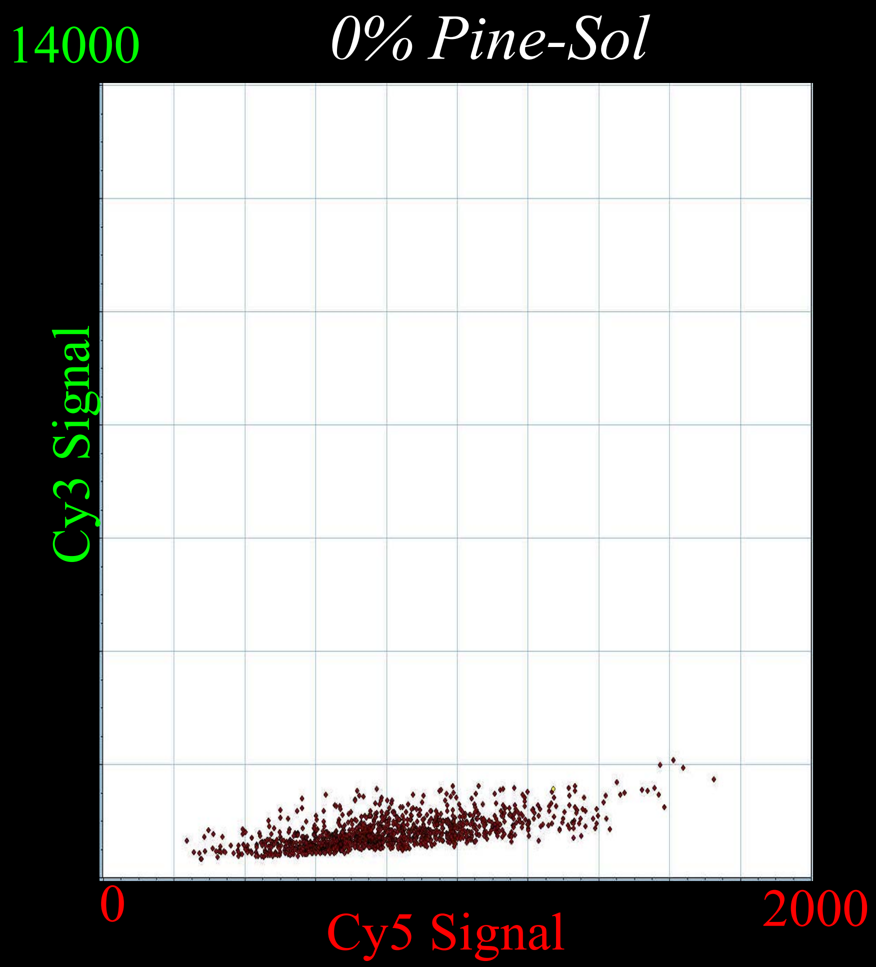
Susceptibility Gene?

Plasmids: Resistance Genes

Genomic DNA: Control

Quantifying Plasmid Levels to Identify Resistance Genes


0.4% Pine-Sol


0% Pine-Sol

Antibiotic Resistance Genes

	Gene	***Function***
	ygcA	Unknown
→	***gabD***	Succinate-semialdehyde dehydrogenase
→	***pheP***	Phenylalanine specific permease
	sucD	Succinyl-CoA synthetase
→	***putA***	Proline dehydrogenase
→	***icdA***	Isocitrate dehydrogenase
→	***fdoG***	Formate dehydrogenase
	metR	Positive regulator for methionine genes
→	***lpxD***	Glucosamine-N-acyltransferase (lipid A)
→	***pheA***	Phenylalanine dehydrogenase
	cysH	Adenylsulfate reductase
	glnA	Glutamine synthetase
	yhgB	Unknown
→	***livG***	High affinity branched chain amino acid transporter
→	***livJ***	High affinity branched chain amino acid transporter
→	***fdhF***	Formate dehydrogenase
	trpE	Anthranilate synthase component (tryptophan)
→	***kdsB***	CMP-3-deoxy-D-manno-octulosonate cytidylyltransferase (lipid A)
	proA	Glutamyl P reductase

Validation Experiments

Transformant	Trait	MIC[1] Ratio[2]	CV[4]	p-value[5]	N[6]
pUC19 Control	Control	1.0[3]	0	NA	29
Library	Res	1.20	0.000	0*[7]	4
Enriched Library	Res	1.40	0.165	0.020	4
pBAD-*hybC*	Null	1.06	0.065	0.135	4
pBAD-*leuC*	Null	1.06	0.065	0.135	4
pBAD-*trpE*	Res	1.14	0.000	0.029	4
pBAD-*livJ*	Res	1.19	0.080	0.015	4
pBAD-*pheP*	Res	1.18	0.112	0.040	9

[1] MIC: Minimum Inhibitory Concentration. [2] Ratio of MIC of the corresponding strain by the MIC value of the pUC19 control. [3] MIC for pUC control was 0.6 +/- 0.063 %(v/v) Pine-Sol in LBA. [4] Coefficient of Variation = SD/mean. Each strain was tested four to nine times with the pUC control tested 29 times. [5] p-value is the one-tailed probability that the mean MIC between the transformant and the pUC19 control were equal using a students means t-test for two samples of unequal variances. Therefore the null hypothesis that the MIC means of the control and the strains reported are the same is rejected with 95% confidence for all strains except of those of the null genes. [6] Number of separate MIC assays. [7] Variance in library MIC assays was zero

Beyond transcriptional studies: New applications of microarrays

1. **Genome-wide screening for genes conferring specific cellular traits (Gill *et al.*, 2002)**
2. **Genome-wide location and function of DNA binding proteins (Ren *et al.*, *Science*, December 22, 2000)**

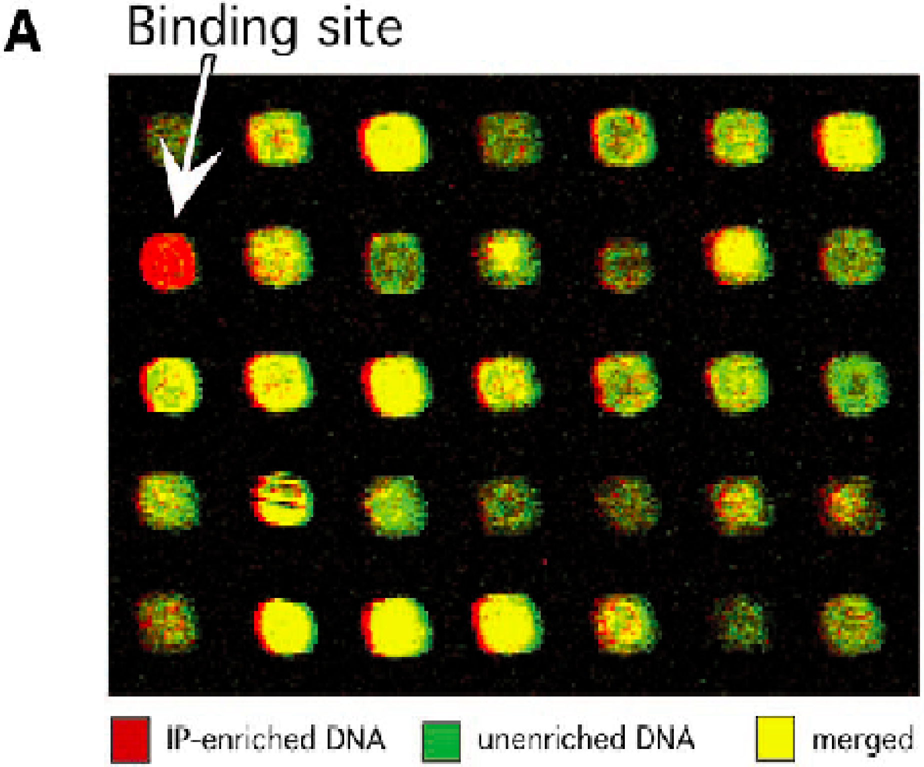
Research Goal

- Understand how DNA binding proteins regulate global gene expression
- Study genes whose expression is directly controlled by Gal4 and Ste12

ChIP Chip

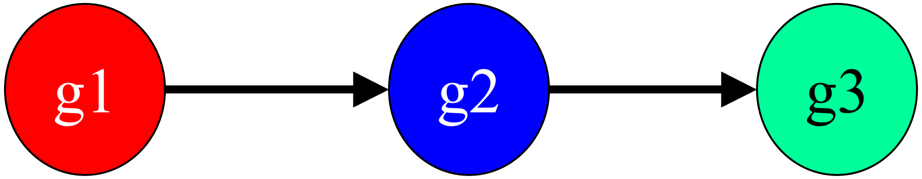
- ChIP: Chromatin Immuno-Precipitation
- Fix cells with formaldehyde and harvest
- Use an antibody to precipitate DNA fragments bound to protein of interest
- Remove cross-links, amplify, label (Cy5)
- Mix with unenriched sample (Cy3)
- Bind to DNA microarray

Example of results



Why?

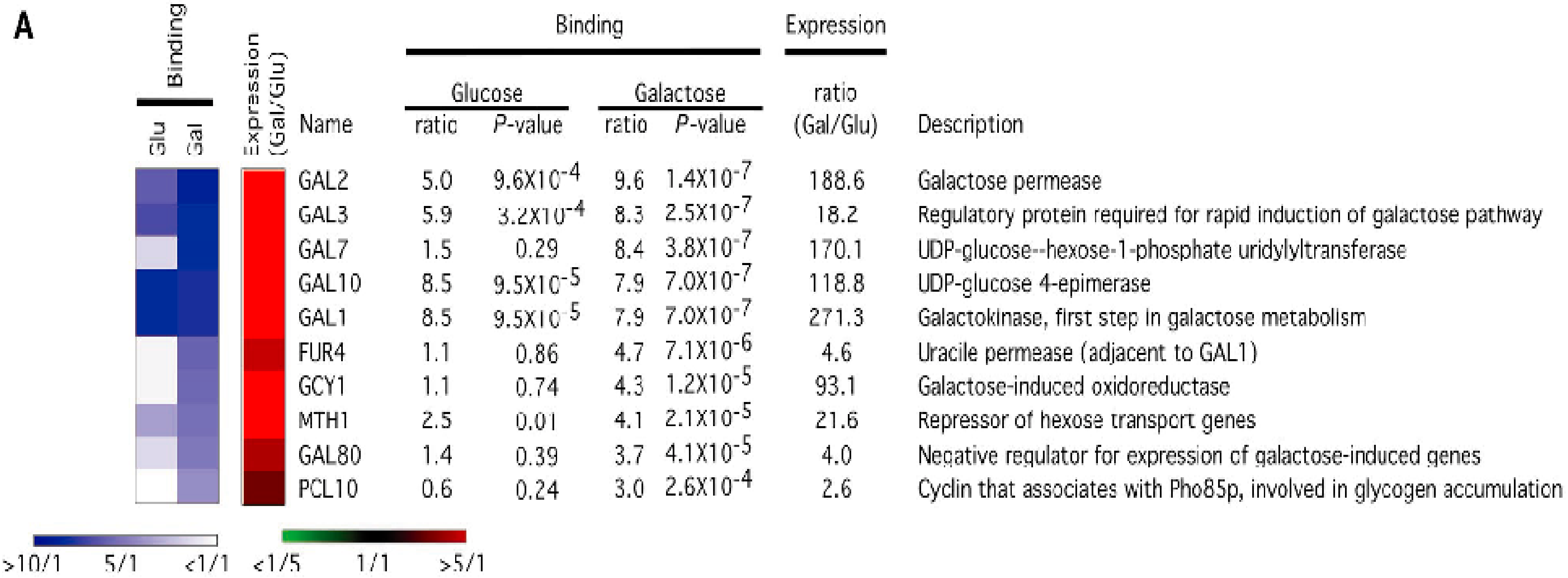
- Microarrays alone cannot distinguish *primary* effects from *secondary* effects
- Identification of where protein binds gives exact interaction



Example: Gal4 in yeast

- Gal4 activates the genes necessary for galactose metabolism
- Find genes which are upregulated in galactose *and* bound by Gal4
- 10 targets found with 0.001 ≥ P value

Results


A
Binding
Glu
Gal
Expression (Gal/Glu)
>10/1
5/1
<1/1
<1/5
1/1
>5/1

	Binding				Expression	
	Glucose		Galactose		ratio	
Name	ratio	*P*-value	ratio	*P*-value	(Gal/Glu)	Description
GAL2	5.0	9.6X10^{-4}	9.6	1.4X10^{-7}	188.6	Galactose permease
GAL3	5.9	3.2X10^{-4}	8.3	2.5X10^{-7}	18.2	Regulatory protein required for rapid induction of galactose pathway
GAL7	1.5	0.29	8.4	3.8X10^{-7}	170.1	UDP-glucose--hexose-1-phosphate uridylyltransferase
GAL10	8.5	9.5X10^{-5}	7.9	7.0X10^{-7}	118.8	UDP-glucose 4-epimerase
GAL1	8.5	9.5X10^{-5}	7.9	7.0X10^{-7}	271.3	Galactokinase, first step in galactose metabolism
FUR4	1.1	0.86	4.7	7.1X10^{-6}	4.6	Uracile permease (adjacent to GAL1)
GCY1	1.1	0.74	4.3	1.2X10^{-5}	93.1	Galactose-induced oxidoreductase
MTH1	2.5	0.01	4.1	2.1X10^{-5}	21.6	Repressor of hexose transport genes
GAL80	1.4	0.39	3.7	4.1X10^{-5}	4.0	Negative regulator for expression of galactose-induced genes
PCL10	0.6	0.24	3.0	2.6X10^{-4}	2.6	Cyclin that associates with Pho85p, involved in glycogen accumulation

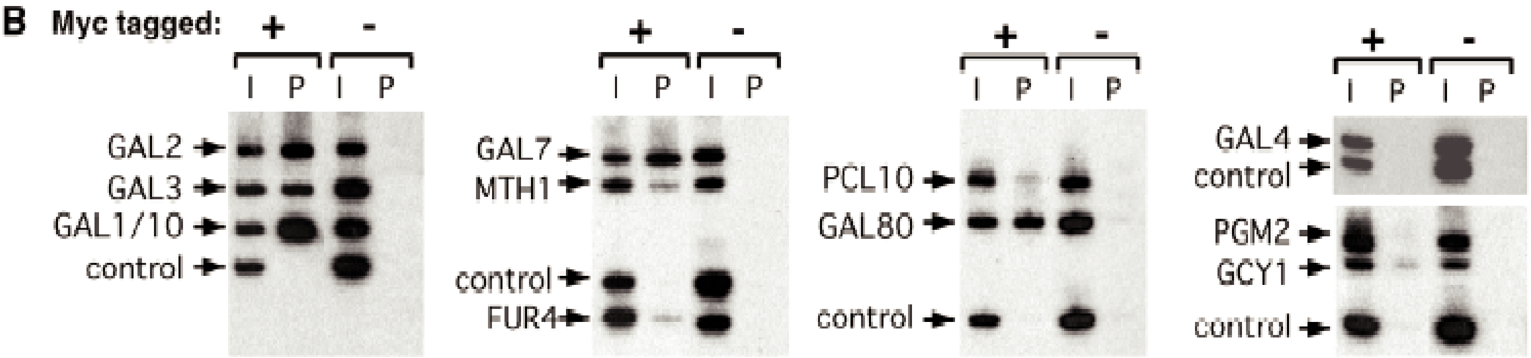
Results, Part II

- 7 genes known to be bound by Gal4
  - GAL1, GAL2, GAL3, GAL7, GAL10, GAL80, GCY1
- 3 others (MTH1, PCL10, FUR4)
- The consensus binding sequence ($CGGN_{11}CCG$) occurs in many other places

  "…sequence alone is not sufficient to account for specificity of Gal4 binding in vivo."

Confirmation

Standard Chromatin IP



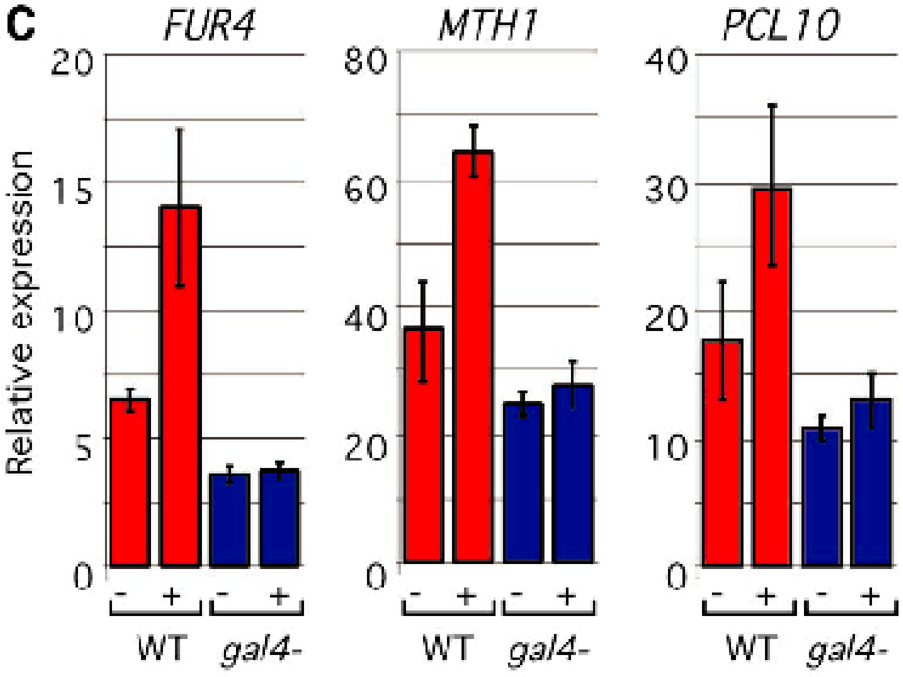
+/-	Strains with or without tagged Gal4

I/P	Unenriched or Enriched DNA

Additional confirmation

RT-PCR to quantify expression of 3 unexpected genes

Wild type vs gal4-

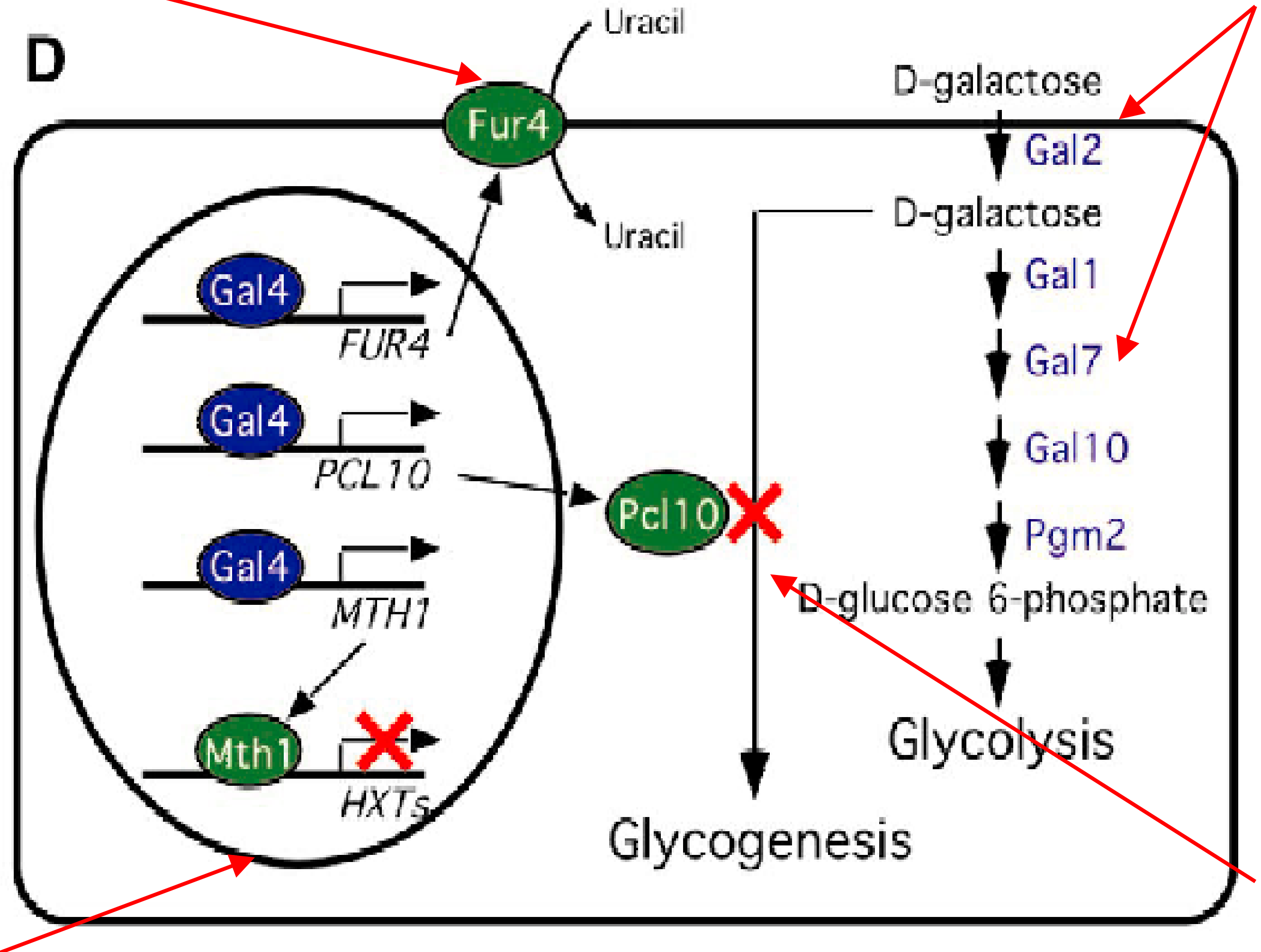


“Galactose-induced expression of FUR4, MTH1, and PCL10 is Gal-4 dependent”

Results, Part III

Increase pyrimadines for UDP (?)

Induce transport and conversion



Maximize energy from galactose

Reduce other transporters

Beyond transcriptional studies: New applications of microarrays

1. **Genome-wide screening for genes conferring specific cellular traits (Gill *et al., PNAS*)**
2. **Genome-wide location and function of DNA binding proteins (Ren *et al., Science,* December 22, 2000)**
3. **Experimental annotation of human genome (Shoemaker *et al., Nature* 409**:922-927, 2001; **Rosetta Inpharmatics)**

WHAT THE PAPER CLAIMS...

Microarray technology:

- High-throughput microarray based **experimental** method to validate predicted exons
- Group the exons into genes by co-regulated expression
- Define full-length mRNA transcripts

Approaches:

1 Exon array approach - *high-throughput method*

2 Tiling array approach - *higher-resolution method*

EXON ARRAY APPROACH

Analysis of human chromosome 22q

- Exon arrays consisting of long oligonucleotide probes derived from the 8,183 predicted exons annotated on chromosome 22q were fabricated
- Hybridization with flourescently labeled cDNAs derived from 69 experiments conducted on pairs of conditions

  (e.g., kidney vs. lung, testes vs uterus, etc.) using Cy3/Cy5 labels

FABRICATION OF EXON ARRAYS

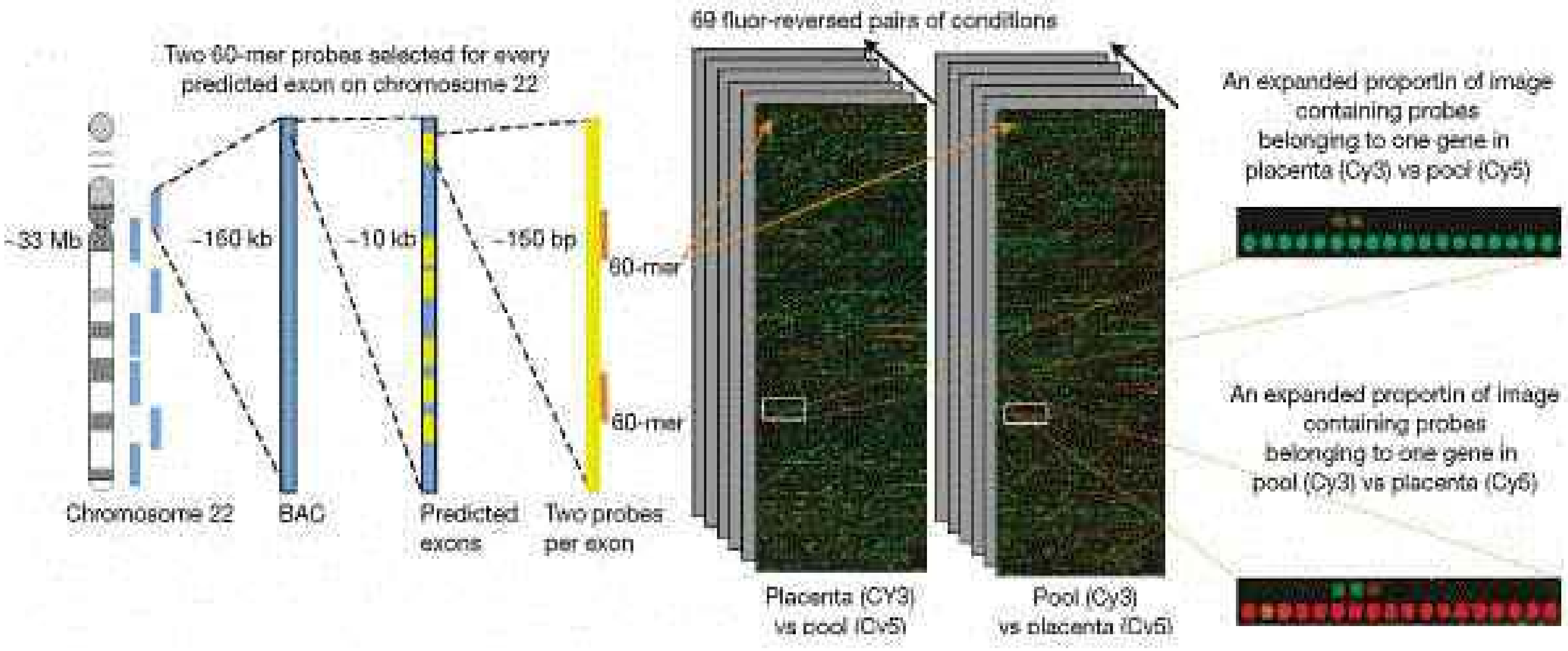
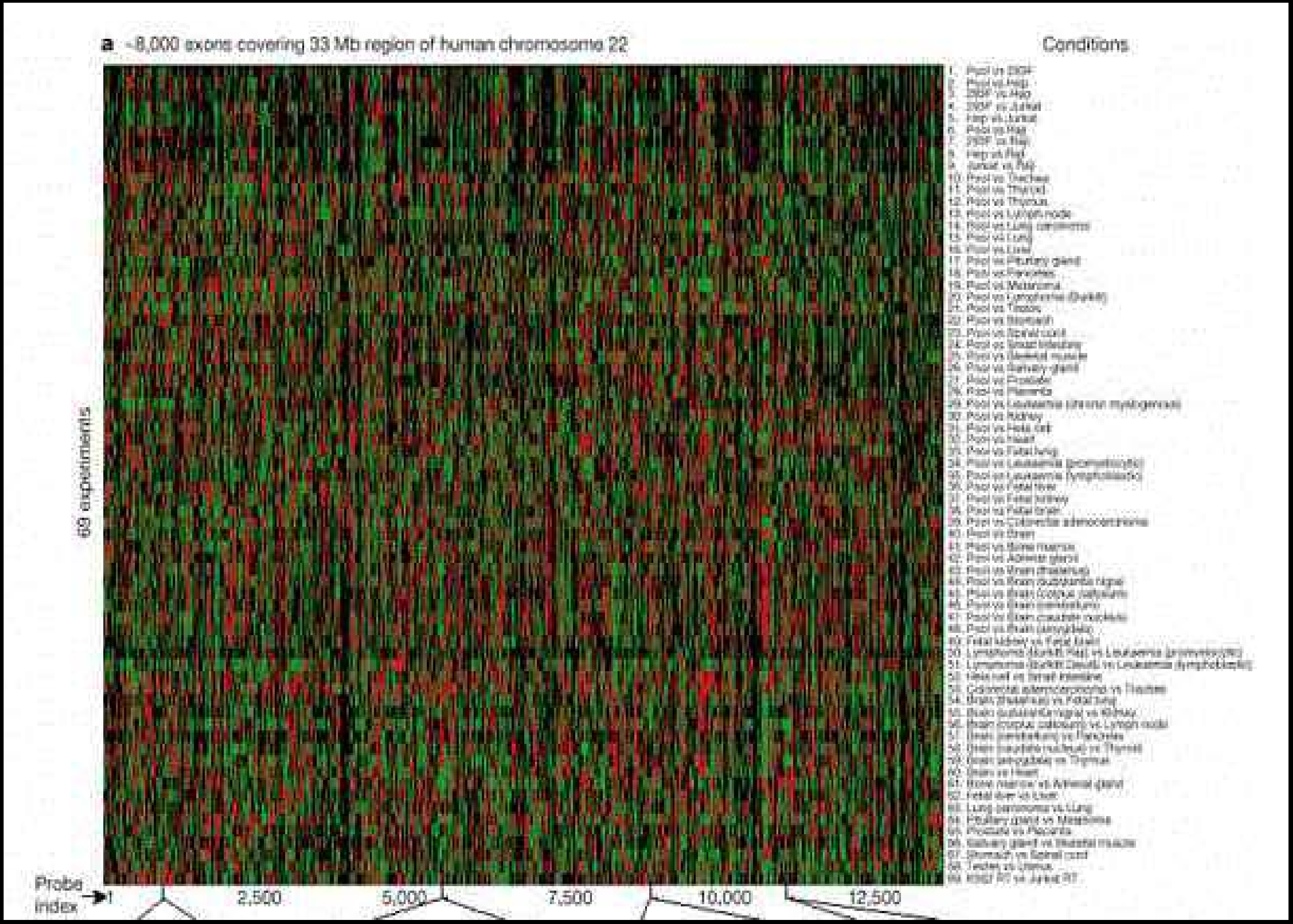
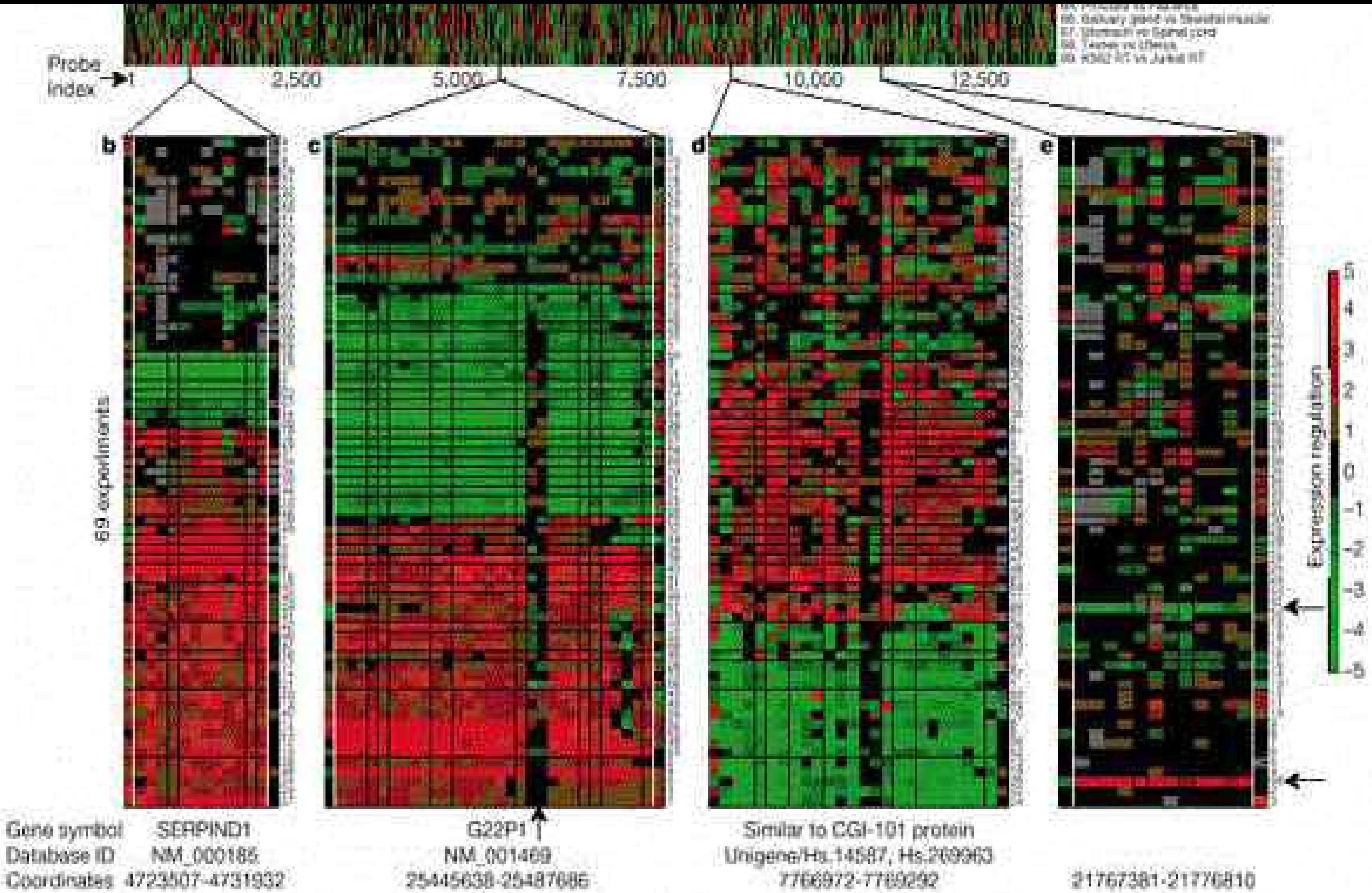

Two 60-mer probes selected for every predicted exon on chromosome 22
69 fluor-reversed pairs of conditions
An expanded proportin of image containing probes belonging to one gene in placenta (Cy3) vs pool (Cy5)
~33 Mb
~160 kb
~10 kb
~150 bp
60-mer
60-mer
Chromosome 22
BAC
Predicted exons
Two probes per exon
An expanded proportin of image containing probes belonging to one gene in pool (Cy3) vs placenta (Cy5)
Placenta (CY3) vs pool (Cy5)
Pool (Cy3) vs placenta (Cy5)

Figure 1 Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).

COMPENDIUM OF EXPERIMENTS



VALIDATION OF GENE BOUNDARIES



SUMMARY OF RESULTS

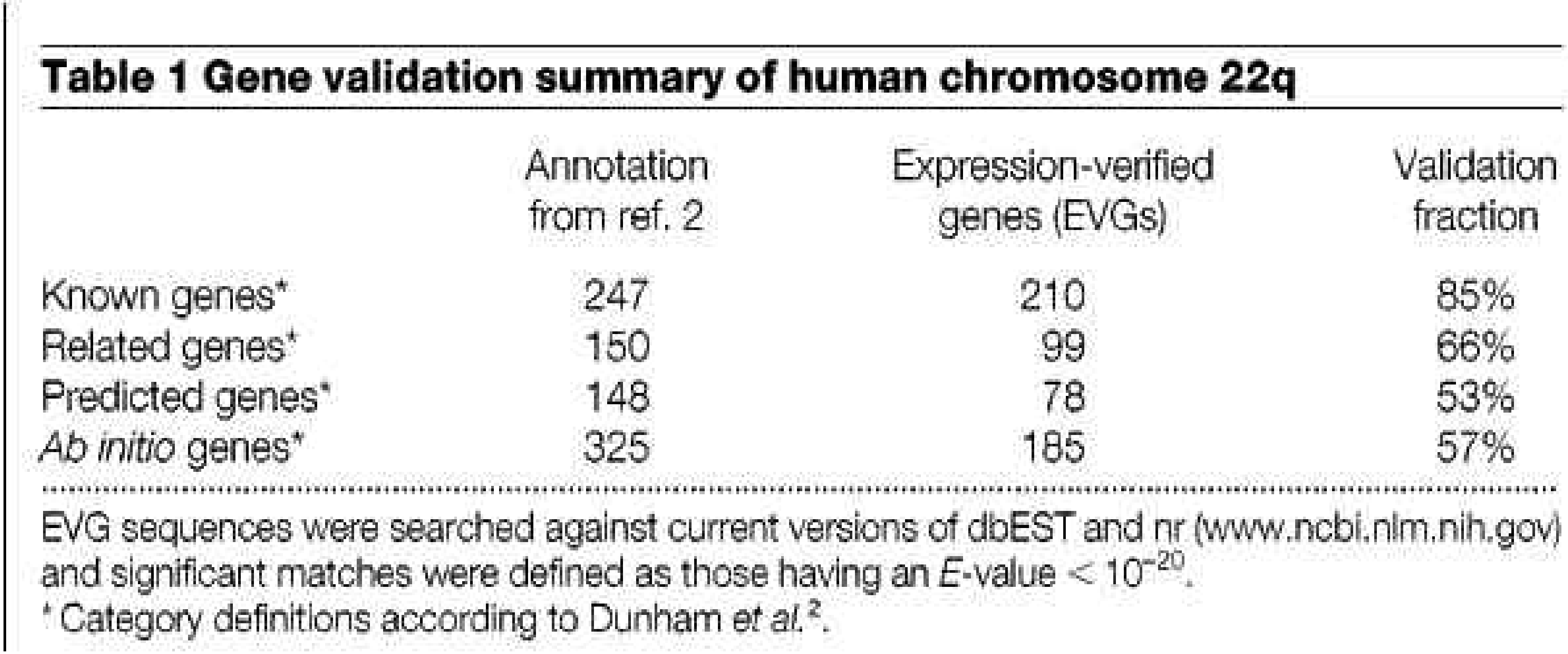


Table 1 Gene validation summary of human chromosome 22q

	Annotation from ref. 2	Expression-verified genes (EVGs)	Validation fraction
Known genes*	247	210	85%
Related genes*	150	99	66%
Predicted genes*	148	78	53%
Ab initio genes*	325	185	57%

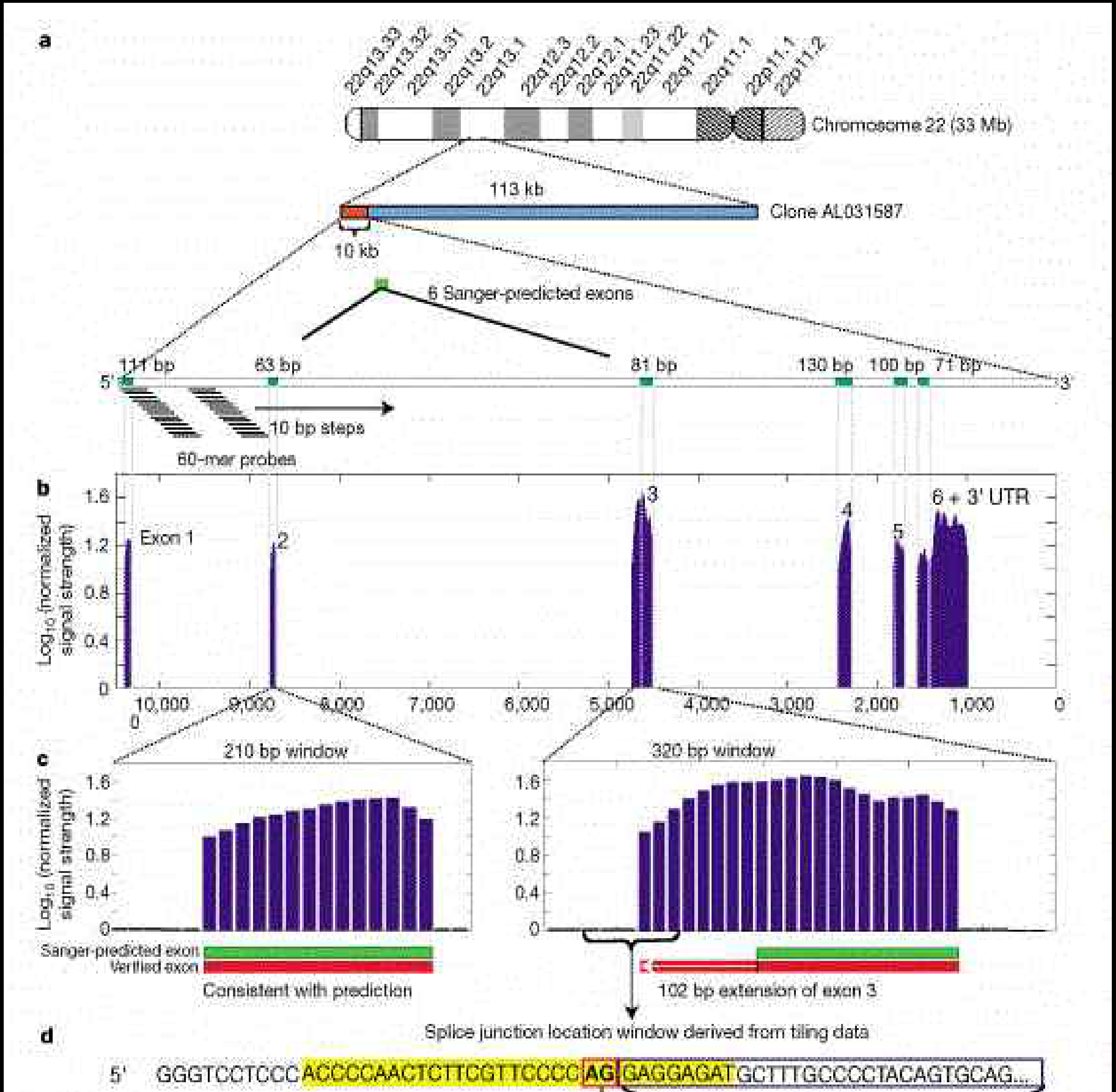
EVG sequences were searched against current versions of dbEST and nr (www.ncbi.nlm.nih.gov) and significant matches were defined as those having an *E*-value $< 10^{-20}$.

* Category definitions according to Dunham *et al.*[2].

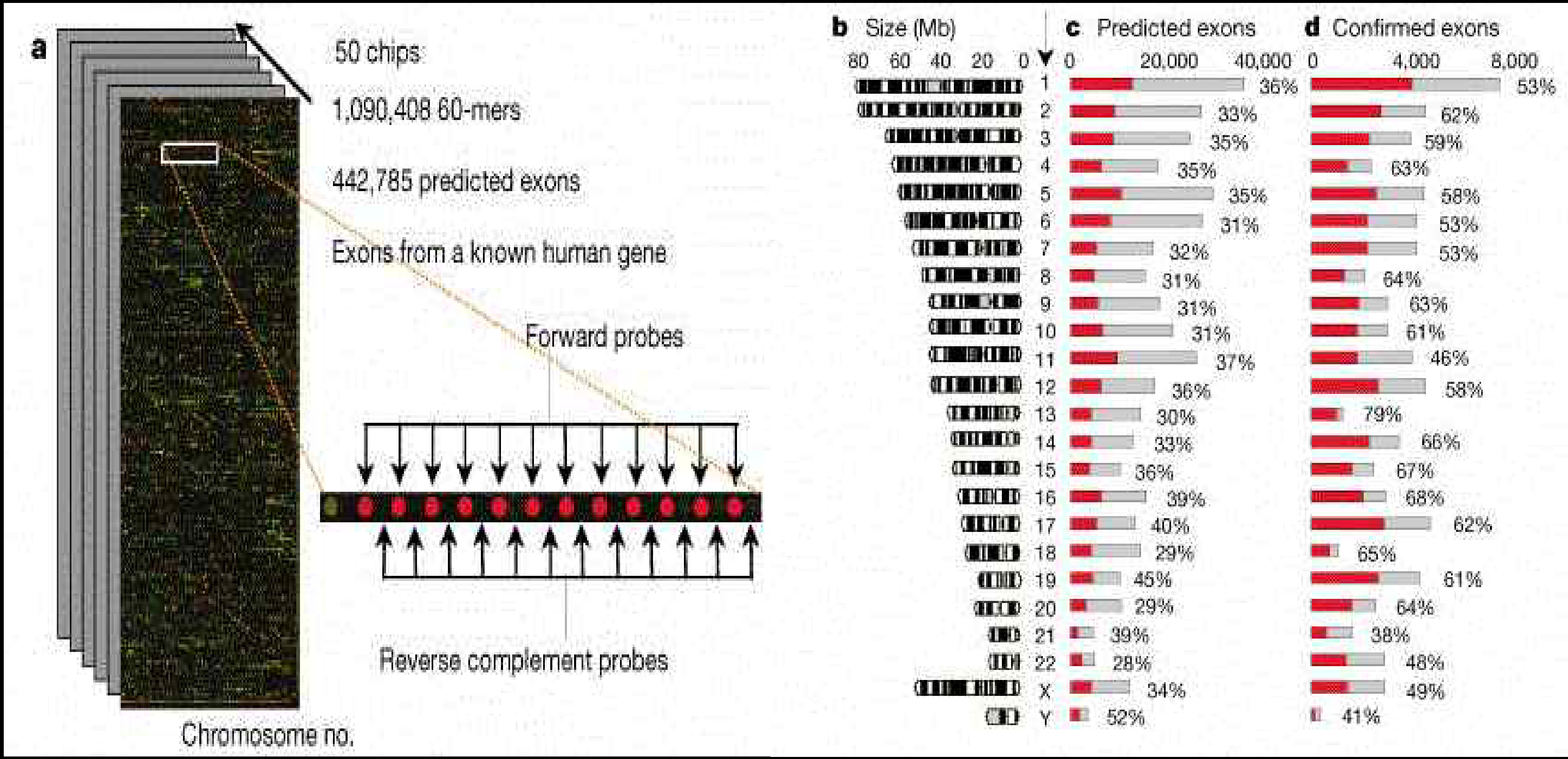
TILING ARRAY APPROACH

- Higher-resolution view of a genomic region of interest
- Can potentially reveal exons not identified by current gene prediction algorithms
- Provides information about alternative splicing
- Can be applied specifically on initial and terminal exons where gene prediction programs are not very accurate
- Does not need any *a priori* information on exon content of genomic sequence

TESTES TRANSCRIPT REFINED USING TILING ARRAYS



HUMAN GENOME SCAN USING EXON ARRAYS



DISCUSSION

- A microarray approach to the simultaneous validation of gene predictions and study of the transcriptome under any number of medically interestingly conditions
- Exon-based approach provides high-throughput screening of diverse cell types, growth conditions and disease states
- Method could be used for rapid annotation of sequence information from the Human Genome project

Beyond transcriptional studies: New applications of microarrays

1. Genome-wide screening for genes conferring specific cellular traits
2. Genome-wide location and function of DNA binding proteins (Ren *et al., Science,* December 22, 2000)
3. Experimental annotation of human genome
4. Large scale identification of secreted and membrane associated gene products (Diehn *et al., Nature Genetics,* 25:58-62 2000)

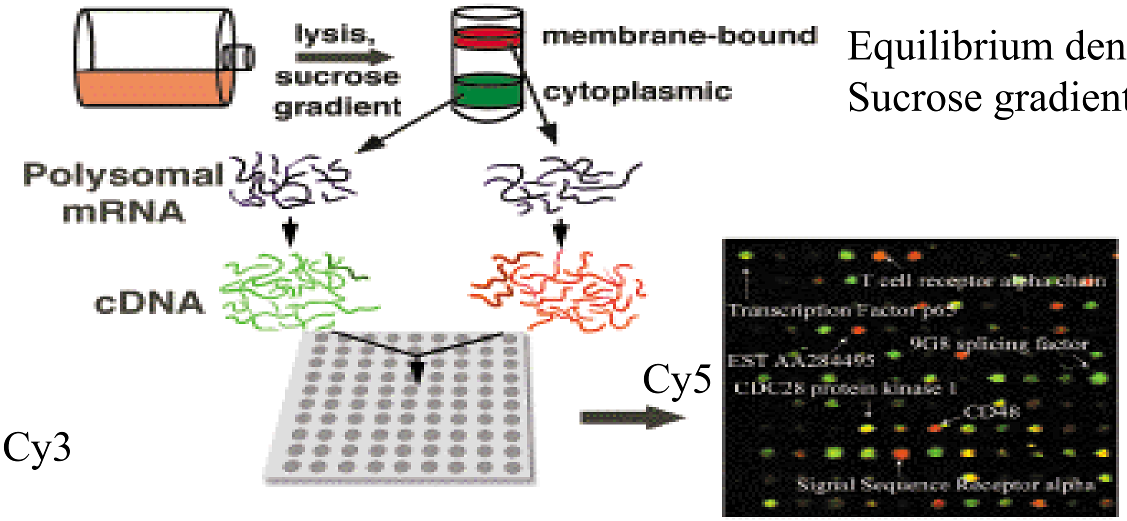
Motivation

- Importance of membrane-associated and secreted proteins
  - Receptors
  - Transporters
  - Adhesion molecules
  - Hormones
  - Cytokines

Current Methods

- Computational
  - Potential amino-terminal membrane-targeting signals
  - Trans-membrane domains
    - Require knowledge of entire coding sequence
- Experimental
  - No large-scale genomic methods

Method for isolating membrane-bound polysomes



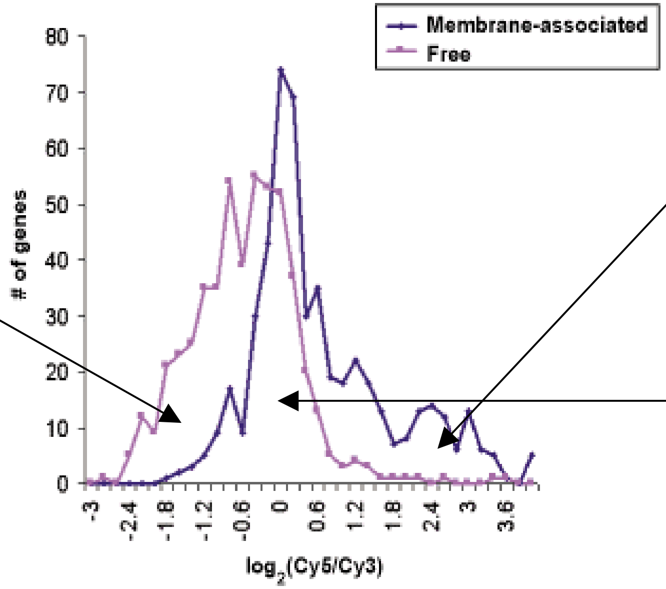
Equilibrium density centrifugation in a Sucrose gradient

Cy5

Cy3

Figure 1. Procedure for isolating membrane-bound polysomes from cell lines. Jurkat cells were hypotonically lysed, and membrane-bound RNA was separated from free RNA by equilibrium density centrifugation in a sucrose gradient. Total RNA was isolated separately from the fractions containing membrane-bound or free RNA. After amplification of mRNA, cDNA was synthesized from membrane-bound and free mRNA and labeled with the fluorescent dyes Cy5 and Cy3, respectively. The cDNAs were hybridized to a DNA microarray and analyzed using standard methodology. The subsection of an array pictured shows the identity of some of the spots from a representative experiment.

Evaluation for genes with known localization



Distinctly cytosolic

Distinctly membrane-bound

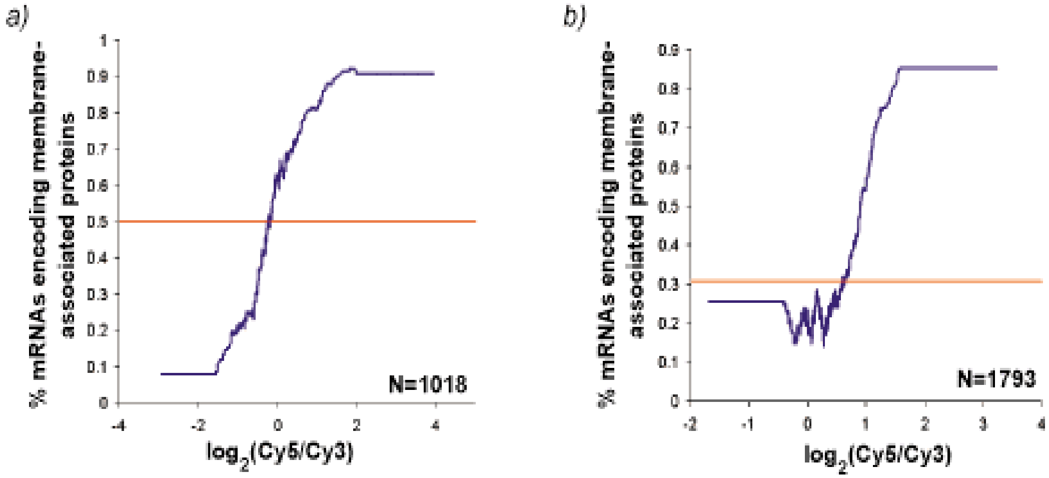
Overlap:

- Confirmed in other organisms and from literature
- No simple classification, biological explanation

Figure 2. Distribution of Jurkat T cell mRNAs coding for proteins with characterized subcellular localization. Information on the empirically determined subcellular localization of proteins was retrieved from the SWISS-PROT database and the literature. Genes were classified into two categories: those whose protein products are membrane-associated (transmembrane, secreted, or ER/Golgi/Vesicle resident) (blue curve) and those whose products are cytosolic (or nuclear) (red curve). The graphs show the number of genes in each class plotted against the log-transformed (base 2) Cy5/Cy3 ratios in bins of 0.2.

Characterization of unclassified genes

Compare with the already characterized genes, in the following graphs.
Fraction = probability



Human Cells (Jurkat)
Window size: 151 genes

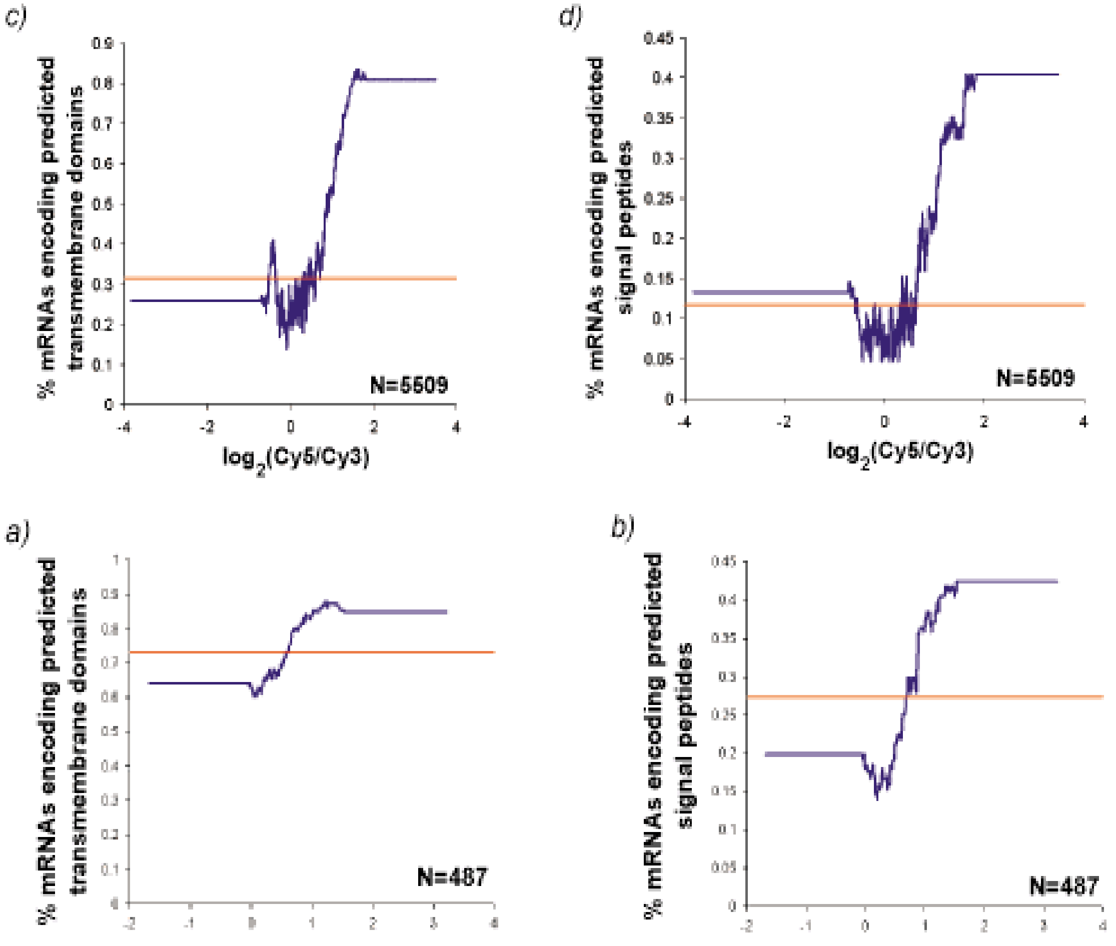
Yeast Cells
Window size: 175 genes

Moving Average Methodology

- Sort genes by the ratio value
- Compute fraction of the genes known to be membrane-associated in a given window of specified size
- Plot this fraction against the expression of the central gene in window

Comparison with computational predictions

Only for Yeast genes, since fully sequenced



Genes enriched in the membrane fraction

Also identify several genes that are membrane-associated but were not computationally predicted

Misclassifications/Exceptions

- Cytosolic protein in expt. membrane fraction
  - HAC1: spliced in cytoplasm by interaction with Ire1p, an ER transmembrane protein
- Cytosolic protein with membrane binding domain
  - Calcineurin B: associated to cytoplasmic side of plasma membrane. Ribosomes may be recruited to PM
- Alternative splice sites
  - The present microarray cannot distinguish between the various forms

Analysis of Microarray Data

1. Internal Validation

- Background
- Normalization
- Confidence intervals

2. Analysis of *Static* expression data

- PCA, Decision trees

3. Analysis of *Dynamic* expression data

- Clustering (Hierarchical, SOM's)

Analysis of Microarray Data

1. Internal Validation

RNA Diagnostics (yield, purity)

Average Yield: 3.5 +/- 0.8 ug RNA/10^8 cells (Cyanobacteria)

Average Purity: $A_{260/280}$ = 1.85 +/- 0.1

Data Filtering (Removes precipitants, ghost spots, weak signals)

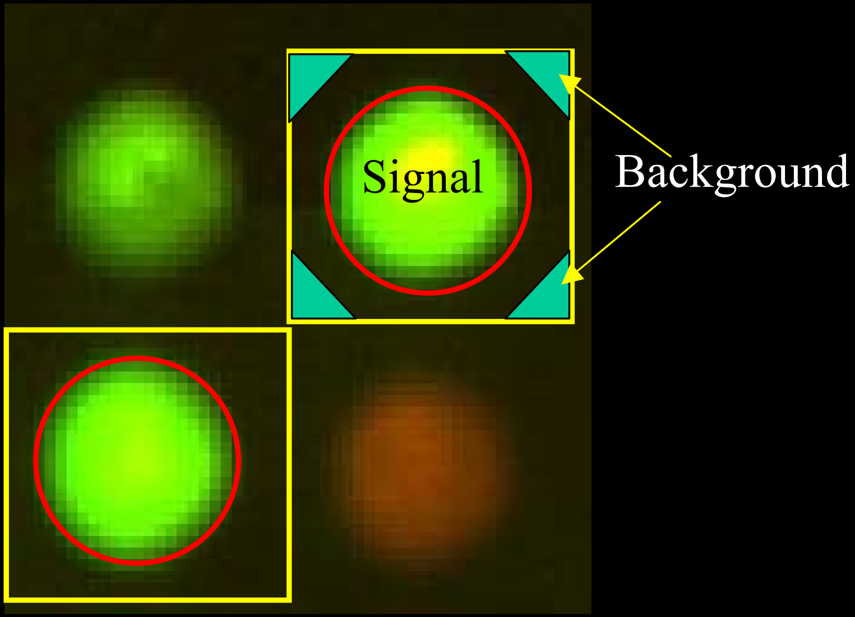
Filter 1: $(\text{Background}_{\text{local}}-\text{Background}_{\text{avg}}) > x\text{SD}_{\text{background}}$

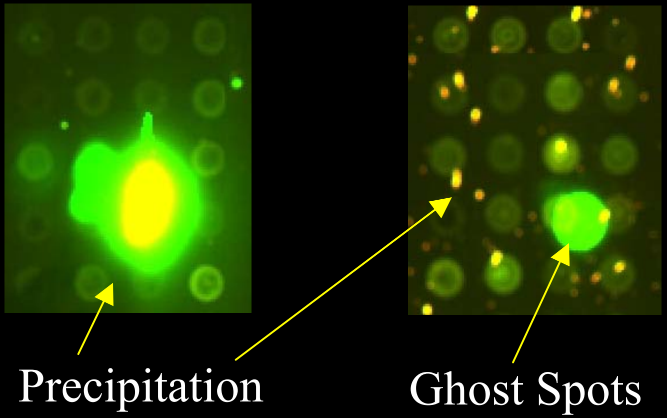
Filter 2: $(\text{Signal}_{\text{local}} - \text{Background}_{\text{local}}) < y\text{SD}_{\text{B'local}}$

Filter1 **$x(\text{SD}_{\text{avg}})$**	**Filter2** **$y(\text{SD}_{\text{B'local}})$**	**S/N Cy3**	**S/N Cy5**	**% Spots** **Retained**
1	0	4.1	2.2	98%
3	0.5	5.3	2.8	91%
1	1	7.8	3.8	75%
2	2	10.3	4.5	57%
0.5	2	6.7	3.4	69%

Internal Validation of Micro-Arraying Methodology

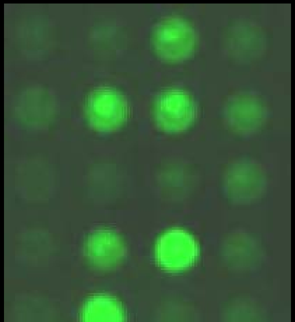
Filtering

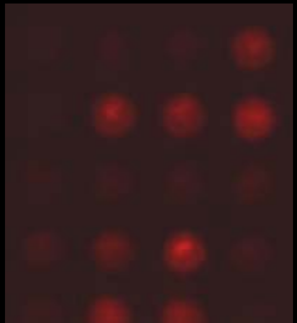




Adjustment

	Cy3	Cy5
Labeling	112 NT/Cy3	293 NT/Cy5
Brightness (Ex*QY)	57,000	70,000





Signal/Noise: S/N= $(\text{Signal}_i - \text{Background}_i)/\sigma_{\text{back}}$

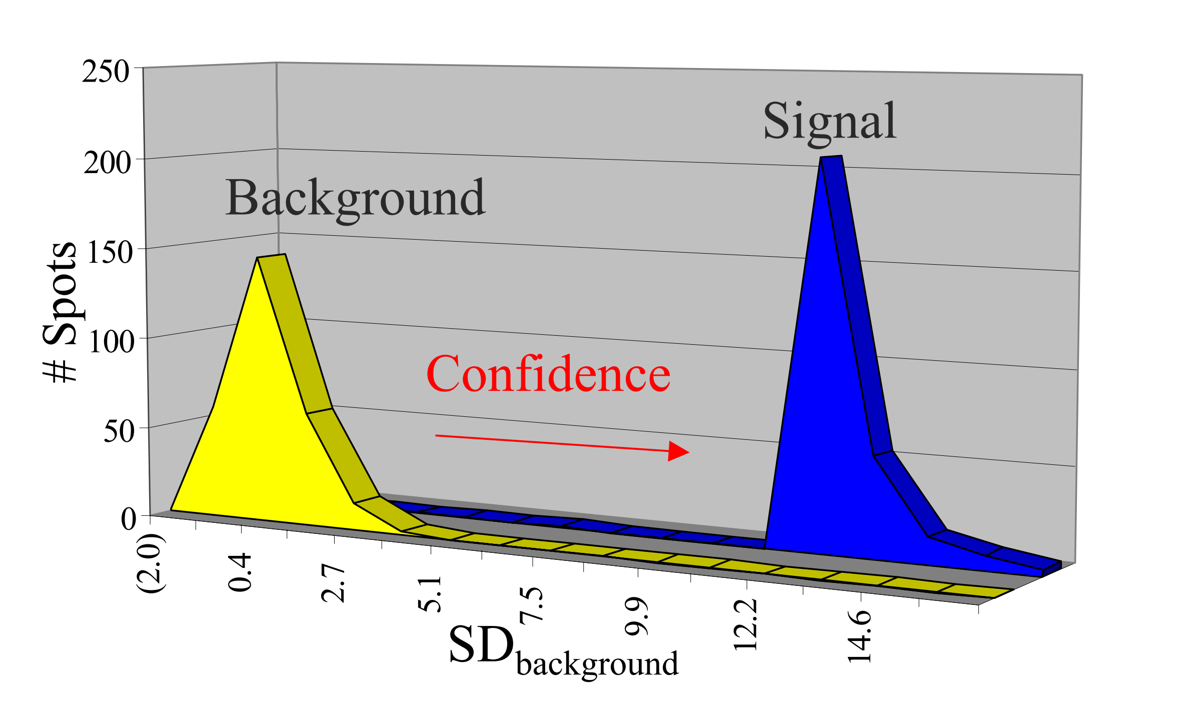
For S/N > 1.96 the signal is diferent than the background with 95% confidence interval (CI)

Labeling Diagnostics

	Extinction Coeff. (EC)	Quantum Yield (QY)	Brightness (EC*QY)	Avg. Incorp’n (NT/Cy)
Cy3:	150,000	0.38	57,000	112 +/- 47
Cy5:	250,000	0.28	70,000	293 +/-154

	Ratio NT_{Cy5}/NT_{Cy3}
Molar	**2.62**
Brightness	**3.21**

Internal Validation of Micro-Arraying Methodology



How Repeatable are Our Results?

CV = 30-45%

95% Confidence Interval = 1.6-1.9 = $\frac{\text{Intensity Sample}}{\text{Intensity Control}}$

Normalization:

1. By the average or total signal for each fluorophore (accounts for variations in brightness and total RNA)
   - Values now are reported as *fractions of total RNA*. This may change under certain conditions (partial arrays)

2. By the average of rRNA genes. They can change too

3. Backgrounds: B_{Cy3} / B_{Cy5}

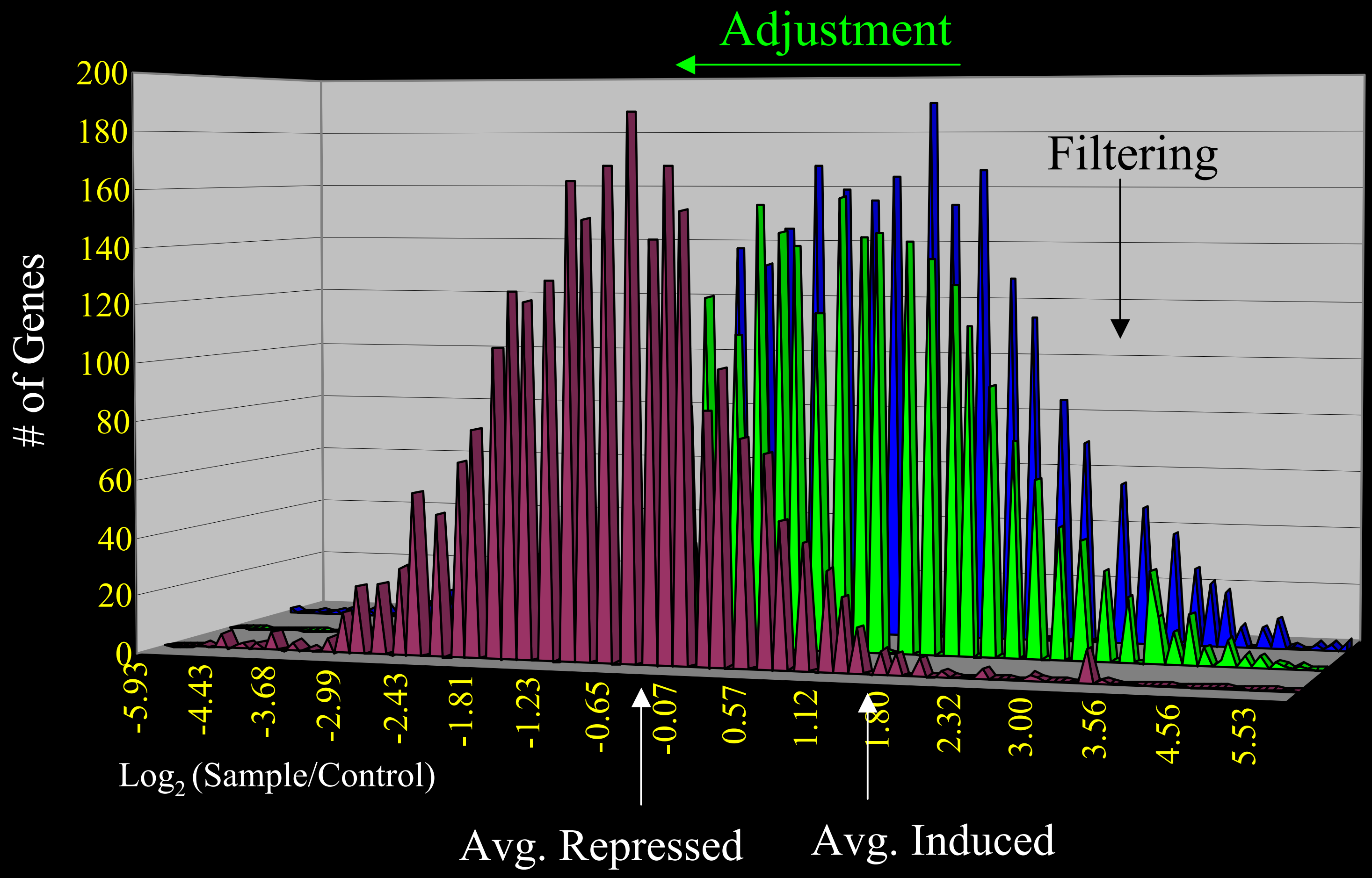
4. Specific brightness: BR_{Cy3} / BR_{Cy5}

5. Total signal: S_{cy3} / S_{Cy5}

6. Total signal minus background: $((S-B)_3$ / $(S-B)_5)$

Is class discovery dependent upon filtering and normalization?

Internal Validation of Micro-Arraying Methodology



🟪 Filtered then Adjusted by Average Signal

🟩 Filtered then Adjusted by Background and Brightness

🟦 No Filter, Adjusted by Background and Brightness

Analysis of Microarray Data

1. Internal Validation
   - Background
   - Normalization
   - Confidence intervals

2. Analysis of *Static* expression data
   - PCA, Decision trees

Key Questions

- Sample classification-*Diagnosis*
- Identification of discriminatory genes

3. Analysis of *Dynamic* expression data
   - Clustering (Hierarchical, SOM's)

Data Analysis and Pattern Classification

Problem-1: Consider N samples and M genes with their corresponding expression levels, e_i , where i = 1, …, M. M_1 of these tissues are characterized as "Healthy", while the other M_2 are labeled as "Pathological". ***Find the set of discriminatory genes*** whose expression levels can diagnose the state, i.e. healthy or pathological, of a new sample tissue.

Feature Space: The space of expression levels for the M genes, i.e. $FS = \{e_1 , e_2 , e_3 , \ldots, e_{M-1} , e_M \}$

Class: A set of genes characterized by the same label, e.g. C_1 = "Healthy" and C_2 = "Pathological".

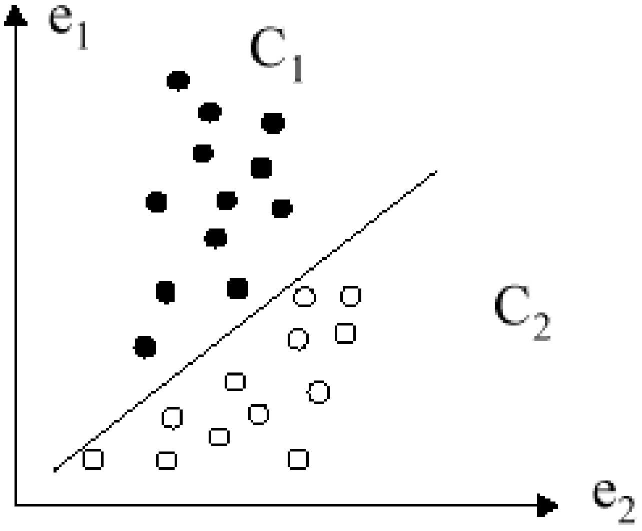
Pattern: The specific M-tuple of expression levels, which characterizes a tissue as belonging to a specific class, i.e. $\mathbf{p}^{(2)} = \{e^{(2)}_1 , e^{(2)}_2 , e^{(2)}_3 , \ldots, e^{(2)}_{M-1} , e^{(2)}_M \}$, Pattern for "Pathological" Tissues.

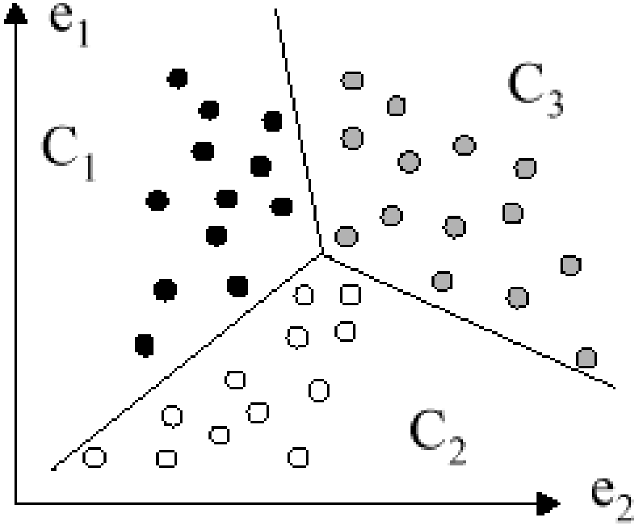
Data Analysis and Pattern Classification

Pattern Classification: The process through which the feature space, FS, is partitioned into K exclusive regions, FS_i i = 1, 2, …, K. Thus,

$$FS^{(i)} \cap FS^{(j)} = 0 \quad \text{and} \quad \cup_{i=1-K} FS^{(i)} = FS$$

Discriminant Functions: $d\,(\mathbf{p}) = d\,(e_1 , e_2 , e_3 , \ldots, e_{M-1} , e_M)$ define the partition of the feature space into the K regions.



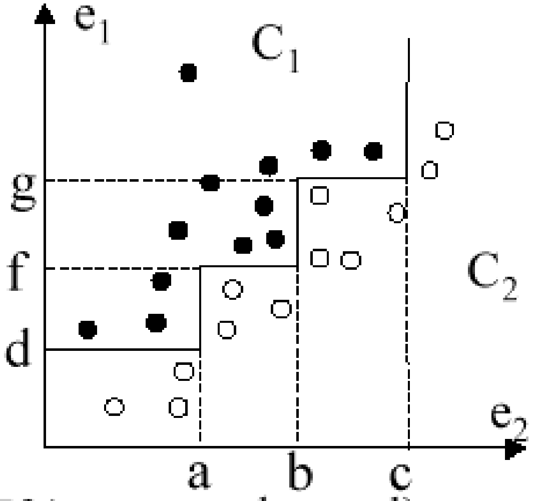


Data Analysis and Pattern Classification

Approaches: Stochastic or Deterministic

Logic-Based:

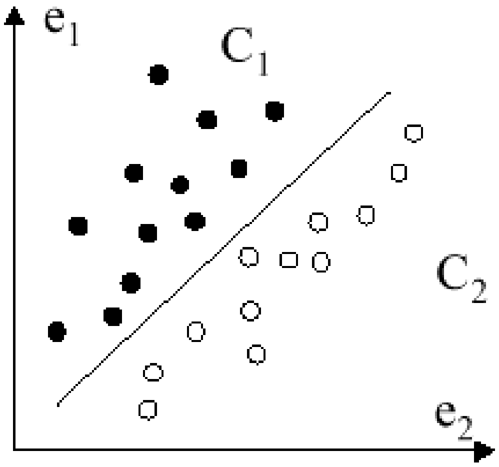
e.g. Decision Trees.



If ($e_2 < a$ and $e_1 < d$) or
($a < e_2 < b$ and $d < e_1 < f$),…
Then Class C_2

Linear:

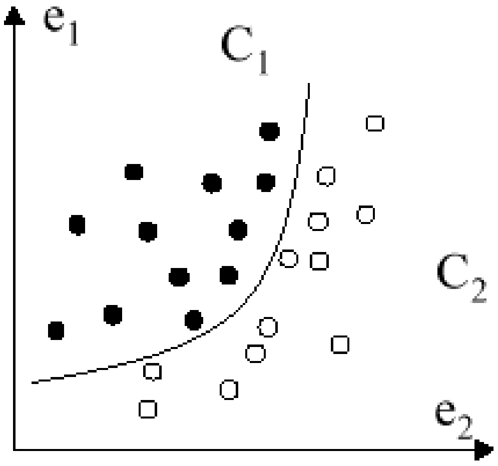
e.g. PCA



If = $k\ e_1 + l\ e_2 > 0$
Then Class C_1

Non-Linear:

e.g. Neural Networks



If = $f(e_1,e_2) > 0$
Then Class C_1

Data Analysis and Pattern Discovery

- **Problem-2:** Consider N samples and M genes with their corresponding expression levels, e_i , where i = 1, …, M. "<u>Discover" the patterns</u> in gene expression levels which are common in a number of samples, i.e. ***find the groups of samples***, each of which is characterized by a common pattern in gene expression and define this common pattern of gene expression levels for each group of samples.

- **Problem-3:** Consider one type of sample and the gene expression levels for M genes over a period of L time points. "<u>Discover" the patterns</u> in gene expression levels, which are common for a particular group of genes, and ***cluster the genes*** with similar patterns into the same group.

Data Analysis and Pattern Classification

Training:

The process through which one determines the discriminant functions, using past examples of "pattern" -"class" associations, i.e. associations between pattern $\mathbf{p}^{(i)} = \{e^{(i)}_1 , e^{(i)}_2 , e^{(i)}_3 , \ldots, e^{(i)}_{M-1} , e^{(i)}_M \}$ and Class $C^{(i)}$

Types of Problems:

- Static: when the gene expression levels represent the expression at a single time.
- Dynamic, or Time-Dependent: when the expression levels are measured over a period of time at various time intervals.
  - Equal sampling intervals.
  - Unequal sampling intervals.

Data Analysis and Pattern Classification

- **Issues to Resolve:**
  - Labeling the various samples
  - *Representation:*Selecting the distinguishing features for classification; particularly important for time-dependent data, e.g. do you use the values, or the time derivatives of expression levels for classification?
  - Selecting the form of the discriminant function
  - Do you have statistically "enough" data for training?
  - Do you have enough data for testing?
  - What is the "noise" in your measurements?
  - What is the sensitivity of the generated discriminant function?
  - What is the robustness of the resulting classification scheme?

Information Theory:Decision Trees in Pattern Classification

Let N be the total number of examples (e.g. samples) and M_i the number of samples in each of the K classes.
The Shannon entropy provides a measure of the information content in the data set,

$$I(M_1, M_2, \ldots, M_K) = \Sigma_{i=1-K} (M_i/M) \log_2 (M_i/M)$$

- If all examples belong in the same class then *I = 0.*
- The smaller the entropy the less variety of classes (more order) in the data set.

Split the data into two groups G_1 and G_2 with $M^{(1)}$ and $M^{(2)}$ examples (samples) in each group. Compute the information content for each group and for the whole set of examples.

Decision Trees in Pattern Classification

$$I(M_1, M_2, \ldots, M_K) = I^{(1)}(M_1^{(1)}, M_2^{(1)}, \ldots, M_K^{(1)}) + I(M_1^{(2)}, M_2^{(2)}, \ldots, M_K^{(2)})$$

$$= \sum_{i=1}^{K} \frac{M_i^{(1)}}{M^{(1)}} \log_2\left(\frac{M_i^{(1)}}{M^{(1)}}\right) + \sum_{i=1}^{K} \frac{M_i^{(2)}}{M^{(1)}} \log_2\left(\frac{M_i^{(2)}}{M^{(2)}}\right)$$

If all the examples in group G_1 belong to class C_1 and all the examples in group G_2 belong to the class C_2, then,

$$M_1^{(1)} = M^{(1)} \;\; and \;\; M_2^{(1)} = M_3^{(1)} = \ldots = M_K^{(1)} = 0$$

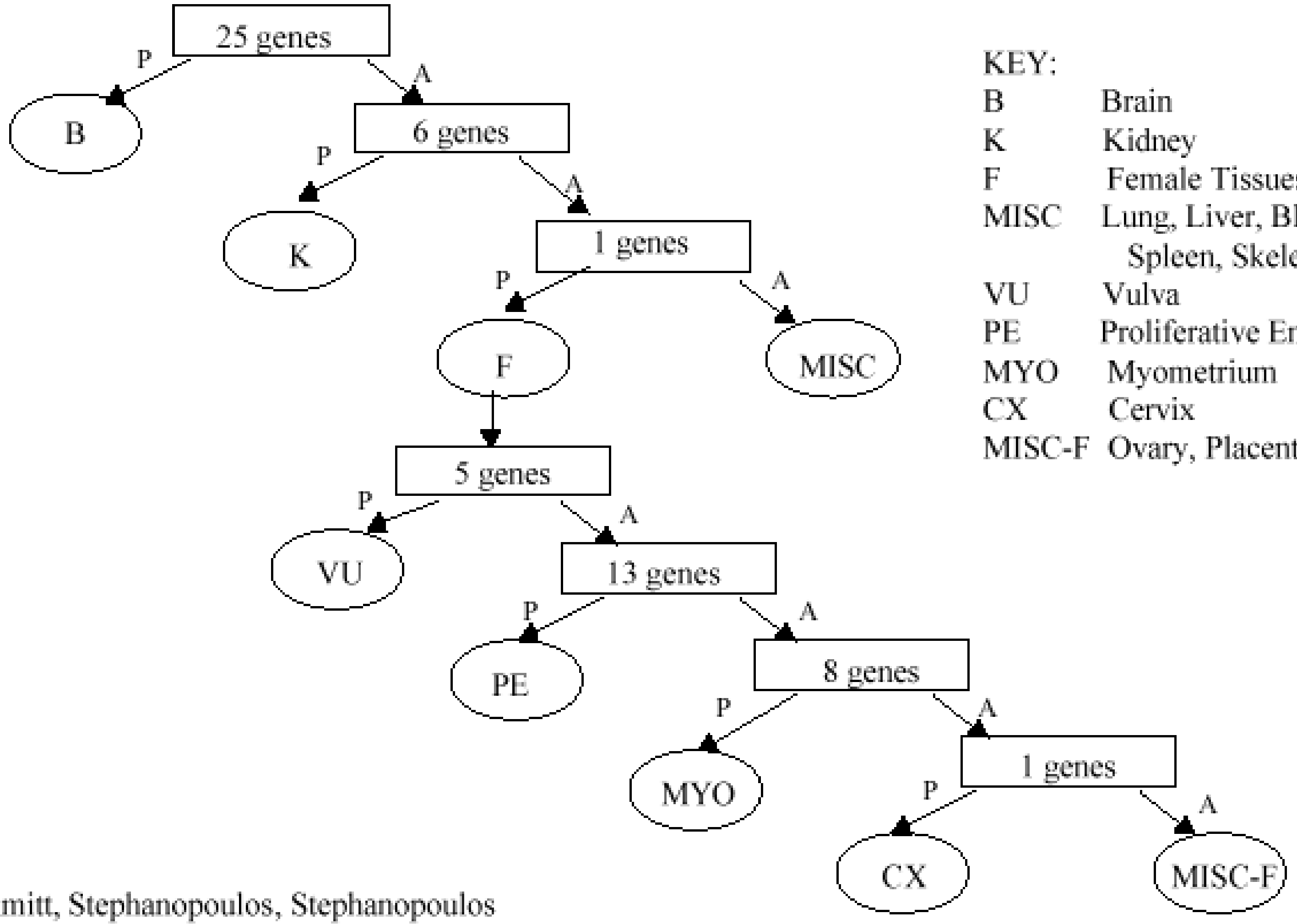
$$and$$

$$M_2^{(2)} = M^{(2)} \;\; and \;\; M_1^{(2)} = M_3^{(2)} = \ldots = M_K^{(2)} = 0$$

and $I^{(1)} = I^{(2)} = 0$, leading to the total $I = 0$.

Therefore, find the genes and their expression levels, which if were used to group the tissues into the K classes would “Minimize *I* “

Discriminating Tree for the Tissues



KEY:

B	Brain
K	Kidney
F	Female Tissues
MISC	Lung, Liver, Blood, Spleen, Skeletal Muscle
VU	Vulva
PE	Proliferative Endometrium
MYO	Myometrium
CX	Cervix
MISC-F	Ovary, Placenta

Misra, Schmitt, Stephanopoulos, Stephanopoulos
BioInformatics and Metabolic Engineering Laboratory,
MIT, 12/3/99

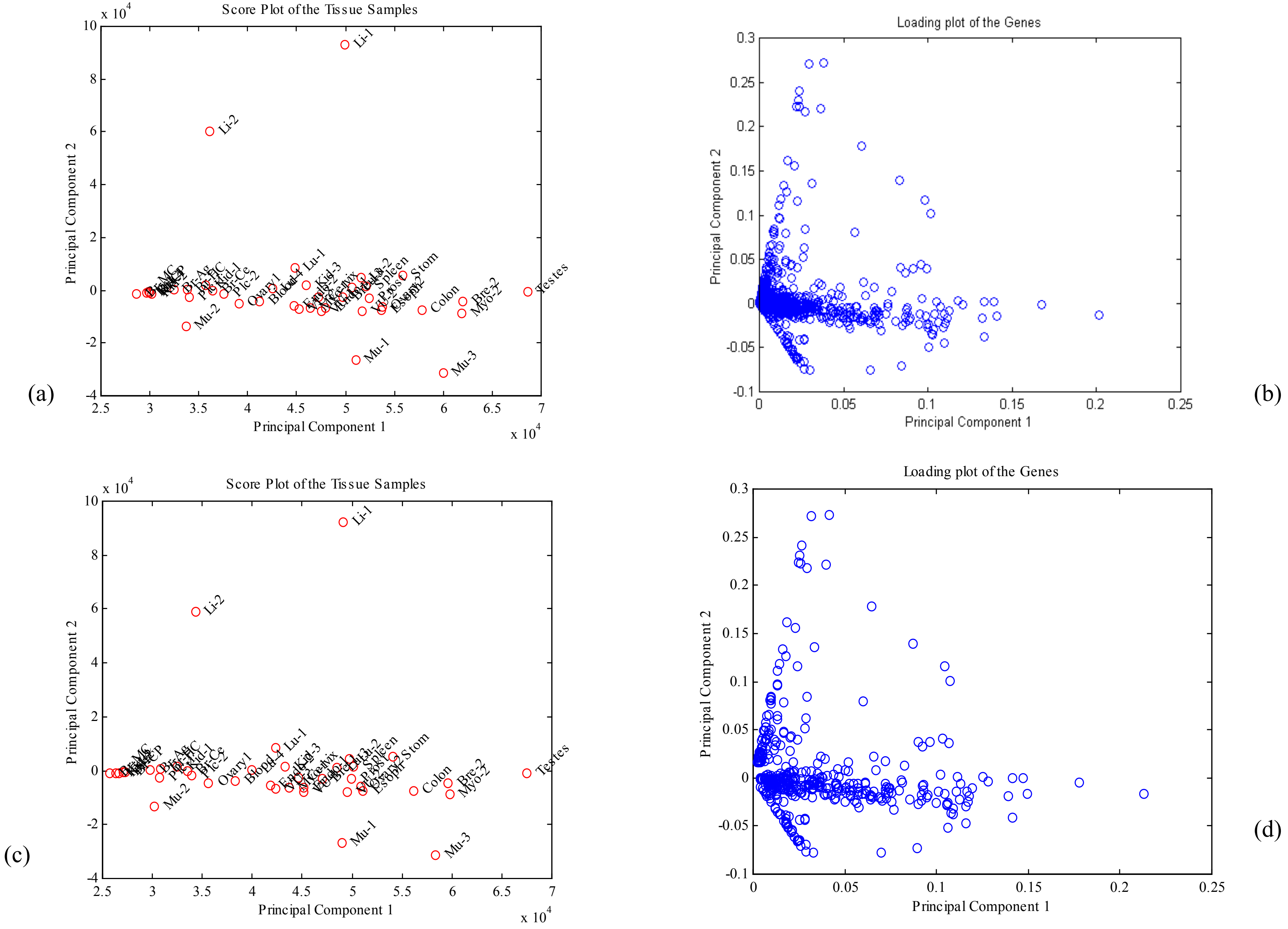


Figure 1: Selection of relevant genes using the loadings on the principal components.

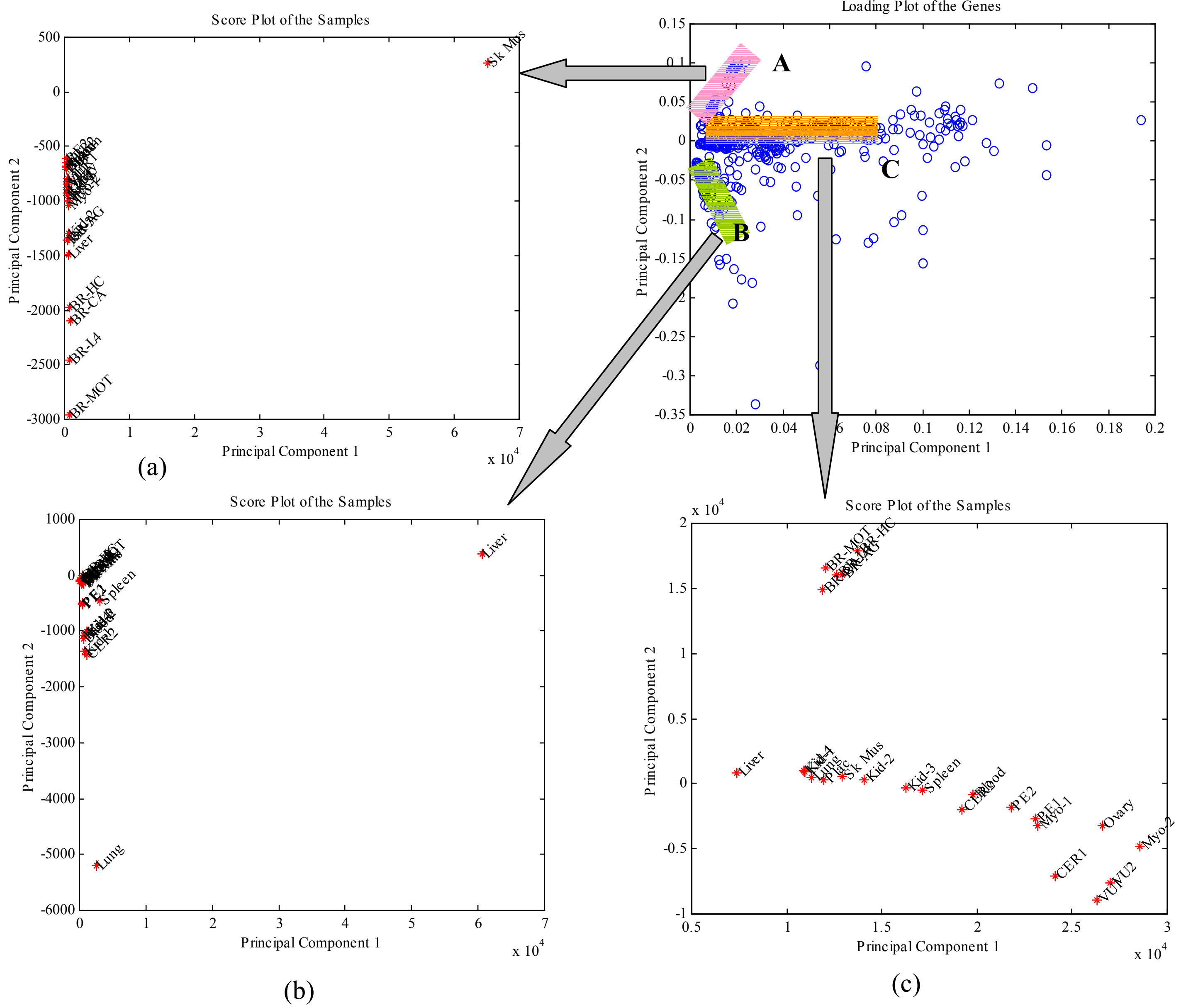
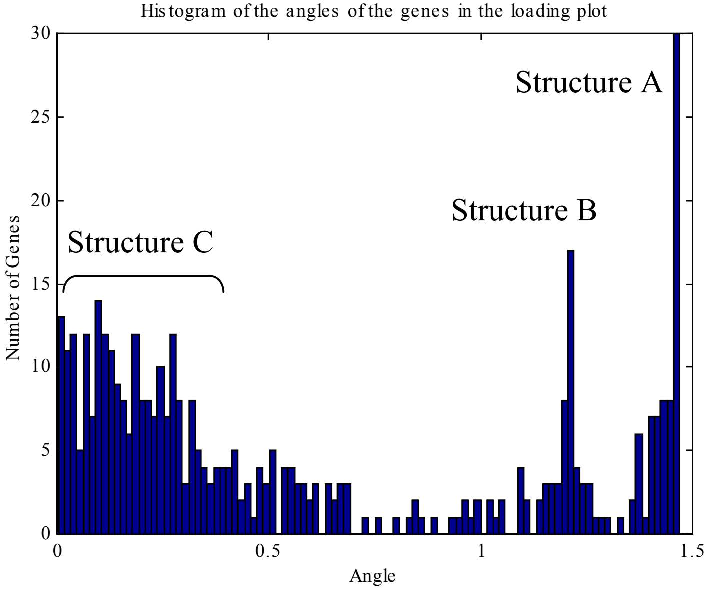
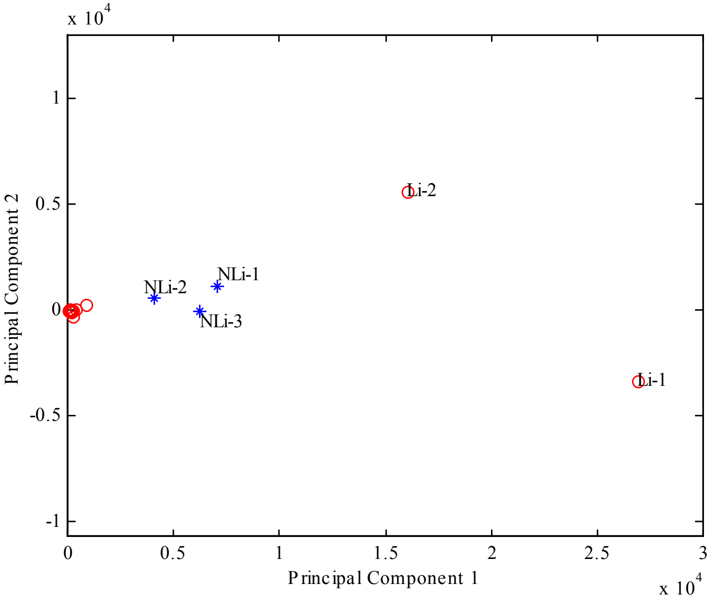
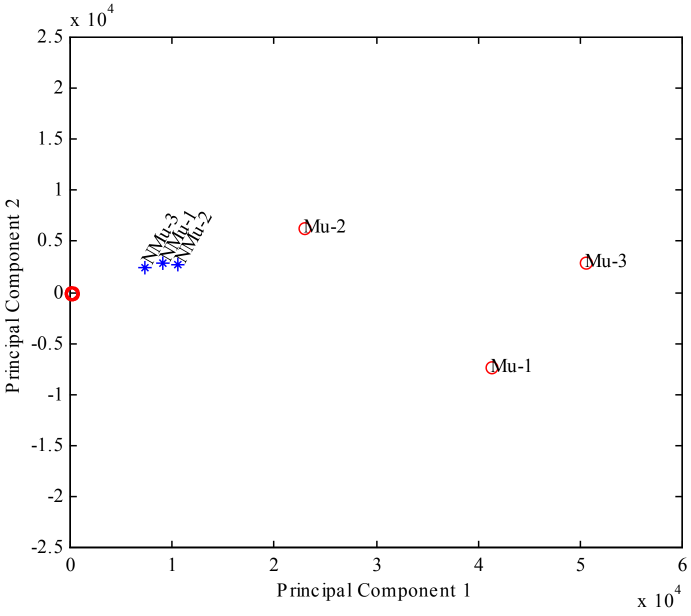
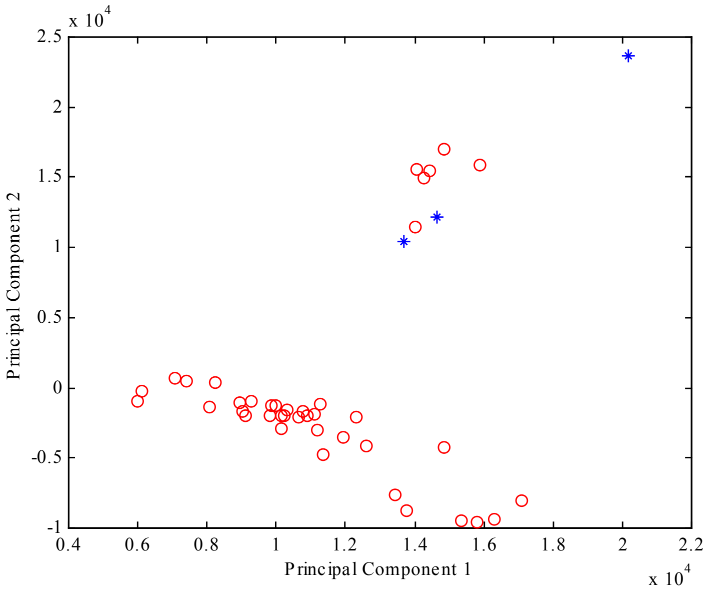


Figure 2: Projection of the samples using the genes in the specific structures observed in the Loading plot. (a) The projection using genes in structure A separated the skeletal muscle tissue. The genes in this structure contained several troponins and skeletal muscle related genes. (b)The projection using genes in structure B separated the liver and lung. These genes contained albumins and apolipoproteins, among others (c) The genes in structure C separated the brain samples from the remaining tissues. The structure contained several brain specific genes and ribosome related genes.


Histogram of the angles of the genes in the loading plot
Structure A
Structure B
Structure C
Number of Genes
Angle
0
5
10
15
20
25
30
0
0.5
1
1.5


x 10^4
Principal Component 2
Principal Component 1
x 10^4
Li-2
Li-1
NLi-1
NLi-2
NLi-3
1
0.5
0
-0.5
-1
0
0.5
1
1.5
2
2.5
3


x 10^4
Principal Component 2
Principal Component 1
x 10^4
NMu-3
NMu-1
NMu-2
Mu-2
Mu-3
Mu-1
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
0
1
2
3
4
5
6


x 10^4
Principal Component 2
Principal Component 1
x 10^4
2.5
2
1.5
1
0.5
0
-0.5
-1
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2