Cancer Genetics Branch

AFTER THE SEQUENCE: WHOLE GENOME APPROACHES TO BIOLOGICAL QUESTIONS

GENE EXPRESSION

GENE VARIATION

GENE FUNCTION

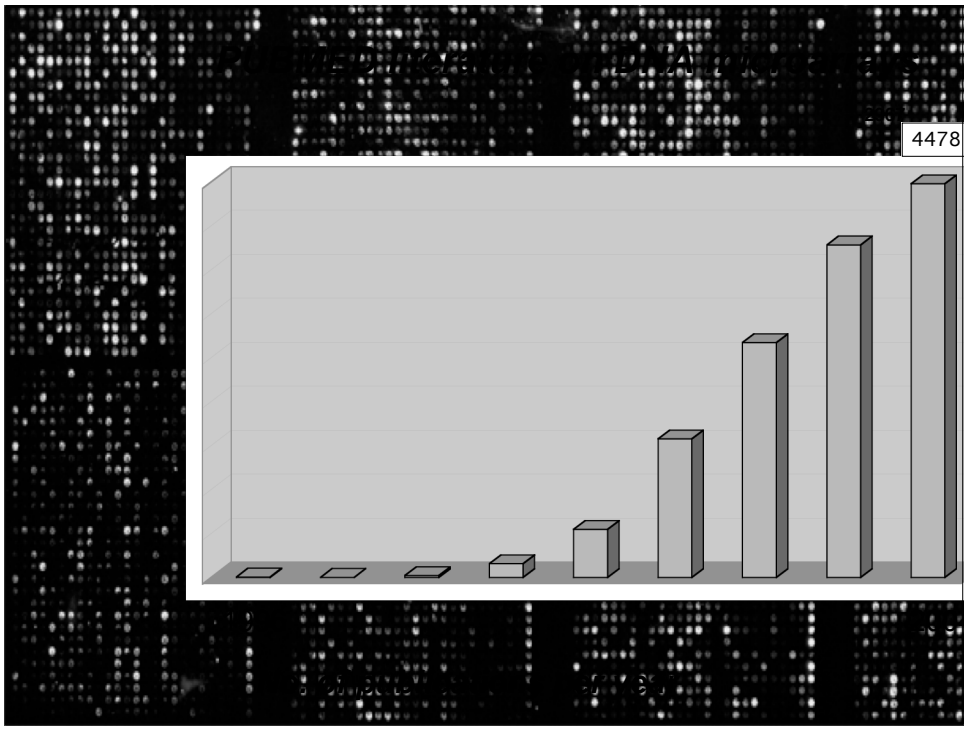
National Human Genome Research Institute

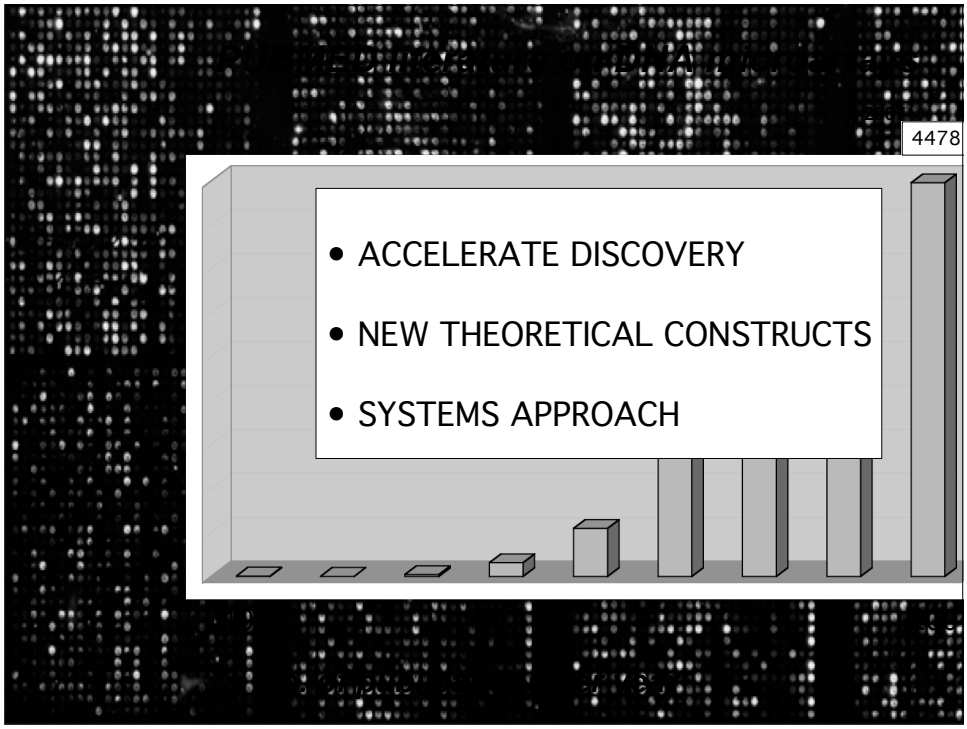
Cancer Genetics Branch

MICROARRAYS PROVIDE A TOOL FOR WHOLE GENOME ANALYSIS

PRIMARY IMPACT: ACCELERATED DISCOVERY AND HYPOTHESIS GENERATION

National Human Genome Research Institute





Cancer Genetics Branch

MICROARRAY TERMINOLOGY

- **Feature--an array element**
- **Probe--a feature corresponding to a defined sequence**
- **Target--a pool of nucleic acids of unknown sequence**

National Human Genome Research Institute

Cancer Genetics Branch

POSSIBLE ARRAY FEATURES

- **Synthetic Oligonucleotides**
- **PCR products from**
  - **Cloned DNAs**
  - **Genomic DNA**
- **Cloned DNA**

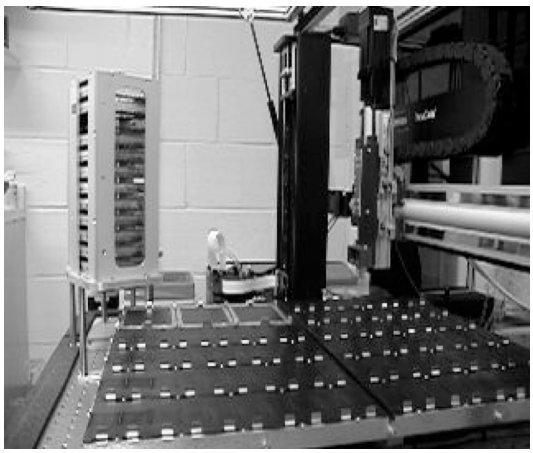
Cancer Genetics Branch

Microarray Manufacture

- **Printing**

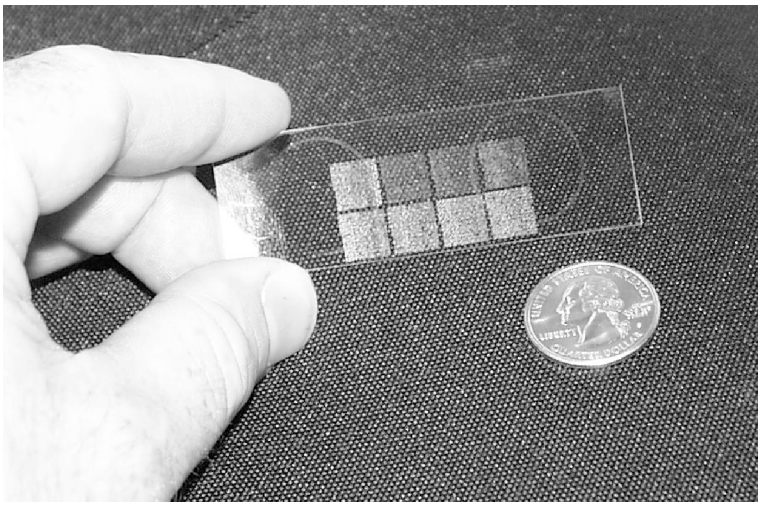
National Human Genome Research Institute

Cancer Genetics Branch



National Human Genome Research Institute

Cancer Genetics Branch



National Human Genome Research Institute

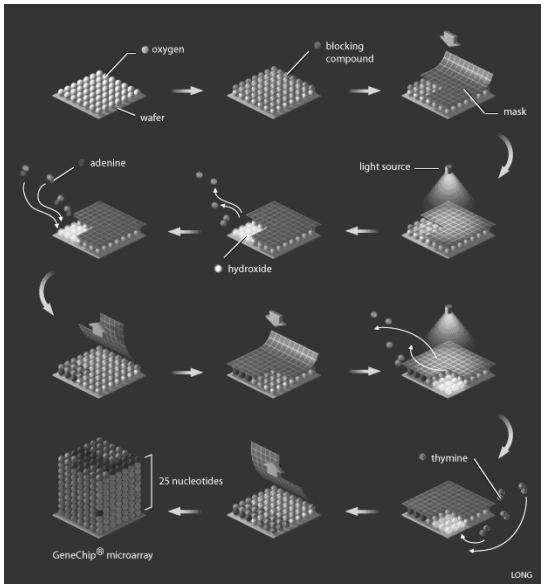
Cancer Genetics Branch

Microarray Manufacture

- **Printing**
- **Synthesis *in situ***

Cancer Genetics Branch

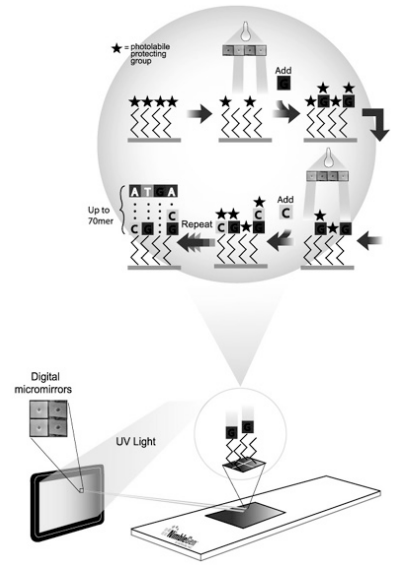
LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS



National Human Genome Research Institute

Cancer Genetics Branch

LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS



National Human Genome Research Institute

Cancer Genetics Branch

MICROARRAY READOUT

- Determine quantity of target bound to each probe in a complex hybridization
- Must have high sensitivity, low background
- High spatial resolution essential
- Dual channel capability
- Fluorescent tags meet these demands

National Human Genome Research Institute

Cancer Genetics Branch

Building Microarrays

- Methods are applicable to any organism
- Sequenced organisms: oligonucleotides
- Unsequenced organisms: cloned DNAs

Cancer Genetics Branch

Building Microarrays

- Density depends on specific technology
- Printing based methods limited to 40-50K
  - In situ synthesis: 100K and up
- Array design is linked to purpose.

National Human Genome Research Institute

Cancer Genetics Branch

Laboratory Essentials

- Arrays
- Scanner
- Software for processing array image
- Software for data analysis and display

Cancer Genetics Branch

DNA Microarray Applications

- Resequencing
- Comparative Genomic Hybridization
- Gene Expression
- Transcription factor localization
- Chromatin/DNA modification

National Human Genome Research Institute

Cancer Genetics Branch

DNA Microarray Applications

- Resequencing
- Comparative Genomic Hybridization
- Gene Expression
- Transcription factor localization
- Chromatin/DNA modification

Cancer Genetics Branch

DNA Microarray Applications

- Resequencing

Mutations

Polymorphisms

National Human Genome Research Institute

Cancer Genetics Branch

SINGLE NUCLEOTIDE POLYMORPHISM

AGGTTACCAGTA

AGGTTGCCAGTA

OCCUR ABOUT 1: 1250 BASES

•Dense SNP maps provide a basis to design microarrays for genome scanning

National Human Genome Research Institute

Cancer Genetics Branch

LABELLING SNPs

Genomic DNA

↓

Reduced complexity PCR product

↓

Label

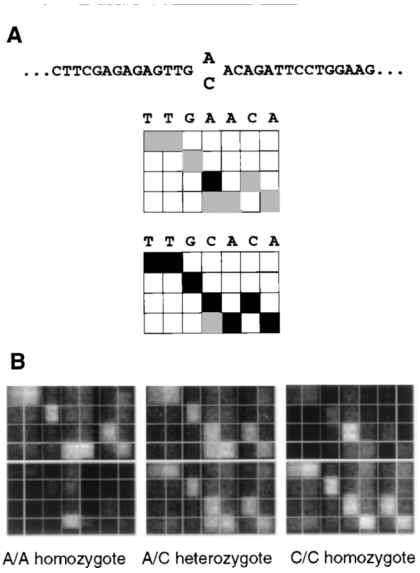
↓ **pool, denature, dilute into buffer**

Hybridize to microarray

National Human Genome Research Institute

Cancer Genetics Branch

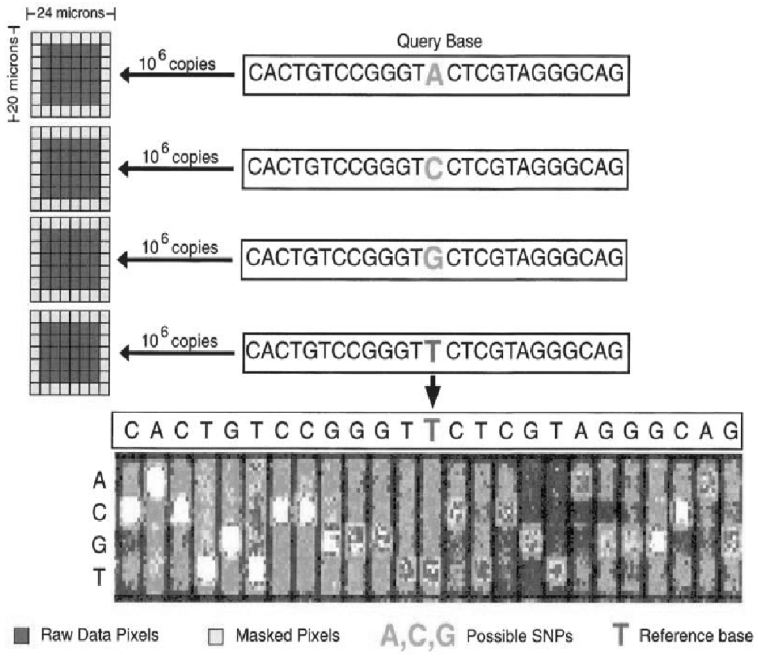
SNP CHIP*



*Wang et al.
Science 280:1077
1998

National Human Genome Research Institute

Cancer Genetics Branch



Cutler DJ et al. Genome Res. 2001 11:1913-25.

National Human Genome Research Institute

Cancer Genetics Branch

ACCURACY OF SNP CHIP

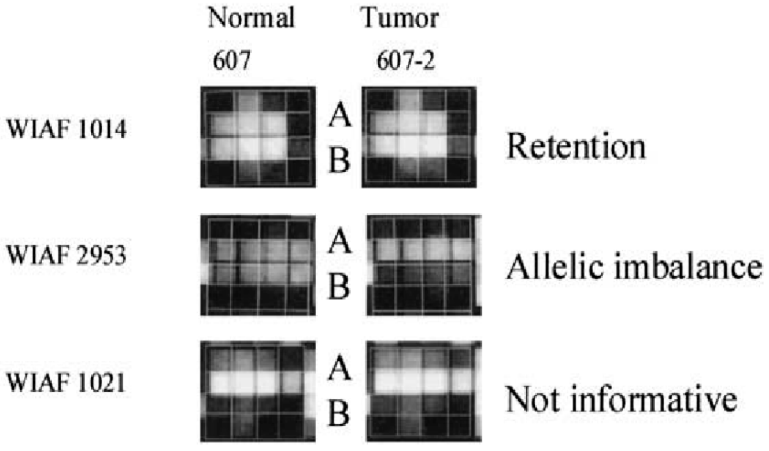
Table 3. ABACUS SNP Detection and Genotyping Accuracy

	Verified	Total Possible
A. Accuracy of autosomal SNPs detection		
Singleton SNPs	17	17
Non-singleton SNPs	91	91
Total SNPs	108	108
B. Number of autosomal SNPs electronically verified		
Number of SNPs electronically verified	371	
C. Accuracy of autosomal genotype calls		
Number of verified homozygous genotype calls	1515	
Number of incorrect homozygous genotype calls	0	
Percent correct homozygote calls	100.00%	
Number of verified heterozygous genotype calls	423	
Number of incorrect heterozygous genotype calls	3	
Percent correct heterozygote calls	99.30%	
D. Accuracy of haploid genotype calls		
Number of bases sequenced (6X coverage)	17,423	
Number of bases different from microarray chip calls	0	
Percent of bases identical	100.00%	

Cutler DJ et al. Genome Res. 2001 11:1913-25.

Cancer Genetics Branch

SNP CHIP FOR ALLELIC IMBALANCE



Primdahl H et al. J Natl Cancer Inst. 2002, 94:216-223

Cancer Genetics Branch

SNP CHIPS

HAVE ACHIEVED HIGH DENSITY

1,586,383 SNPS

HINDS ET AL. SCIENCE 307:1072 (2005)

COMMERCIAL CHIPS AVAILABLE: 100,000 SNPS

SOON TO INCREASE

VIABLE OPTION FOR:
GENOTYPING.
CANCER ALLELIC IMBALANCE.

Cancer Genetics Branch

ROLE OF SNP CHIPS IN RESEQUENCING CODING AND FUNCTIONAL SNPS

TECHNICAL CHALLENGE FOR LARGE SCALE ANALYSIS

AMPLICHIP CYP450 NOW FDA APPROVED

(31 POLYMORPHISMS IN
2D6 AND 2C19 P450 GENES)

LIKELY TO BE OF GROWING CLINICAL AND RESEARCH SIGNIFICANCE

National Human Genome Research Institute

Cancer Genetics Branch

DNA Microarray Applications

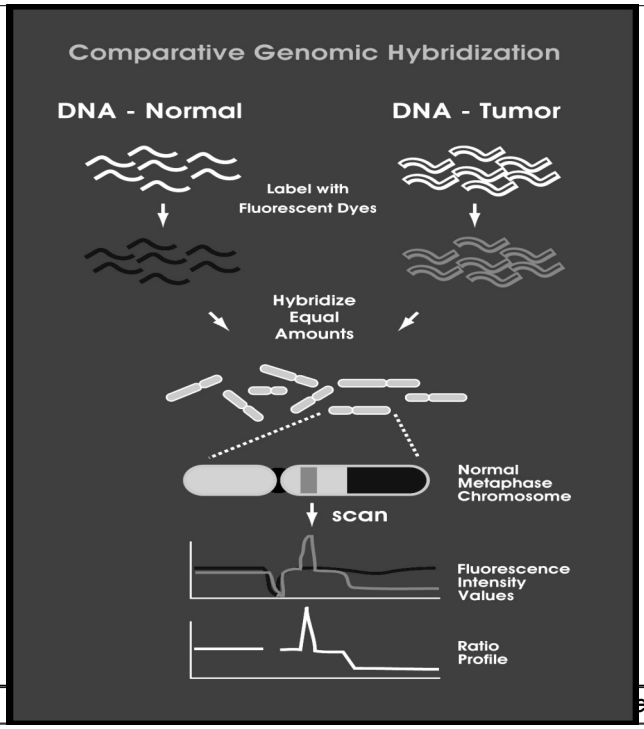
- **Resequencing**
- **Comparative Genomic Hybridization**
- **Gene Expression**
- **Transcription factor localization**
- **Chromatin/DNA modification**

Cancer Genetics Branch

COMPARATIVE GENOMIC HYBRIDIZATION

- Method for gene copy number determination.
- Useful in cancer research to localize regions containing candidate oncogenes (gains) and tumor suppressor genes (losses).
- Useful in hereditary disease research to localize regions containing constitutional gains or losses of chromosome segments.

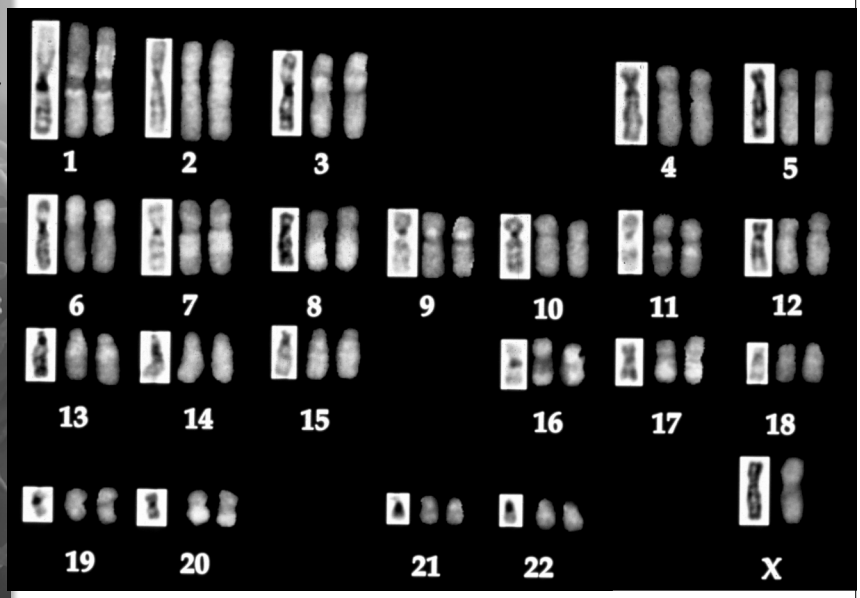
Cancer Genetics Branch



earch Institute

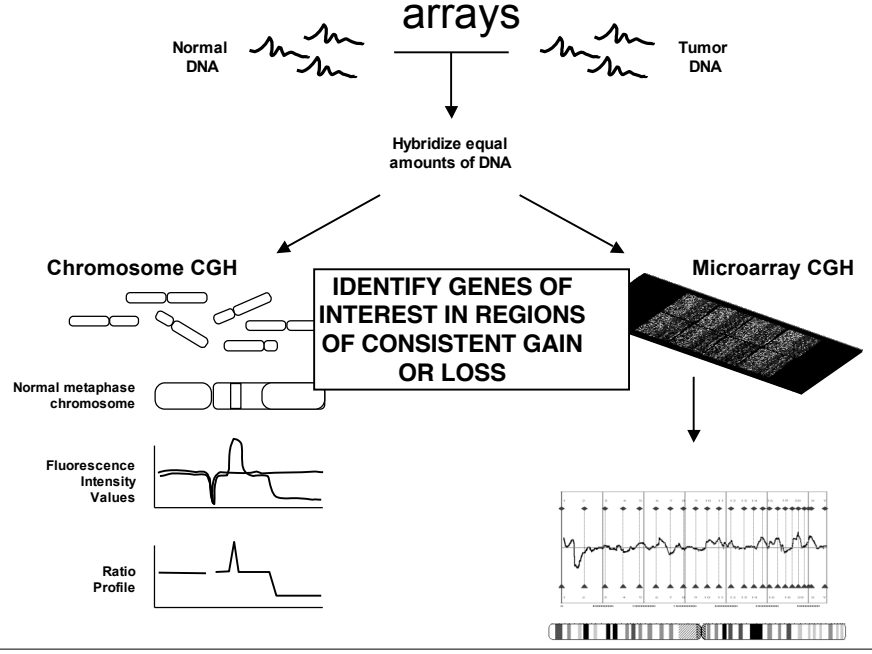
Cancer Genetics Branch

COMPARATIVE GENOMIC HYBRIDIZATION



Cancer Genetics Branch *National Human Genome Research Institute*

Comparative Genomic Hybridization (CGH) arrays



Cancer Genetics Branch

PLATFORMS FOR ARRAY BASED COMPARATIVE GENOMIC HYBRIDIZATION (CGH)

- BACs
- cDNAs
- Oligonucleotides

National Human Genome Research Institute

Cancer Genetics Branch

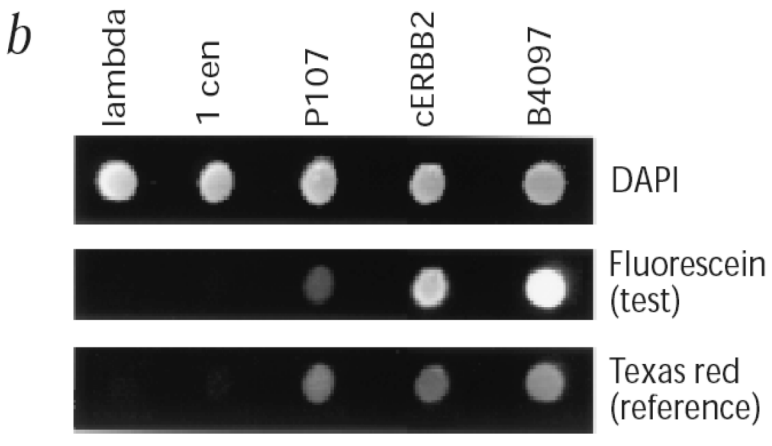
ARRAY CGH

- HIGH RESOLUTION.
- SIMPLIFIED IMAGE ANALYSIS.
- HIGH THROUGHPUT.
- OLIGO STRATEGY ALLOWS GENOME BASED DESIGN.

National Human Genome Research Institute

Cancer Genetics Branch

CGH BAC ARRAYS

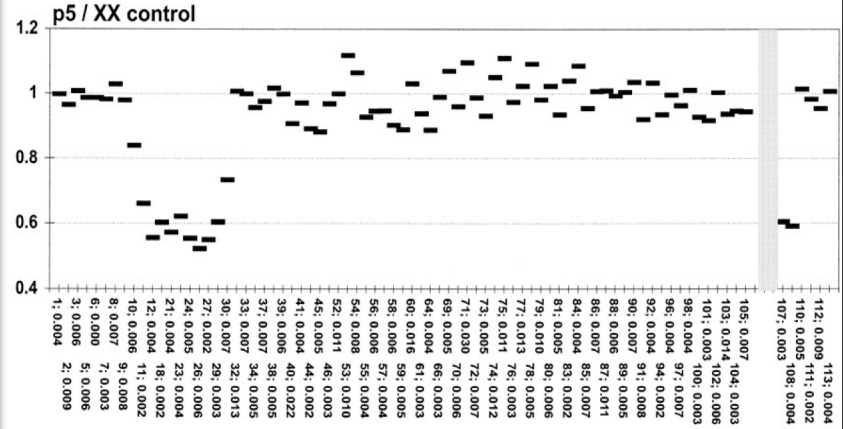


Pinkel D et al., Nature Genetics 20, 207 - 211 ,1998.

National Human Genome Research Institute

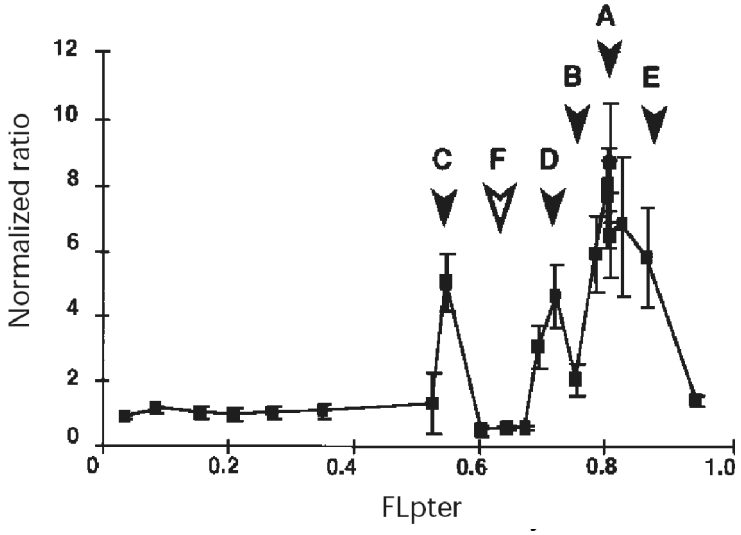
Cancer Genetics Branch

CGH BAC ARRAYS



Bruder CE et al., Hum Mol Genet. 2001;10:271-82.

Cancer Genetics Branch

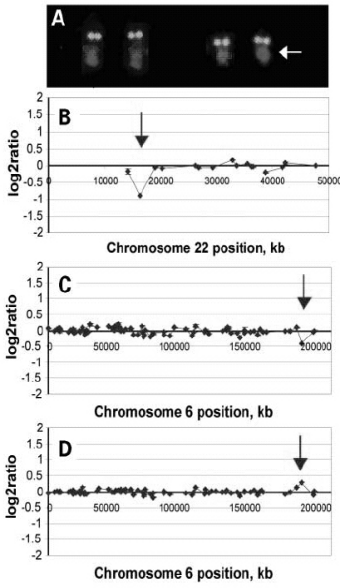


Pinkel et al. Nat Gen 20:207

National Human Genome Research Institute

Cancer Genetics Branch

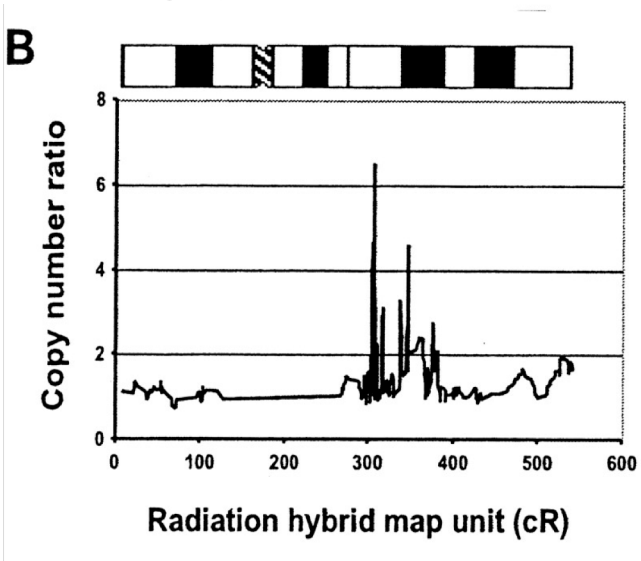
CGH BAC ARRAYS



Albertson and Pinkel Hum Mol Genet 2003 12:145

Cancer Genetics Branch

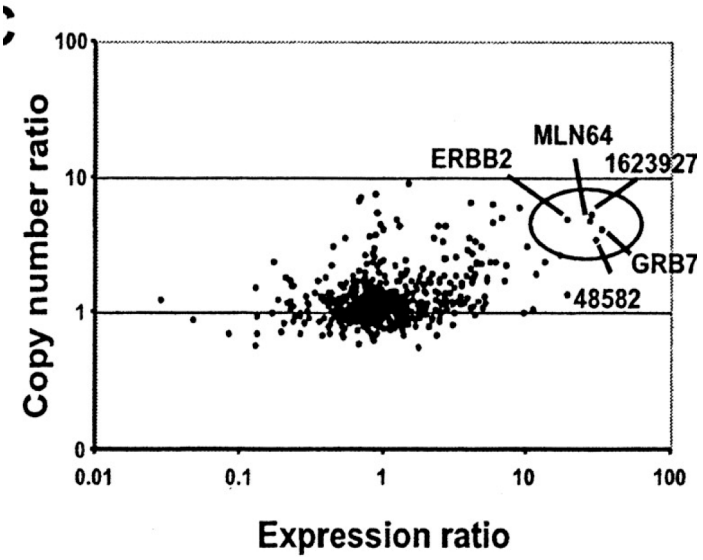
CGH cDNA



Kauraniemi P et al., Cancer Res. 2001 ;61:8235-40.

Cancer Genetics Branch

CGH cDNA

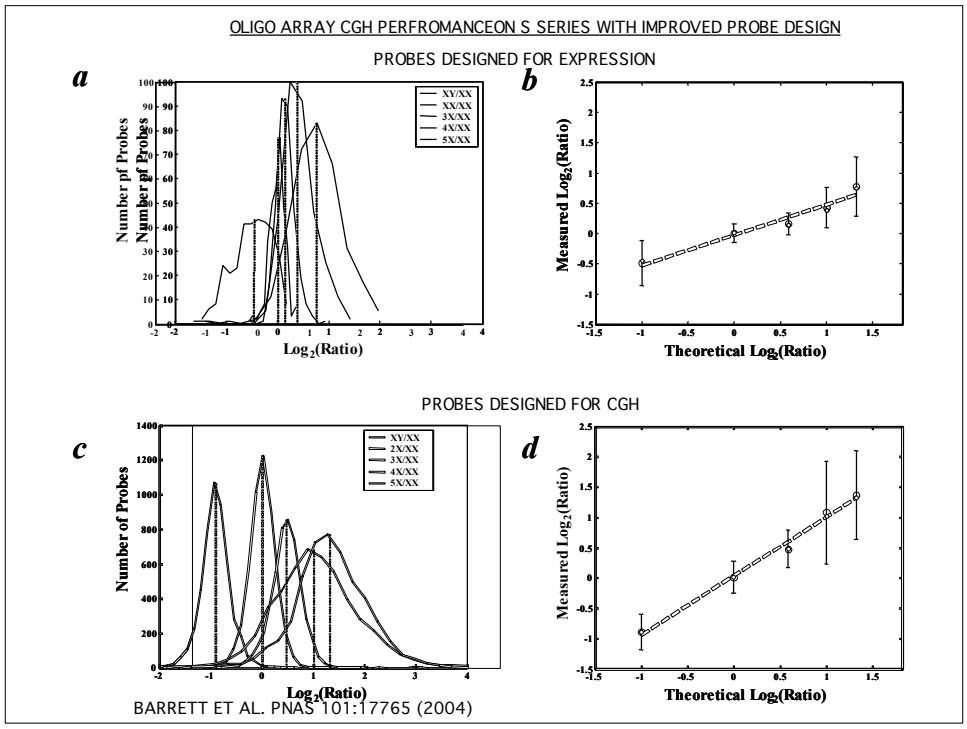


Kauraniemi P et al., Cancer Res. 2001; 61:8235-40

Cancer Genetics Branch

OLIGONUCLEOTIDE BASED CGH

- No bacterial cultures.
- Flexible in silico design.
- Resolution limited only by feature density
- Challenge: complex hybridization



BARRETT ET AL. PNAS 101:17765 (2004)

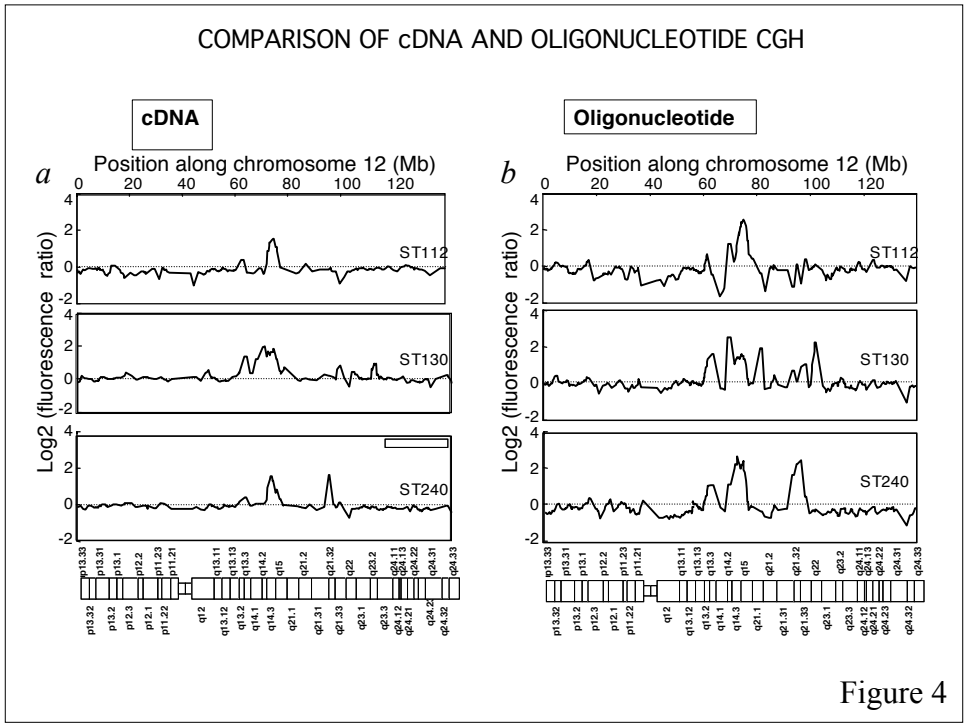
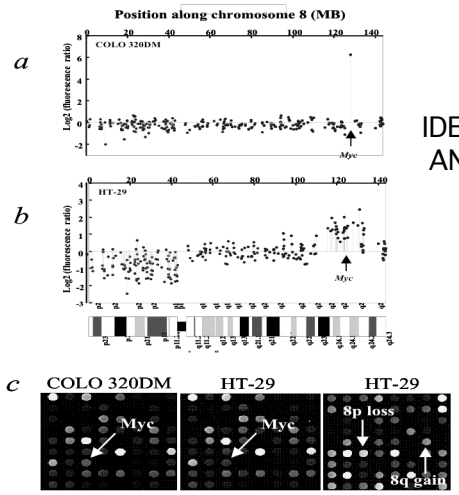


Figure 4

Cancer Genetics Branch

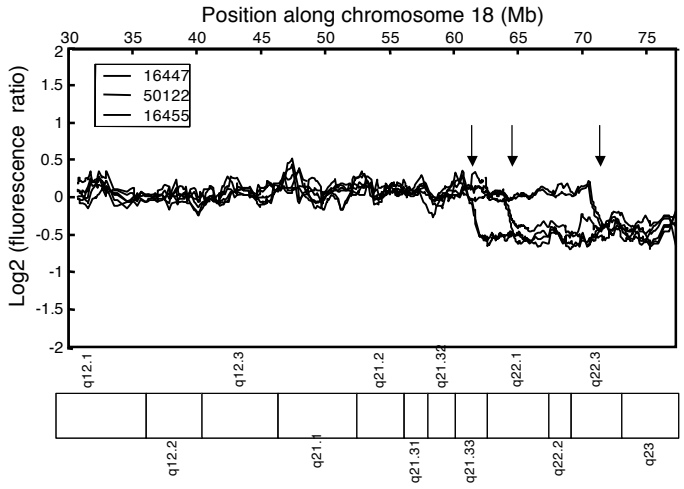


IDENTIFICATION OF GAINS AND LOSSES IN CANCER CELLS

Figure 1. Barrett et al.

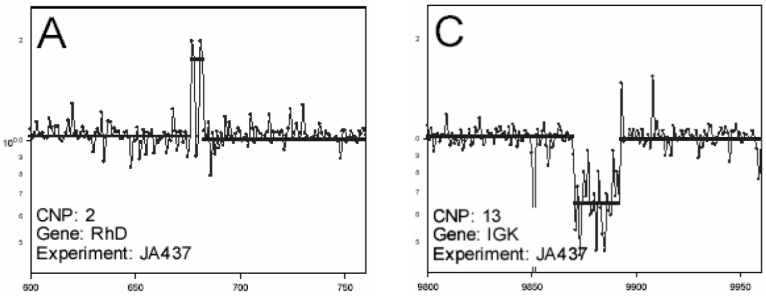
National Human Genome Research Institute

LOCATING CONSTITUTIONAL DELETIONS



Cancer Genetics Branch

HIGH DENSITY OLIGO ARRAYS FOR DETECTING COPY NUMBER POLYMORPHISM



Sebat et al., Science 2004;305:525.

Cancer Genetics Branch

DNA Microarray Applications

- Resequencing
- Comparative Genomic Hybridization
- Gene Expression
- Transcription factor localization
- Chromatin/DNA modification

Cancer Genetics Branch

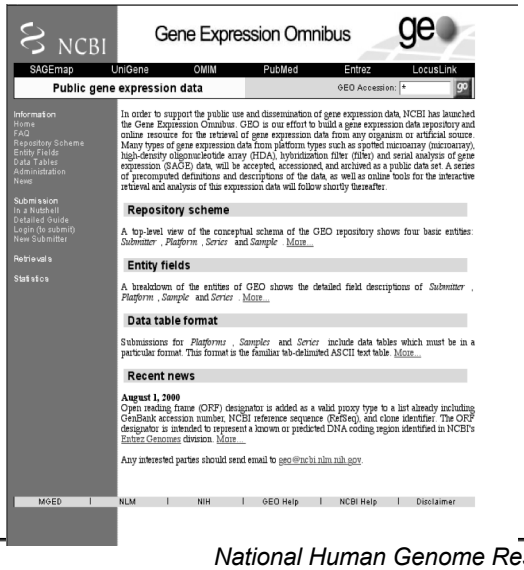
Gene Expression ProfilingTechnologies

- **cDNA library sequencing**
- **Serial analysis of gene expression (SAGE)**
- **MPSS (massively parallel signature sequencing)**
- **Microarray hybridization**

Cancer Genetics Branch

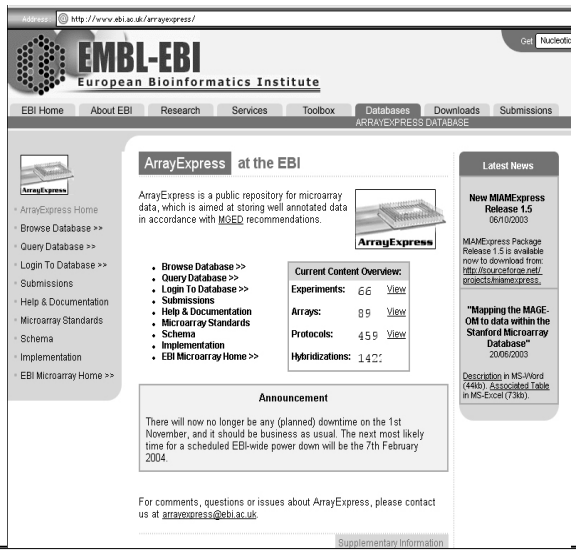
Accessing Expression Data

•Individual Lab and Journal Sites



Cancer Genetics Branch

Accessing Expression Data



Cancer Genetics Branch

Publishing Expression Data

•MIAME standard

Minimum Information about a Microarray Experiment

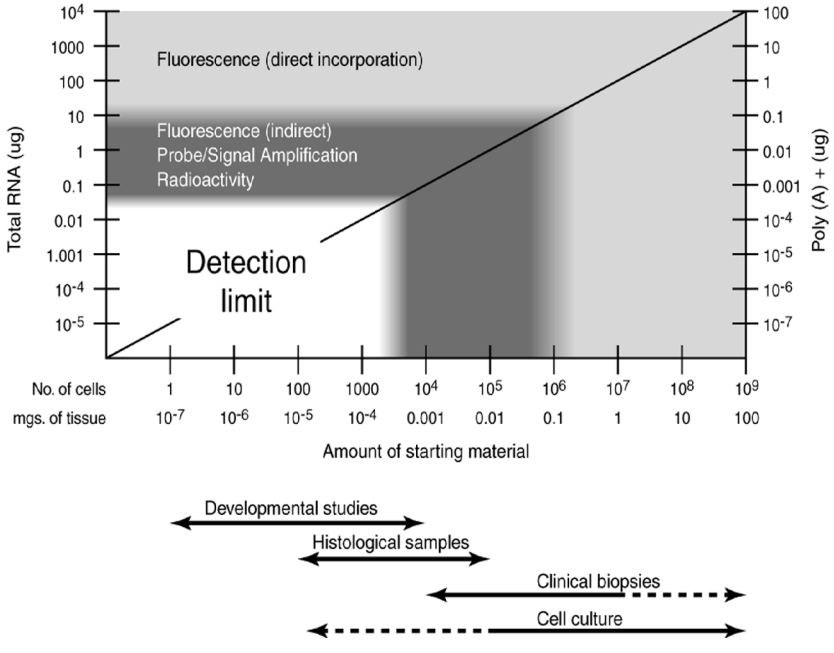
Format required by many journals

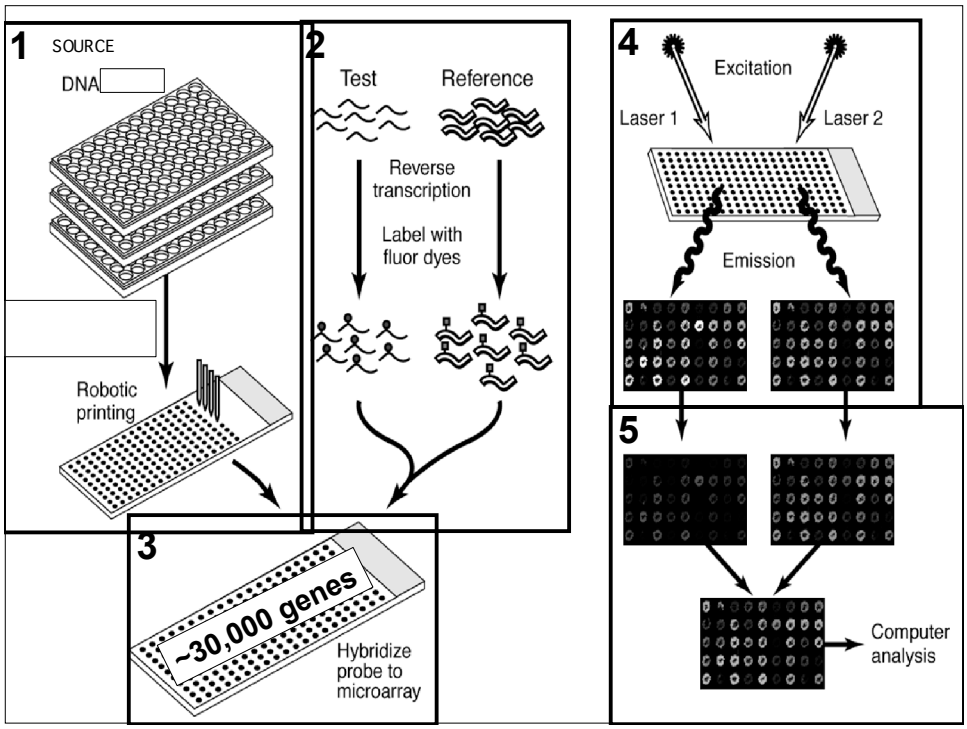
Cancer Genetics Branch

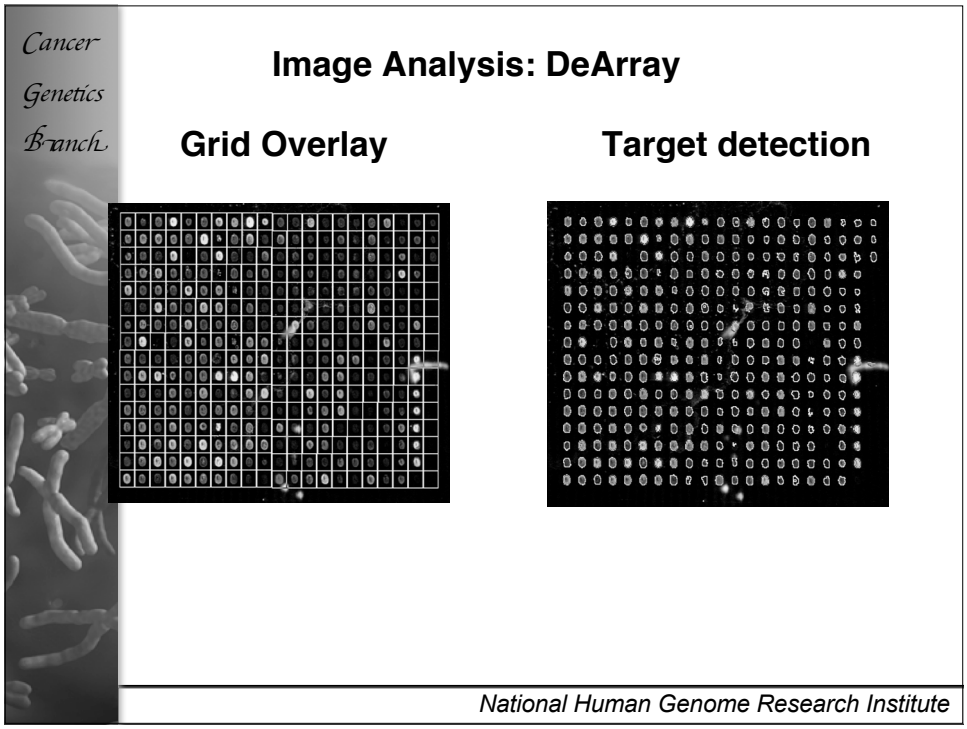
STRATEGIES FOR SIGNAL GENERATION FROM mRNA

- **Fluorochrome conjugated cDNA**
- **Ligand substituted nucleotides with secondary detection (e.g. biotin-streptavidin)**
- **Radioactivity**
- **RNA amplification**

National Human Genome Research Institute

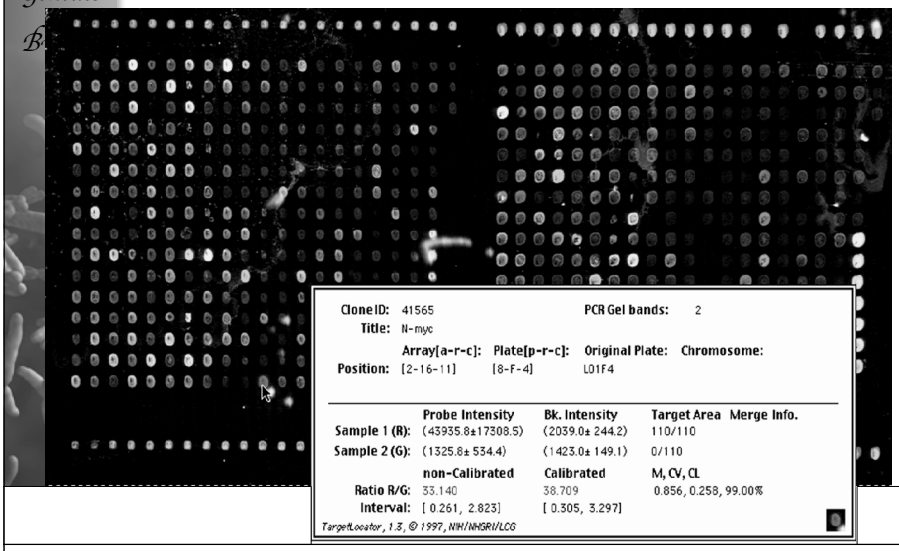



1
SOURCE
DNA
Robotic printing
2
Test
Reference
Reverse transcription
Label with fluor dyes
3
~30,000 genes
Hybridize probe to microarray
4
Excitation
Laser 1
Laser 2
Emission
5
Computer analysis


Cancer Genetics Branch
Image Analysis: DeArray
Grid Overlay
Target detection
National Human Genome Research Institute

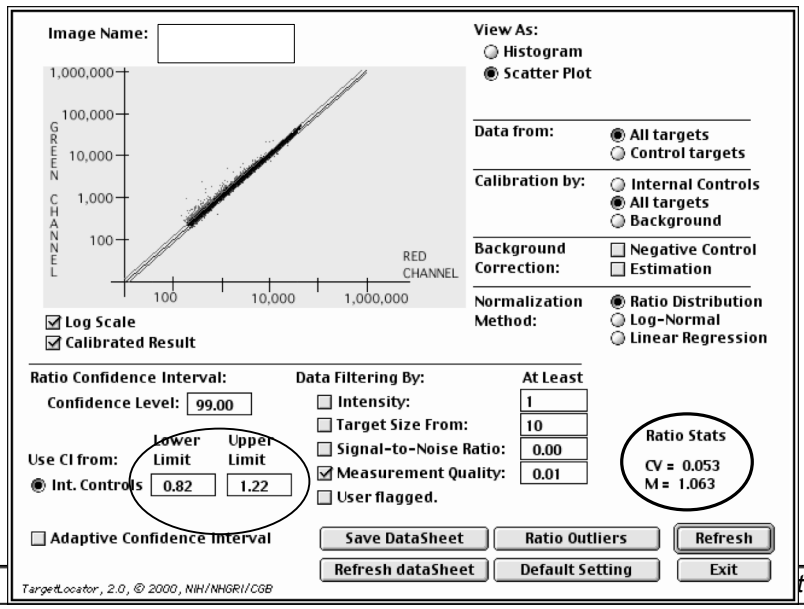
Cancer Genetics Branch

Image Analysis: DeArray

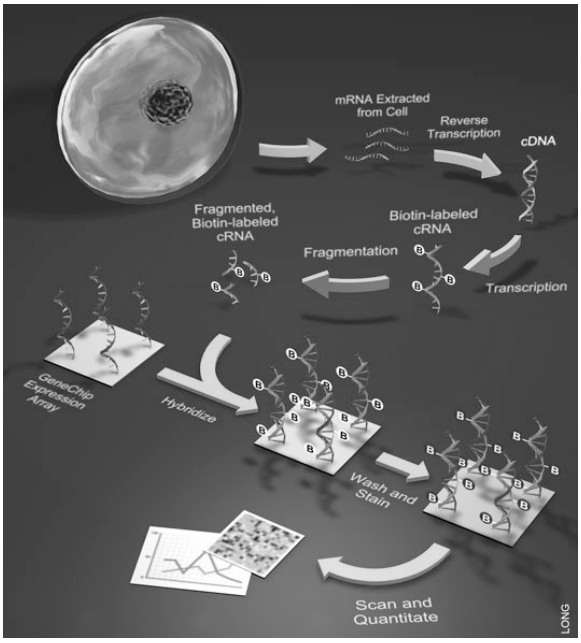


Cancer Genetics Branch

DATA QUALITY IS CRITICAL



Cancer Genetics Branch



ONE COLOR HYBRIDIZATION ON AN OLIGO ARRAY

National Human Genome Research Institute

Cancer Genetics Branch

Output of Microarray Analysis:

expression ratio
(2 color hybridization)

or

relative expression level
(1 color hybridization)

Both types of data can be analyzed with essentially the same tools.

National Human Genome Research Institute

Cancer Genetics Branch

APPLICATIONS OF EXPRESSION ARRAYS

•Expression profiling

Power arises from increasing sample number

•Direct comparisons (Induction)

Biological system critical

•Genome Annotation

Cancer Genetics Branch

A RECURRING PROBLEM

Disease Genes

Transcription factors

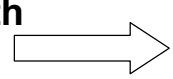
Hormones/growth factors

Drugs

Toxins

Infectious agents

Physical agents



?????

<u>Downstream Genes</u>

•Direct targets

•Indirect targets

Cancer Genetics Branch

EXPRESSION DATA ANALYSIS

- **Large amount of data**
- **Requires visualization and analysis tools**

Cancer Genetics Branch

EXPRESSION DATA ANALYSIS

- **Check quality of individual experiments**
  - **Preprocessing**
    - **Normalization**
    - **Remove genes which are not accurately measured**
    - **Remove genes which are similarly expressed in all samples**
  - **Unsupervised Clustering**
  - **Supervised Clustering**

National Human Genome Research Institute

Cancer Genetics Branch

Unsupervised Clustering

How do genes and samples organize into groups?

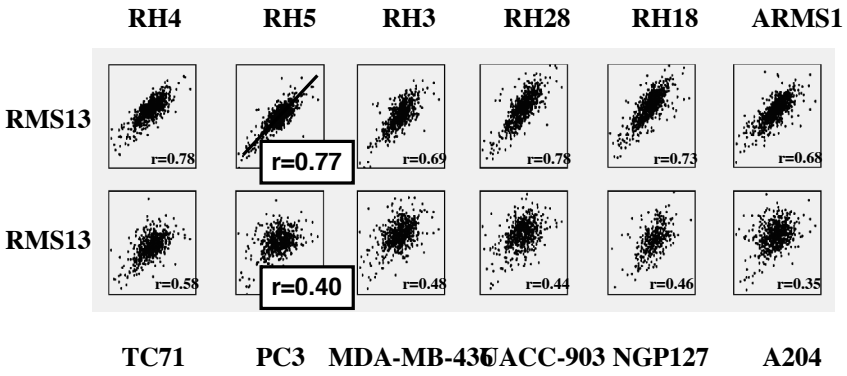
Powerful method of data display.

Does <u>not</u> prove the validity of groups.

- **Clustered Samples Are Biologically Similar**
- **Clusters of Co-expressed genes**
  - **May be functionally related**
  - **May be enriched for pathways**

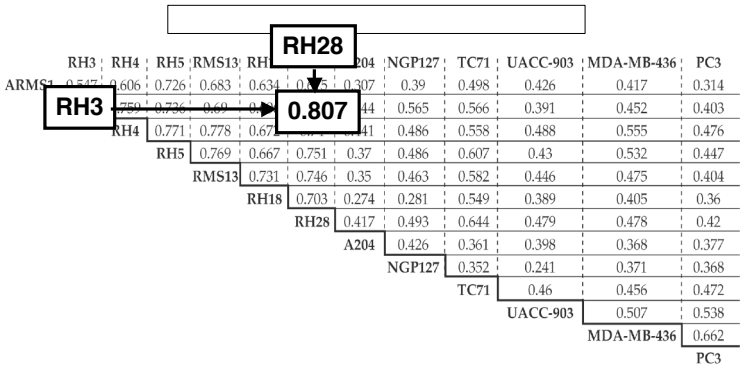
National Human Genome Research Institute

UNSUPERVISED CLUSTERING IS BASED ON A GLOBAL SIMILARITY METRIC



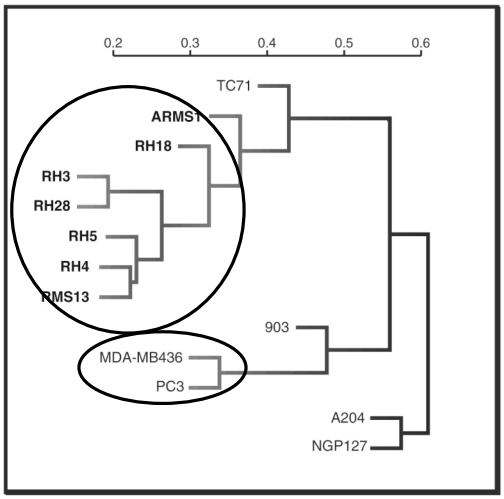
Matrix of Pearson Correlation Coefficients Distance Map

78 pair-wise comparisons

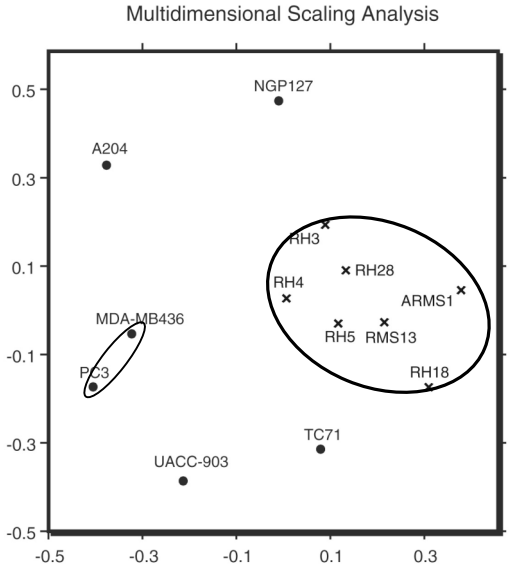


	RH3	RH4	RH5	RMS13	RH18	RH28	A204	NGP127	TC71	UACC-903	MDA-MB-436	PC3
ARMS1	0.547	0.606	0.726	0.683	0.634	0.[illegible]	0.307	0.39	0.498	0.426	0.417	0.314
RH3		0.759	0.736	0.69	0.[illegible]	0.807	0.44	0.565	0.566	0.391	0.452	0.403
RH4			0.771	0.778	0.67[illegible]		0.41	0.486	0.558	0.488	0.555	0.476
RH5				0.769	0.667	0.751	0.37	0.486	0.607	0.43	0.532	0.447
RMS13					0.731	0.746	0.35	0.463	0.582	0.446	0.475	0.404
RH18						0.703	0.274	0.281	0.549	0.389	0.405	0.36
RH28							0.417	0.493	0.644	0.479	0.478	0.42
A204								0.426	0.361	0.398	0.368	0.377
NGP127									0.352	0.241	0.371	0.368
TC71										0.46	0.456	0.472
UACC-903											0.507	0.538
MDA-MB-436												0.662
PC3												

Hierarchical Clustering Dendrogram



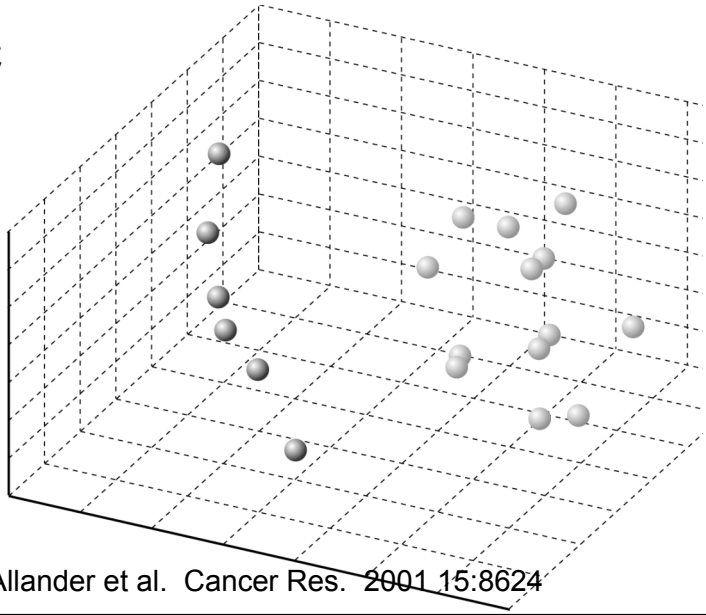
DATA DISPLAY BY MULTIDIMENSIONAL SCALING



Cancer Genetics Branch

C

MDS PLOT

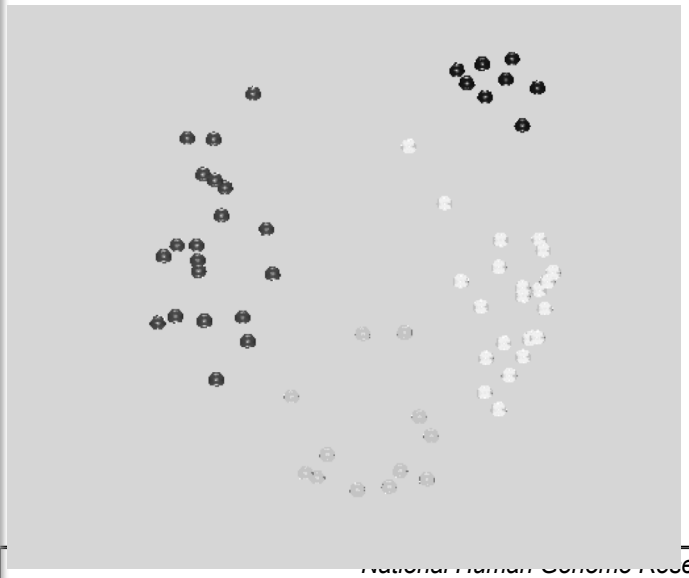


Allander et al. Cancer Res. 2001 15:8624

National Human Genome Research Institute

Cancer Genetics Branch

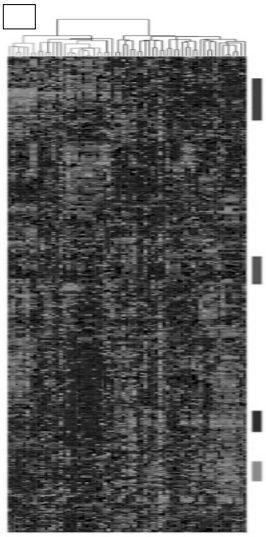
MULTIDIMENSIONAL SCALING



-Lymphoma
RMS
NBL
-EWS

Cancer Genetics Branch

CLUSTERING GENES AND SAMPLES



Perou et al. Nature 2000 406:747

Cancer Genetics Branch

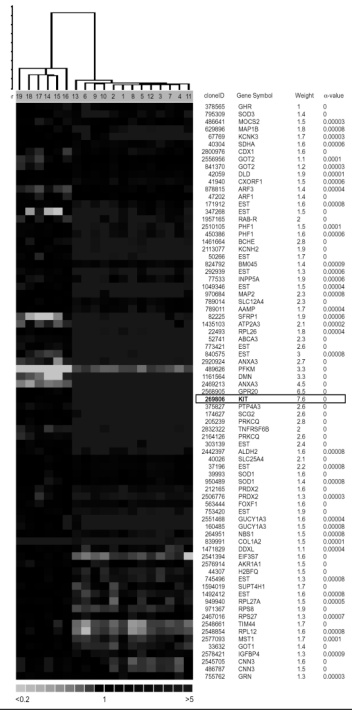
Supervised Clustering

What genes distinguish samples in selected groups from each other?

- **Choice of groups can be based on any known property of the samples.**
- **Many possible underlying methods: t-test or F-statistic frequently used.**
- **Output includes ranked gene list.**
- **Leads to the development of classifiers which can be applied to unknown samples.**
- **Must address the problem of false discovery due to multiple comparisons and discrepancy between sample/gene numbers.**

National Human Genome Research Institute

Cancer Genetics Branch



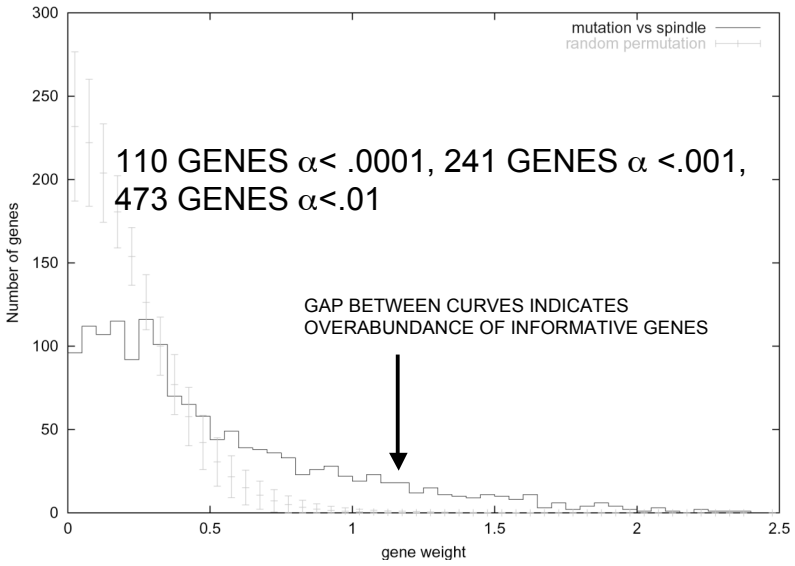
HIERARCHICAL CLUSTERING OF SAMPLES/GENES USING THE GENES SELECTED BY SUPERVISED ANALYSIS

Allander et al. Cancer Res. 2001 15:8624

al Human Genome Research Institute

Cancer Genetics Branch

OVERABUNDANCE OF INFORMATIVE GENES DEMONSTRATED BY RANDOM PERMUTATION TEST



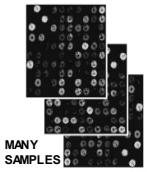
Allander et al. Cancer Res. 2001 15:8624

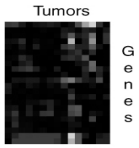
National Human Genome Research Institute

Cancer Genetics Branch

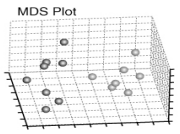
GENOMICS FROM BENCH TO BEDSIDE

WHOLE GENOME

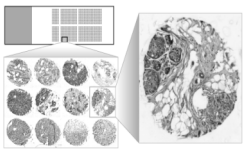


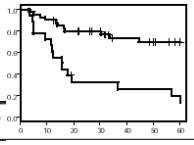


GENE SELECTION



GENE VALIDATION





ASSAY DEVELOPMENT

National Human Genome Research Institute

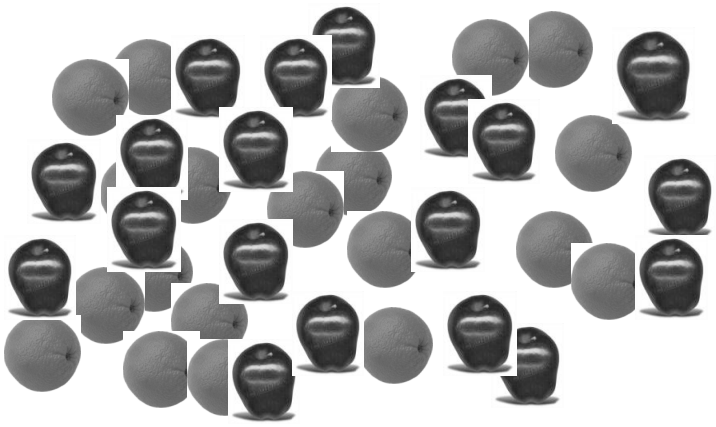
Cancer Genetics Branch

SIGNAL STRENGTH VARIES IN TISSUE PROFILING EXPERIMENTS

THE MOST INTERESTING QUESTIONS TEND TO BE ASSOCIATED WITH WEAKER SIGNAL.

Cancer Genetics Branch

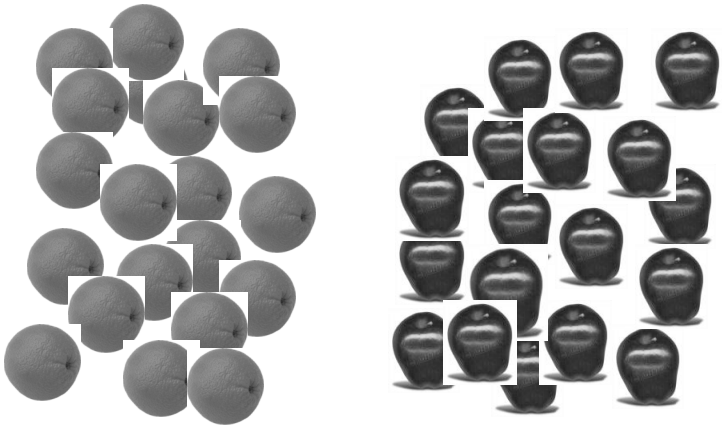
CONSIDER A SAMPLE SET



National Human Genome Research Institute

Cancer Genetics Branch

CONSIDER A SAMPLE SET



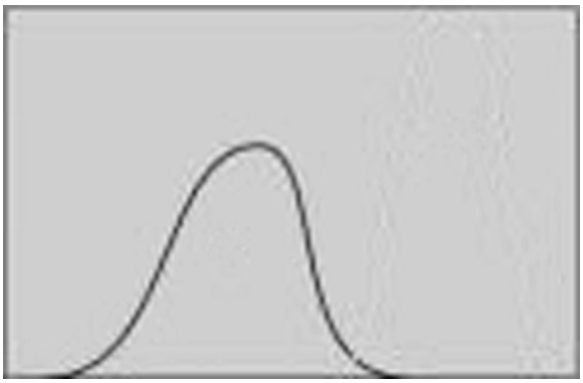
THESE ARE EASY TO DISTINGUISH BY ONE MEASUREMENT PER INDIVIDUAL.

National Human Genome Research Institute

Cancer Genetics Branch

CONSIDER A SAMPLE SET

TUMORS



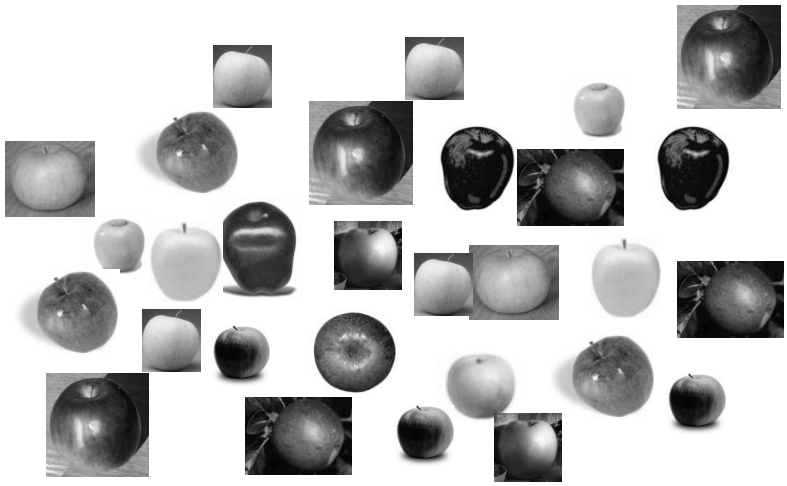
EXPRESSION LEVEL
(HIGHLY INFORMATIVE GENE)

THESE ARE EASY TO DISTINGUISH BY ONE MEASUREMENT PER INDIVIDUAL.

National Human Genome Research Institute

Cancer Genetics Branch

CONSIDER A SAMPLE SET

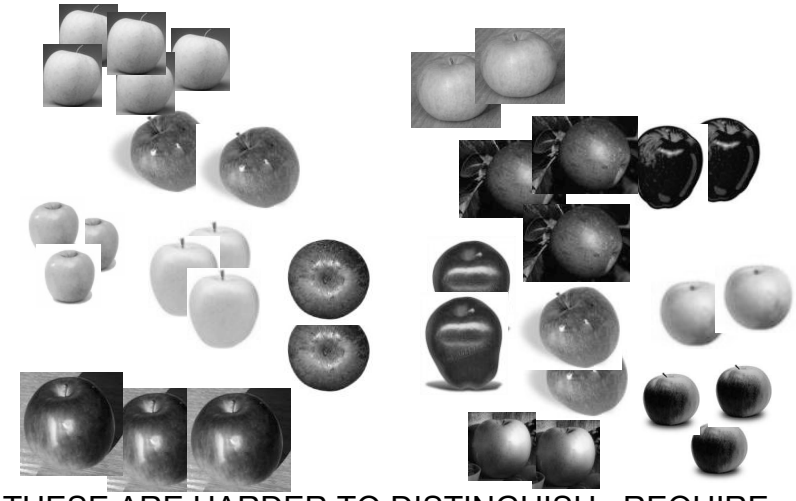


THESE ARE HARDER TO DISTINGUISH. REQUIRE MORE THAN ONE MEASUREMENT PER INDIVIDUAL.

National Human Genome Research Institute

Cancer Genetics Branch

CONSIDER A SAMPLE SET



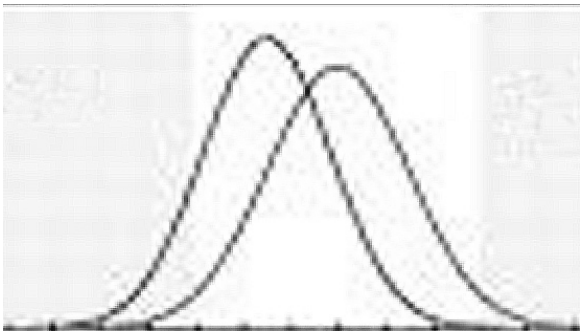
THESE ARE HARDER TO DISTINGUISH. REQUIRE MORE THAN ONE MEASUREMENT PER INDIVIDUAL.

National Human Genome Research Institute

Cancer Genetics Branch

CONSIDER A SAMPLE SET

TUMORS



EXPRESSION LEVEL
(POORLY INFORMATIVE GENE)

THESE ARE HARDER TO DISTINGUISH. REQUIRE MORE THAN ONE MEASUREMENT PER INDIVIDUAL.

Cancer Genetics Branch

WE CAN TELL APPLES FROM ORANGES.

CAN WE DISTINGUISH DIFFERENT KINDS OF APPLES?

National Human Genome Research Institute

Cancer Genetics Branch

A CONTINUUM OF POSSIBLE OUTCOMES FROM MICROARRAY RESEARCH

• SOME FEATURES WILL SEPARATE TUMORS EASILY INTO CLASSES, AND MIGHT BE REDUCED TO SINGLE GENE TESTS, IMPLEMENTED IN A CONVENTIONAL FASHION.

• OTHERS WILL BE MORE DIFFICULT, AND REQUIRE MULTIPLE GENE MEASUREMENTS.

• MANY CLINICALLY RELEVANT FEATURES APPEAR TO FALL WITHIN THIS DIFFICULT GROUP.

National Human Genome Research Institute

Cancer Genetics Branch

A CONTINUUM OF POSSIBLE OUTCOMES FROM MICROARRAY RESEARCH

• SOME GENES WILL SHOW DIFFERENCES BETWEEN GROUPS OF SAMPLES BY CHANCE ALONE.

• THERE MAY BE NO ONE GENE WHICH SEPARATES GROUPS RELIABLY.

• FIND THE MOST INFORMATIVE GENES AND USE THEM IN COMBINATION .

National Human Genome Research Institute

Cancer Genetics Branch

RISK OF OVERFITTING IN CLINICAL STUDIES WITH SMALL SAMPLE SETS

NEED INDEPENDENT VALIDATION SETS.

Cancer Genetics Branch

MICROARRAY STUDIES GENERATE ORGANIZED LIST OF GENES

- Often cryptic and hard to interpret.
- Hypothesis generating, but this is often rather subjective.
- Seldom provide strong evidence for a specific mechanism.
- Expression data is intrinsically limited.

Cancer Genetics Branch

GETTING BEYOND GENE LISTS

- Optimal use of gene annotations.
- Optimizing use of public data.
- Incorporating data from model systems.
- Linking expression data to sequence.
- Adding other types of genome scale data.

National Human Genome Research Institute

Cancer Genetics Branch

WHAT SHOULD YOU LOOK FOR IN A CLINICAL MICROARRAY STUDY?

ARE MICROARRAY TECHNOLOGIES READY TO BE IMPLEMENTED IN CLINICAL PRACTICE?

National Human Genome Research Institute

Cancer Genetics Branch

WHAT TO LOOK FOR IN CLINICAL CORRELATIVE STUDIES USING MICROARRAYS

- WELL DEFINED QUESTION AND PATIENT SAMPLE.
- HIGH QUALITY ARRAY MEASUREMENTS (HARD TO ASSESS WITHOUT REFERENCE TO PRIMARY DATA---SHOULD BE MADE PUBLIC).
- APPROPRIATE AND RIGOROUS STATISTICAL ANALYSIS OF ARRAY DATA.
- FORMAL CLASSIFIER THAT CAN BE APPLIED TO NEW SAMPLES.
- VALIDATION SAMPLE SET.

Cancer Genetics Branch

WHAT TO LOOK FOR IN CLINICAL CORRELATIVE STUDIES USING MICROARRAYS

- **GOAL SHOULD BE TO SEEK AND VALIDATE CLINICALLY RELEVANT SIGNATURES WITHIN DEFINED PATIENT GROUPS FOR WHICH NO CURRENT FEATURES ADEQUATELY ANSWER THE CLINICAL QUESTION POSED.**

National Human Genome Research Institute

Cancer Genetics Branch

EXPRESSION PROFILING IN THE CLINIC?

PROBLEMS:

- SPECIALIZED TECHNOLOGY
- RNA IS UNSTABLE
- FROZEN TISSUE NOT PART OF USUAL OR SAMPLE FLOW

Cancer Genetics Branch

EXPRESSION PROFILING IN THE CLINIC?

OPTIONS:

- REFERENCE LABORATORIES
- RNA PRESERVATIVES
- USE OF PARAFFIN EMBEDDED MATERIALS.

National Human Genome Research Institute

Cancer Genetics Branch

EXPRESSION PROFILING IN THE CLINIC?

- COMMERCIAL TESTS BEGINNING TO APPEAR.
- NOT FDA APPROVED
- LIMITED CLINICAL VALIDATION
- ADDITIONAL CLINICAL STUDIES NEEDED

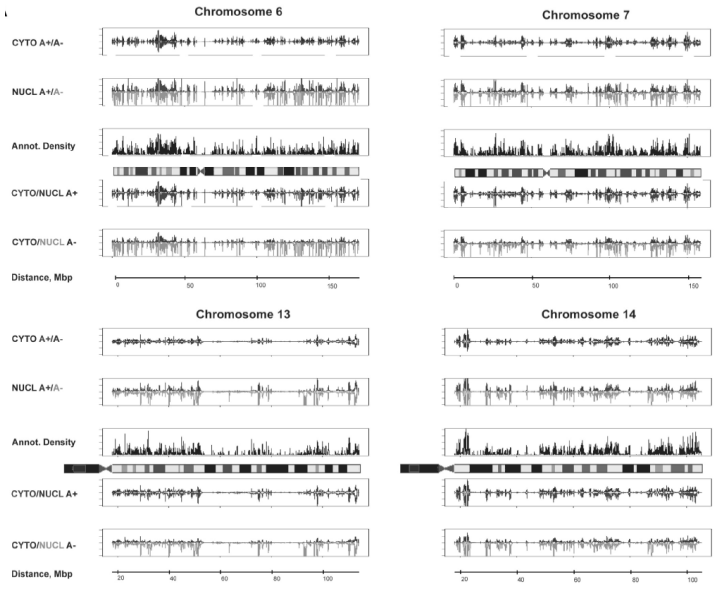
Cancer Genetics Branch

DNA Microarray Applications

- Resequencing
- Comparative Genomic Hybridization
- Gene Expression
- Transcription factor localization
- Chromatin/DNA modification

Cancer Genetics Branch

Scanning Chromosomes with Tiling Path Arrays

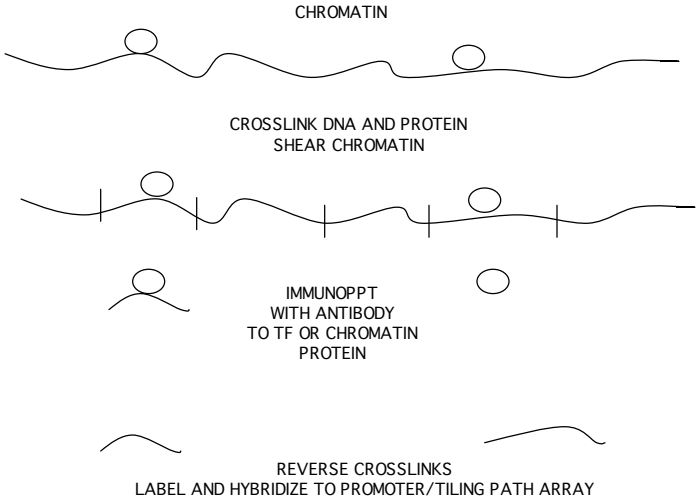


Cheng et al Science March 29, 2005

National Human Genome Research Institute

Cancer Genetics Branch

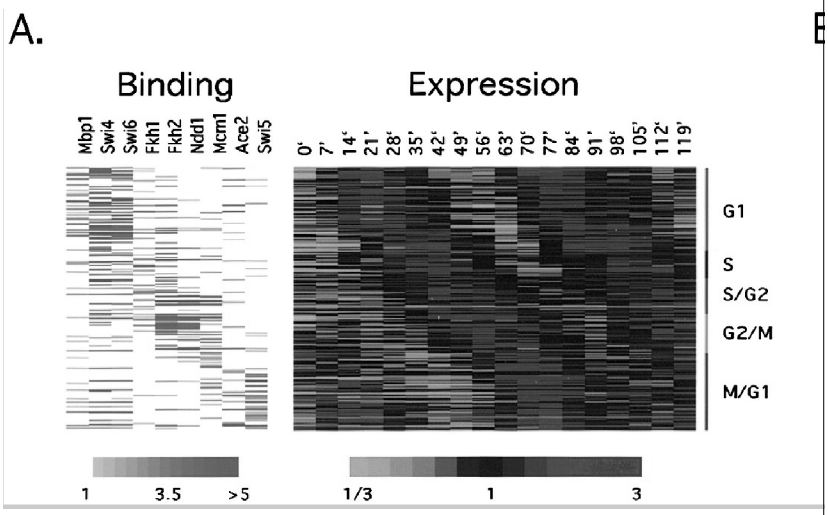
TRANSCRIPTION FACTOR LOCALIZATION ON ARRAYS



National Human Genome Research Institute

Cancer Genetics Branch

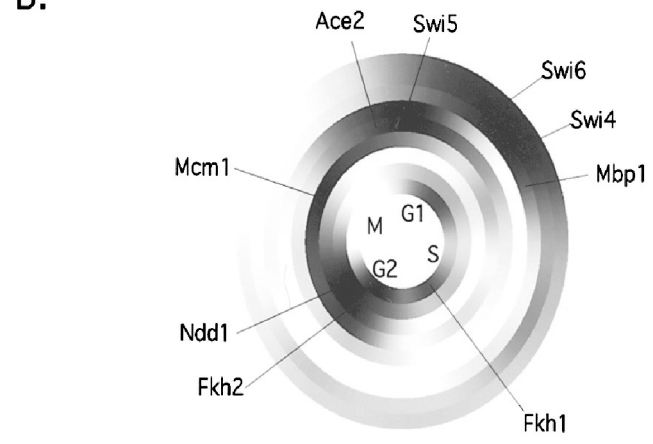
Promoter Occupancy During Yeast Cell Cycle



Simon I Cell. 2001 Sep 21;106(6):697-708.

Cancer Genetics Branch

Promoter Occupancy During Yeast Cell Cycle

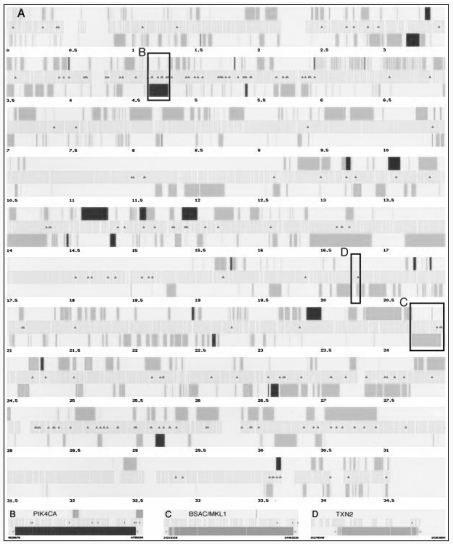


Simon I Cell. 2001 Sep 21;106(6):697-708.

National Human Genome Research Institute

Cancer Genetics Branch

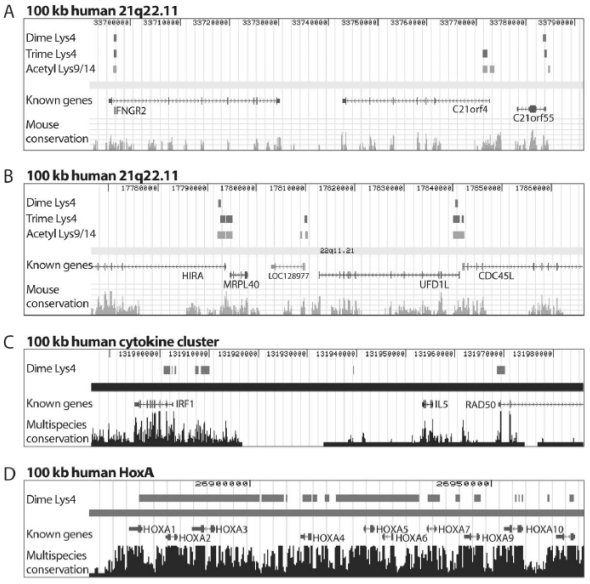
NFKB Binding to Chromosome 22



Martone et al. PNAS. 2004 100:12247.

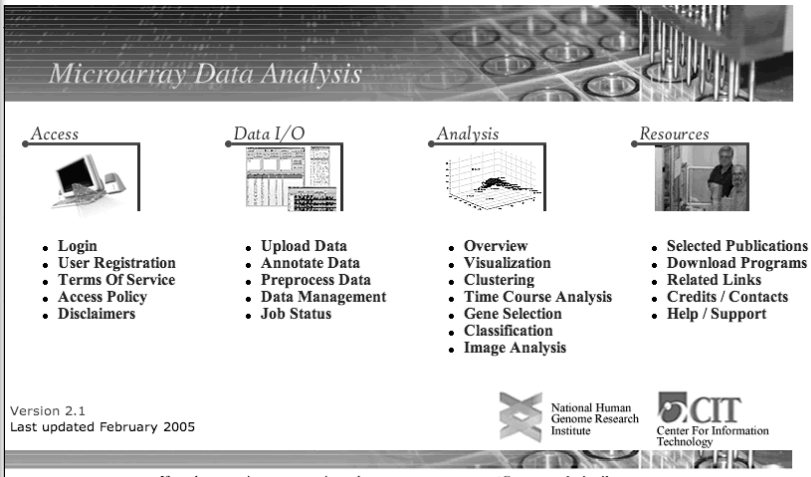
Cancer Genetics Branch

CHROMATIN MODIFICATION BY CHIP CHIP



Bernstein et al. Cell 2005 120:169.

Cancer Genetics Branch



www.arrayanalysis.nih.gov

National Human Genome Research Institute

Cancer Genetics Branch

Selected Web Sites for Microarrays

Non-Profit

NHGRI http://research.nhgri.nih.gov/microarray/
- The National Human Genome Research Institute microarray website

MGED http://www.mged.org/
- The Microarray Gene Expression Data (MGED) Society is an international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments.

NCBI http://ncbi.nih.gov/geo/
- The Gene Expression Omnibus is a gene expression and hybridization array data repository, as well as a curated, online resource for gene expression data browsing, query and retrieval. GEO was the first fully public high-throughput gene expression data repository, and became operational in July 2000.

EBI http://www.ebi.ac.uk/microarray/index.html
- The microarray informatics group at the EBI addresses the problem(s) of managing, storing and analyzing microarray data.

TIGR http://www.tigr.org/tdb/microarray/
- The Institute for Genomic Research

Academic

Stanford http://cmgm.stanford.edu/pbrown/mguide/
- The Brown Lab's complete guide to microarraying for the molecular biologist.

Stanford http://genome-www5.stanford.edu/MicroArray/SMD/
- The Stanford microarray database

UCSF http://www.microarrays.org/index.html
- A public source for microarray protocols and software.

MIT http://www-genome.wi.mit.edu/cancer/
- Focuses on genomic and computational solutions to problems in cancer biology and cancer medicine.

National Human Genome Research Institute