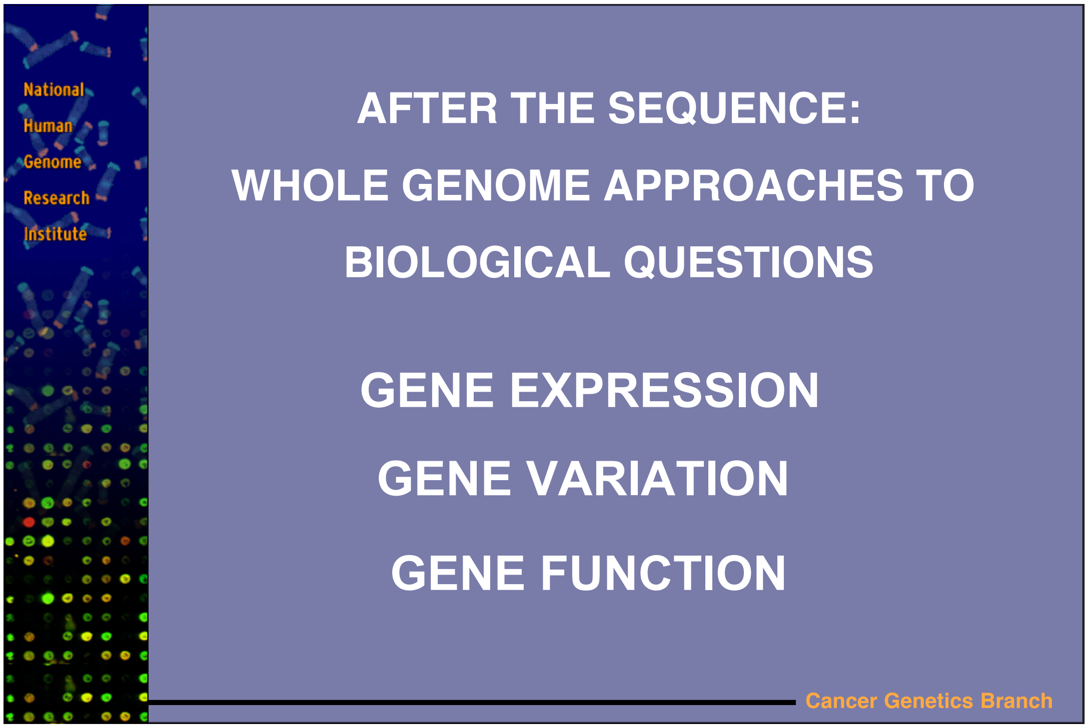
National
Human
Genome
Research
Institute

MICROARRAY TECHNOLOGY

Paul Meltzer, Ph.D.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

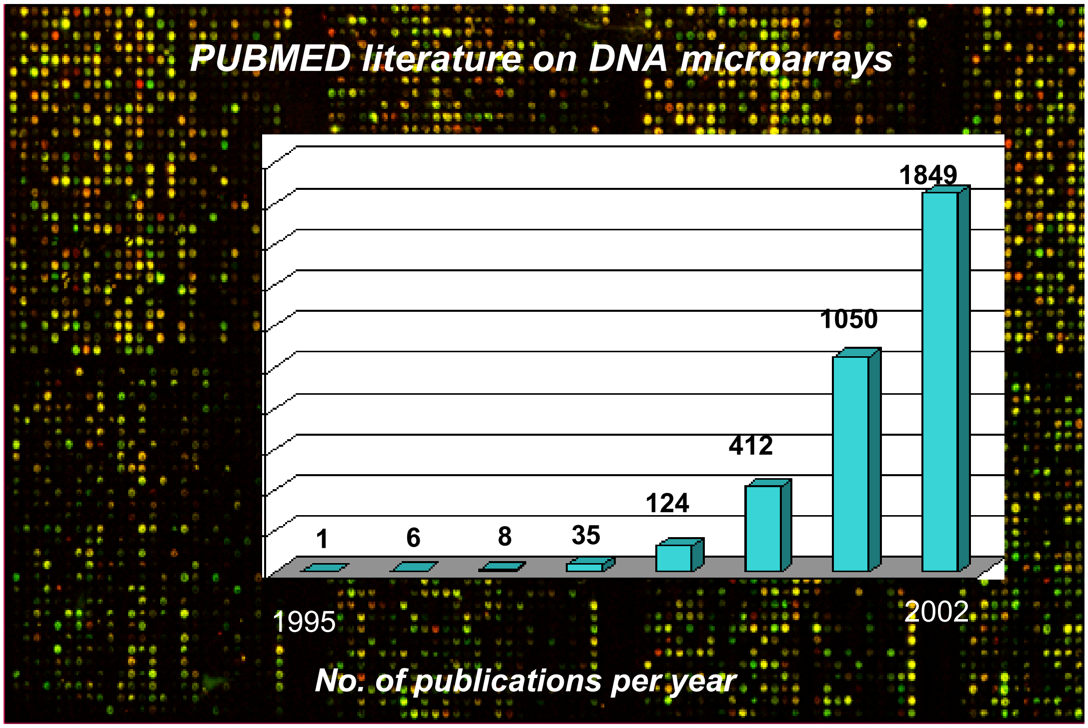
AFTER THE SEQUENCE:
WHOLE GENOME APPROACHES TO BIOLOGICAL QUESTIONS

GENE EXPRESSION

GENE VARIATION

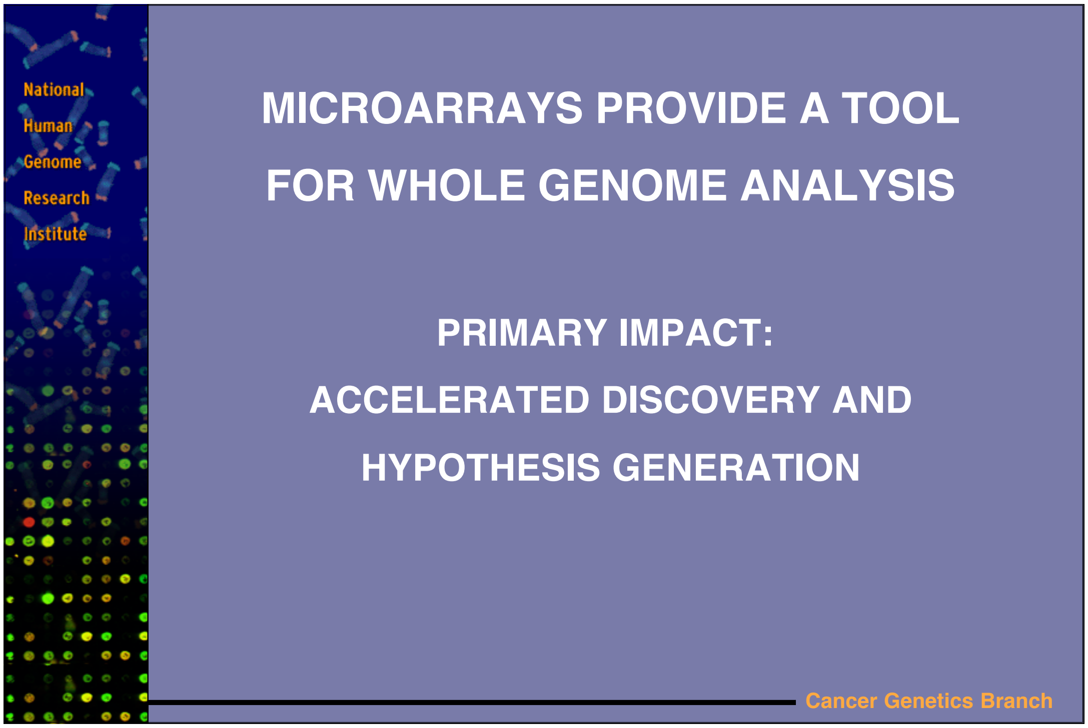
GENE FUNCTION

Cancer Genetics Branch



PUBMED literature on DNA microarrays

No. of publications per year



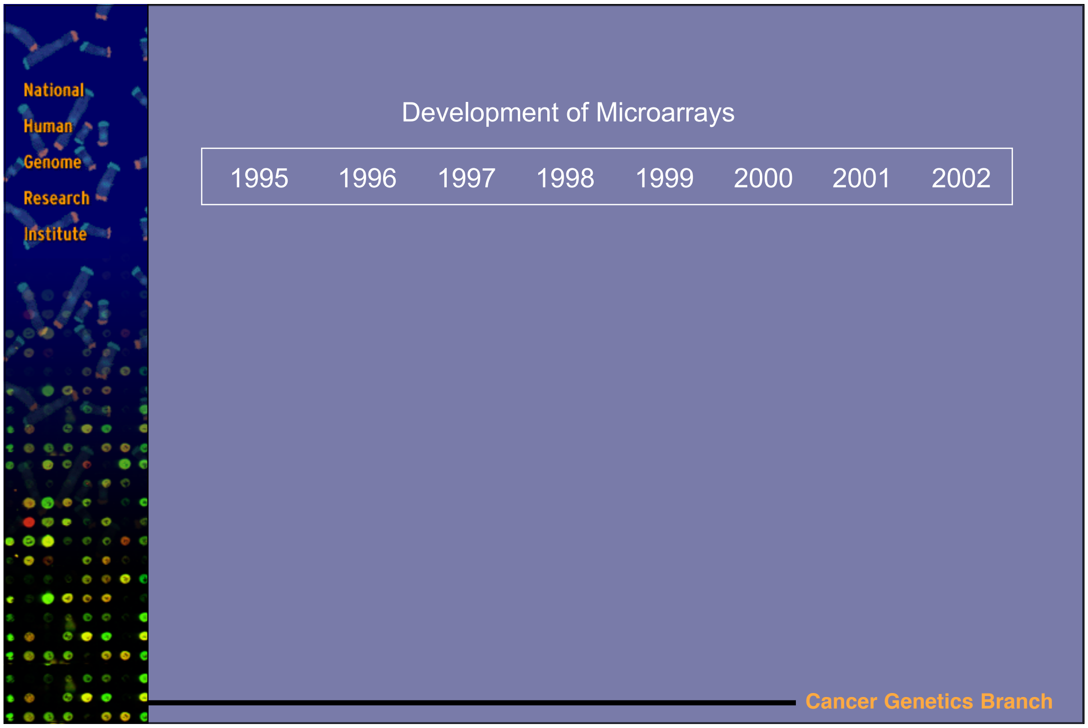
National
Human
Genome
Research
Institute

MICROARRAYS PROVIDE A TOOL FOR WHOLE GENOME ANALYSIS

PRIMARY IMPACT:

ACCELERATED DISCOVERY AND HYPOTHESIS GENERATION

Cancer Genetics Branch

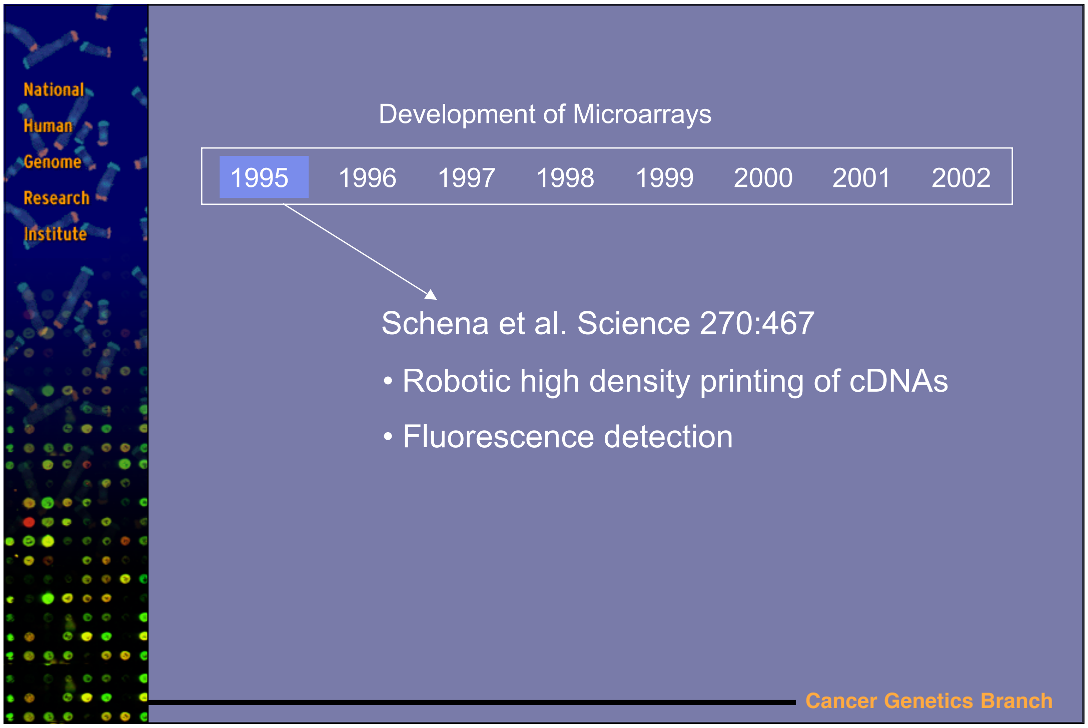


National
Human
Genome
Research
Institute

Development of Microarrays

1995 1996 1997 1998 1999 2000 2001 2002

Cancer Genetics Branch



National
Human
Genome
Research
Institute

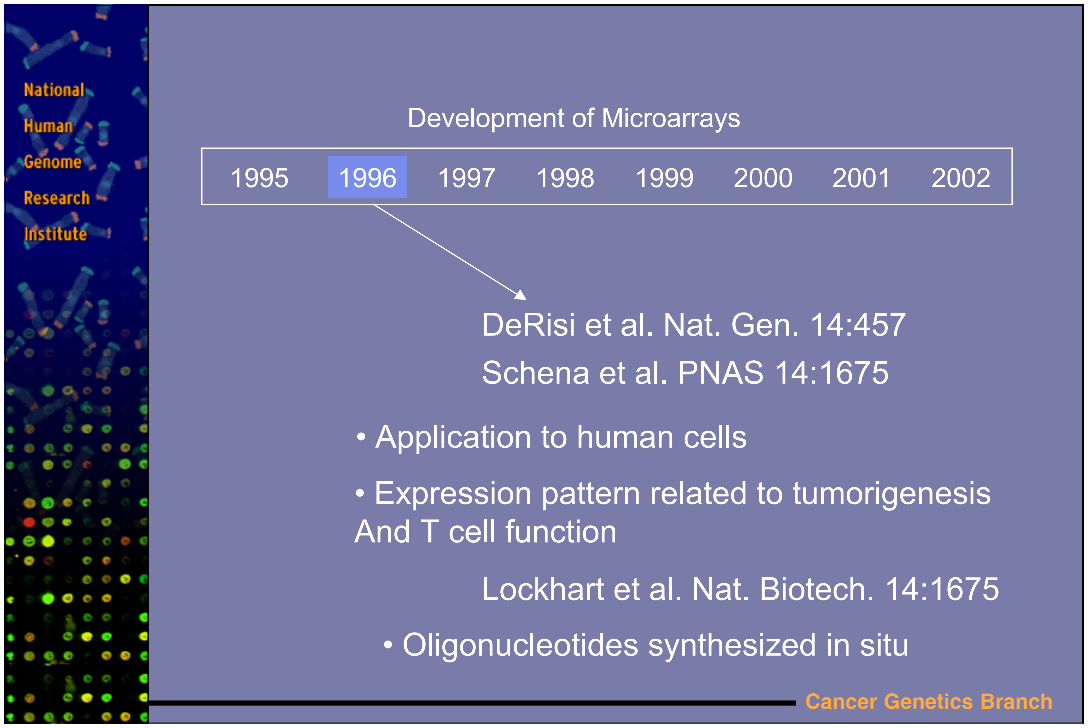
Development of Microarrays

1995 1996 1997 1998 1999 2000 2001 2002

Schena et al. Science 270:467

- Robotic high density printing of cDNAs
- Fluorescence detection

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Development of Microarrays

1995 1996 1997 1998 1999 2000 2001 2002

DeRisi et al. Nat. Gen. 14:457

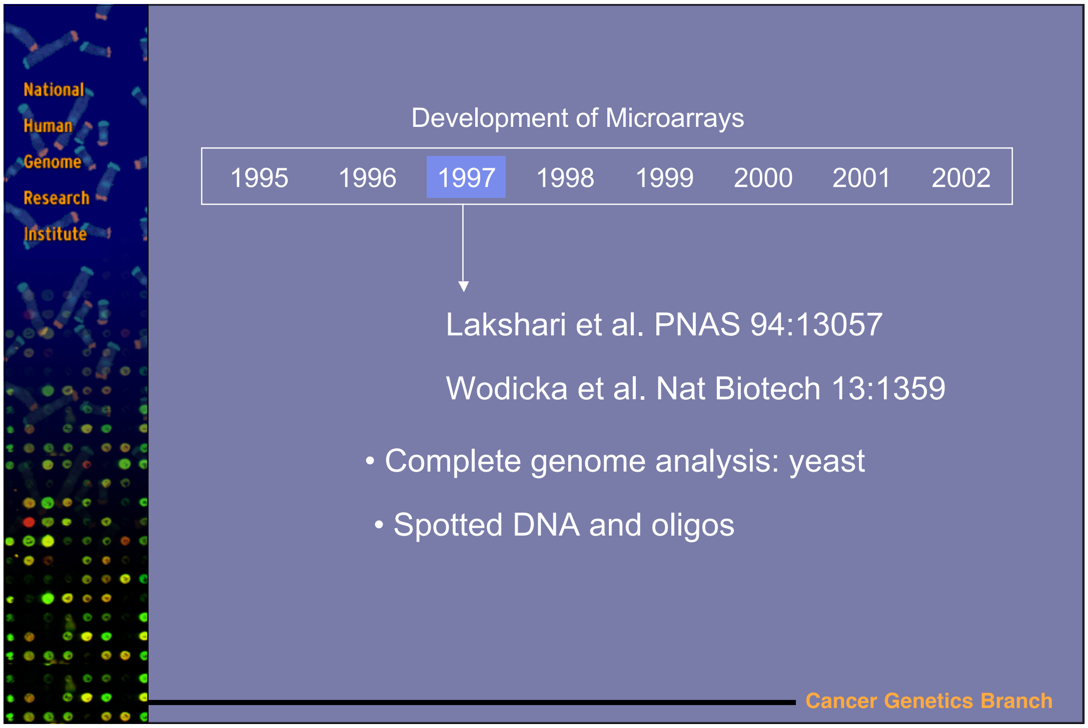
Schena et al. PNAS 14:1675

- Application to human cells
- Expression pattern related to tumorigenesis And T cell function

Lockhart et al. Nat. Biotech. 14:1675

- Oligonucleotides synthesized in situ

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Development of Microarrays

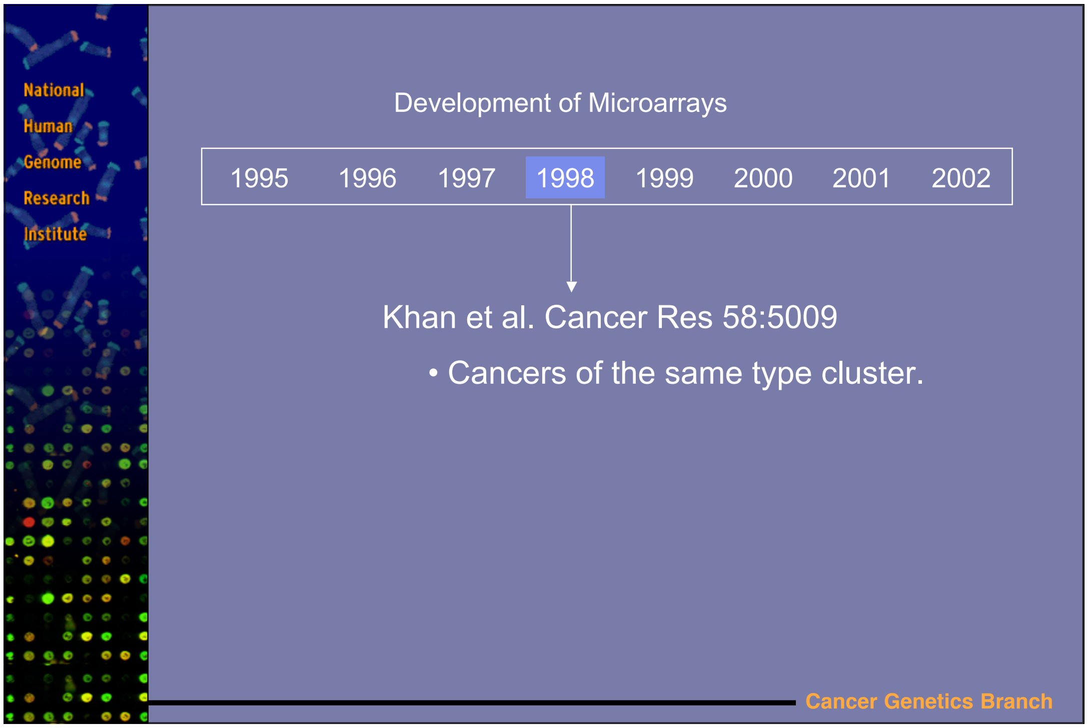
1995 1996 1997 1998 1999 2000 2001 2002

Lakshari et al. PNAS 94:13057

Wodicka et al. Nat Biotech 13:1359

- Complete genome analysis: yeast
- Spotted DNA and oligos

Cancer Genetics Branch



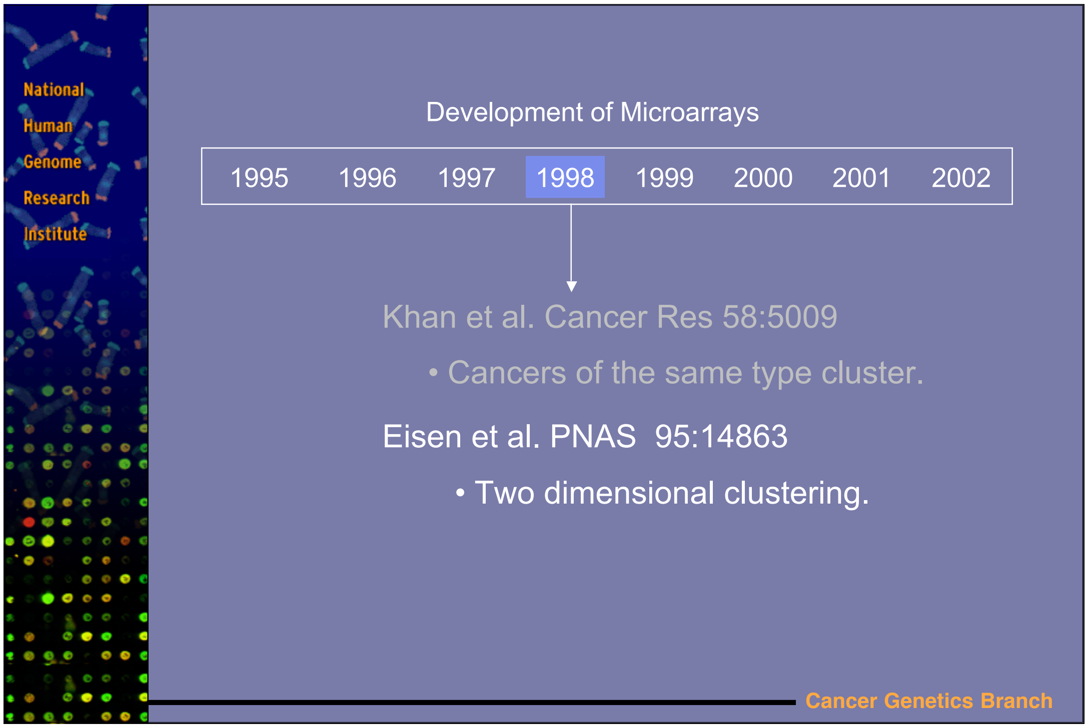
Development of Microarrays

1995 1996 1997 **1998** 1999 2000 2001 2002

Khan et al. Cancer Res 58:5009

- Cancers of the same type cluster.

Cancer Genetics Branch



Development of Microarrays

1995 1996 1997 **1998** 1999 2000 2001 2002

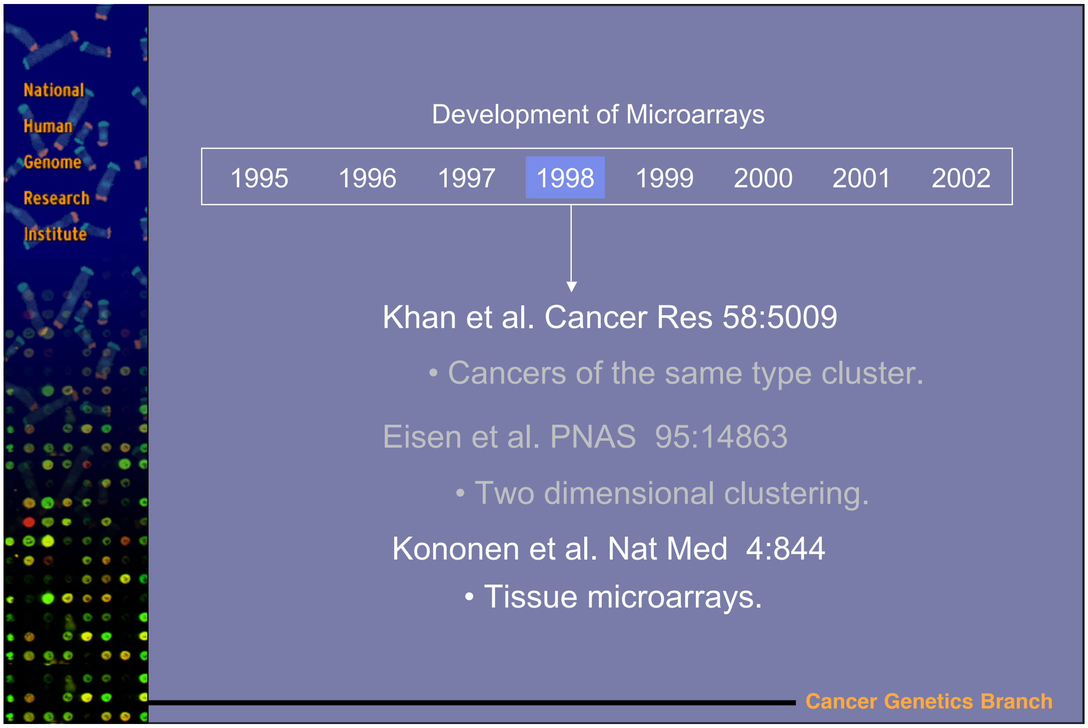
Khan et al. Cancer Res 58:5009

- Cancers of the same type cluster.

Eisen et al. PNAS 95:14863

- Two dimensional clustering.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Development of Microarrays

1995 1996 1997 1998 1999 2000 2001 2002

Khan et al. Cancer Res 58:5009

- Cancers of the same type cluster.

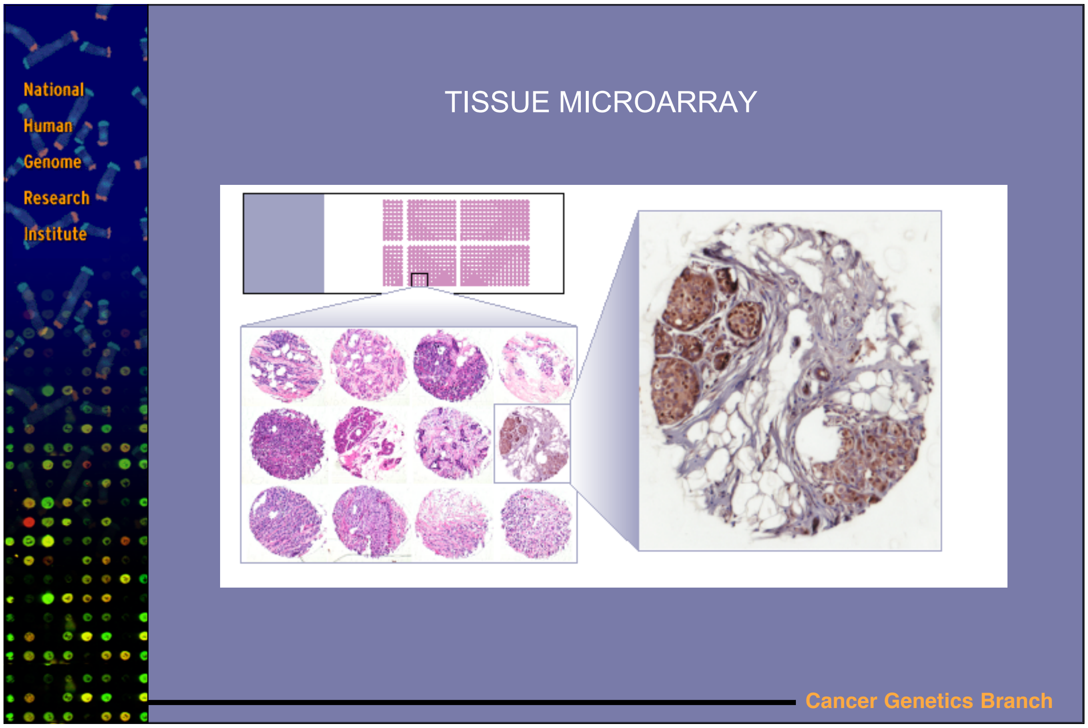
Eisen et al. PNAS 95:14863

- Two dimensional clustering.

Kononen et al. Nat Med 4:844

- Tissue microarrays.

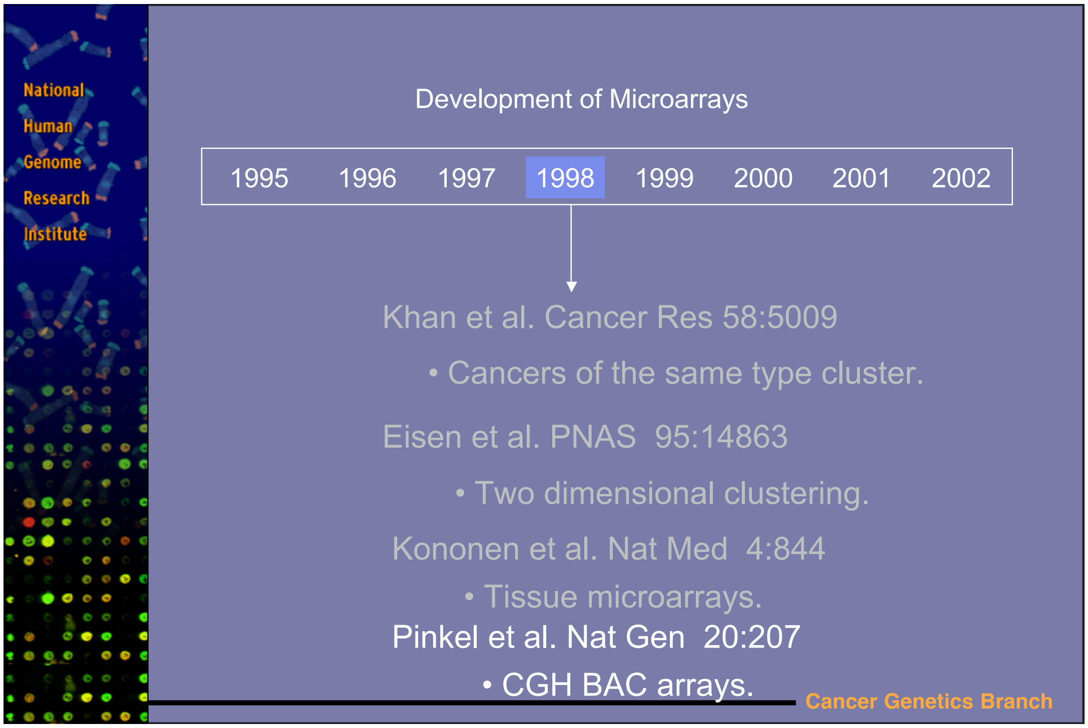
Cancer Genetics Branch



National
Human
Genome
Research
Institute

TISSUE MICROARRAY

Cancer Genetics Branch



Development of Microarrays

1995 1996 1997 **1998** 1999 2000 2001 2002

Khan et al. Cancer Res 58:5009

- Cancers of the same type cluster.

Eisen et al. PNAS 95:14863

- Two dimensional clustering.

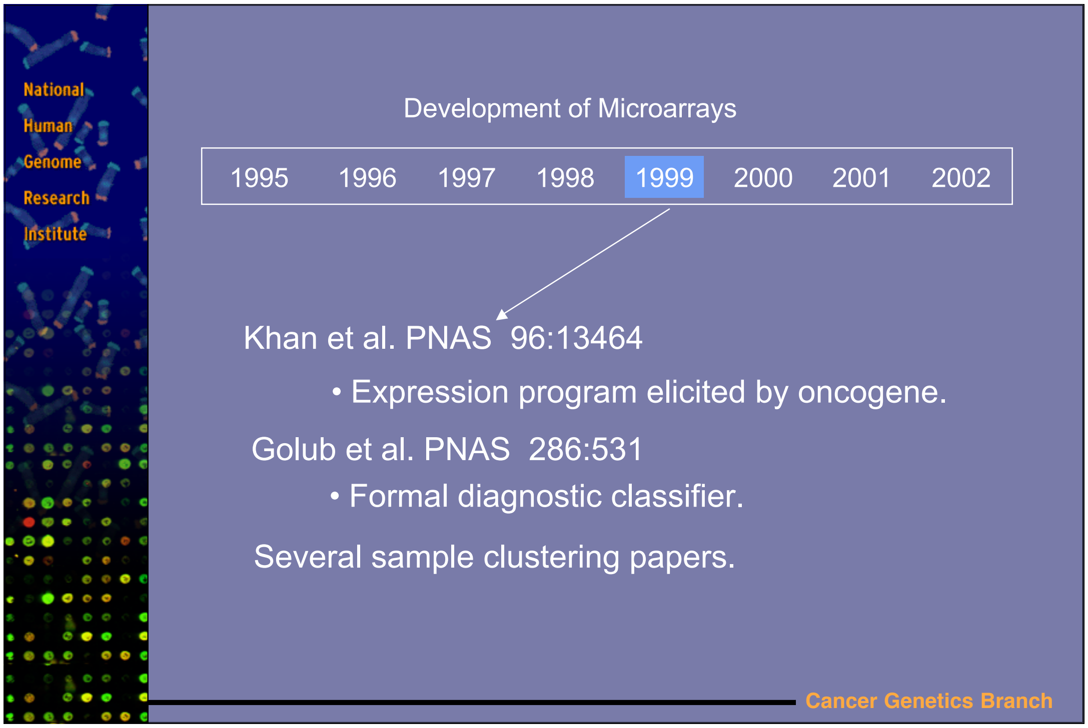
Kononen et al. Nat Med 4:844

- Tissue microarrays.

Pinkel et al. Nat Gen 20:207

- CGH BAC arrays.

Cancer Genetics Branch



Development of Microarrays

1995 1996 1997 1998 **1999** 2000 2001 2002

Khan et al. PNAS 96:13464

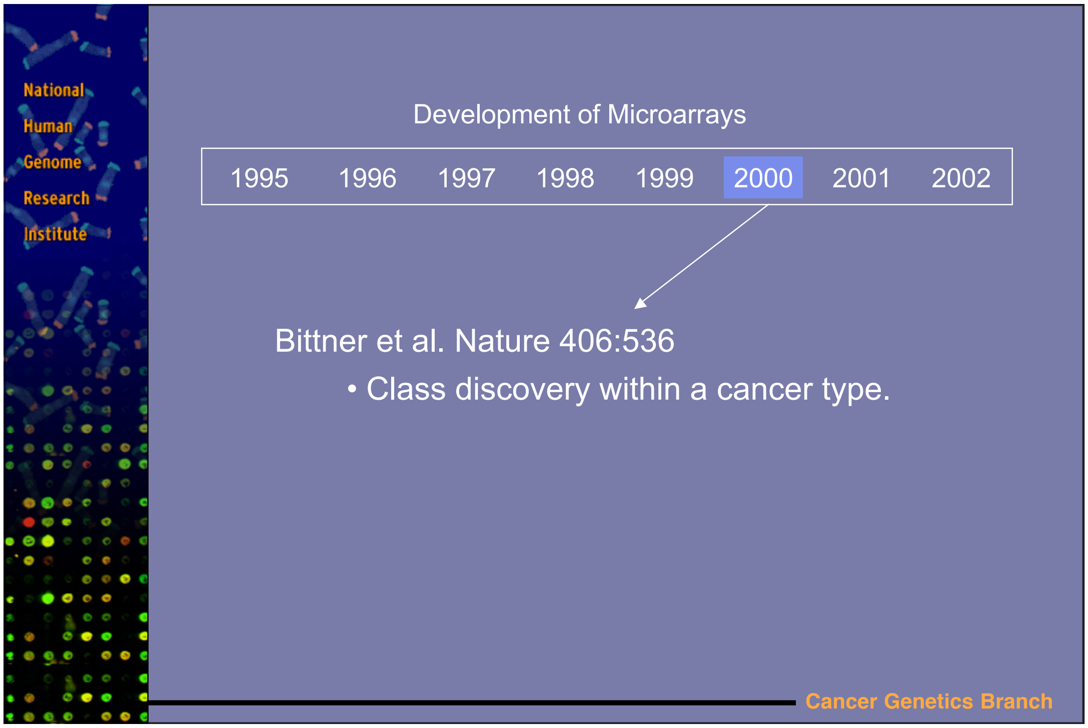
- Expression program elicited by oncogene.

Golub et al. PNAS 286:531

- Formal diagnostic classifier.

Several sample clustering papers.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

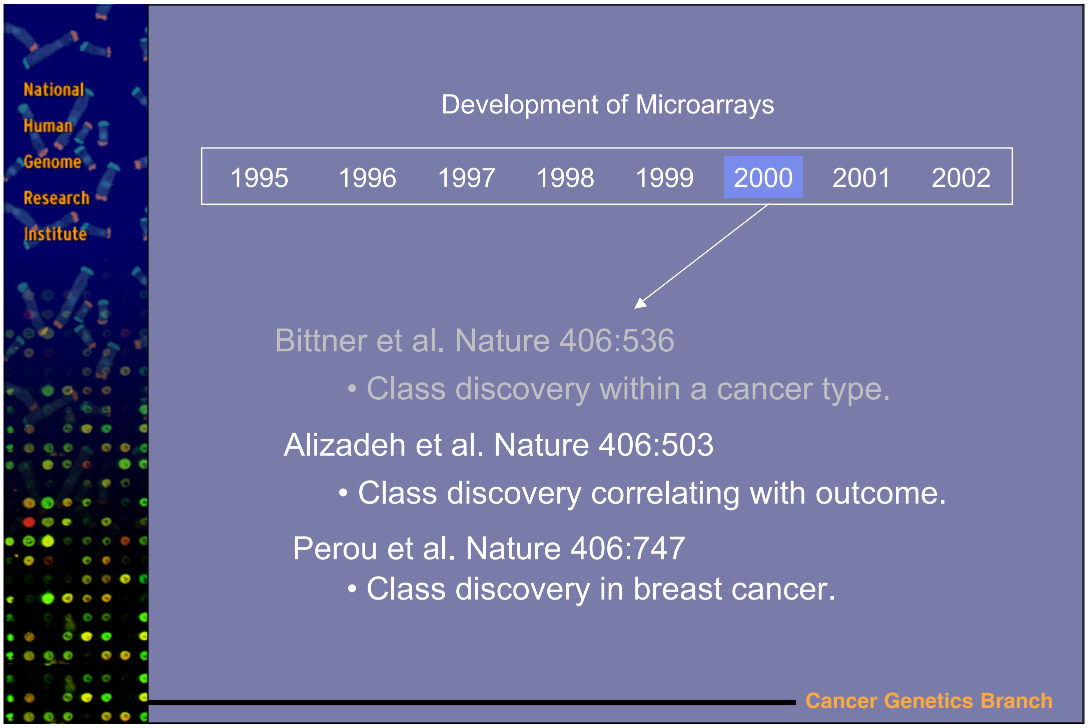
Development of Microarrays

1995 1996 1997 1998 1999 **2000** 2001 2002

Bittner et al. Nature 406:536

- Class discovery within a cancer type.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Development of Microarrays

1995 1996 1997 1998 1999 **2000** 2001 2002

Bittner et al. Nature 406:536

- Class discovery within a cancer type.

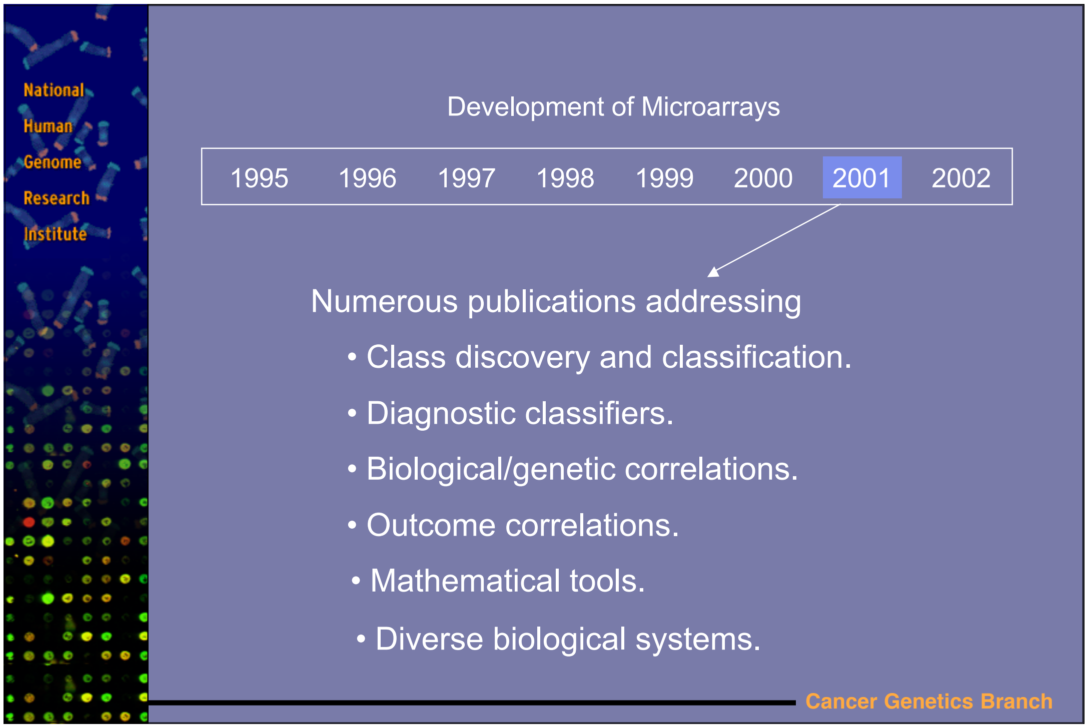
Alizadeh et al. Nature 406:503

- Class discovery correlating with outcome.

Perou et al. Nature 406:747

- Class discovery in breast cancer.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

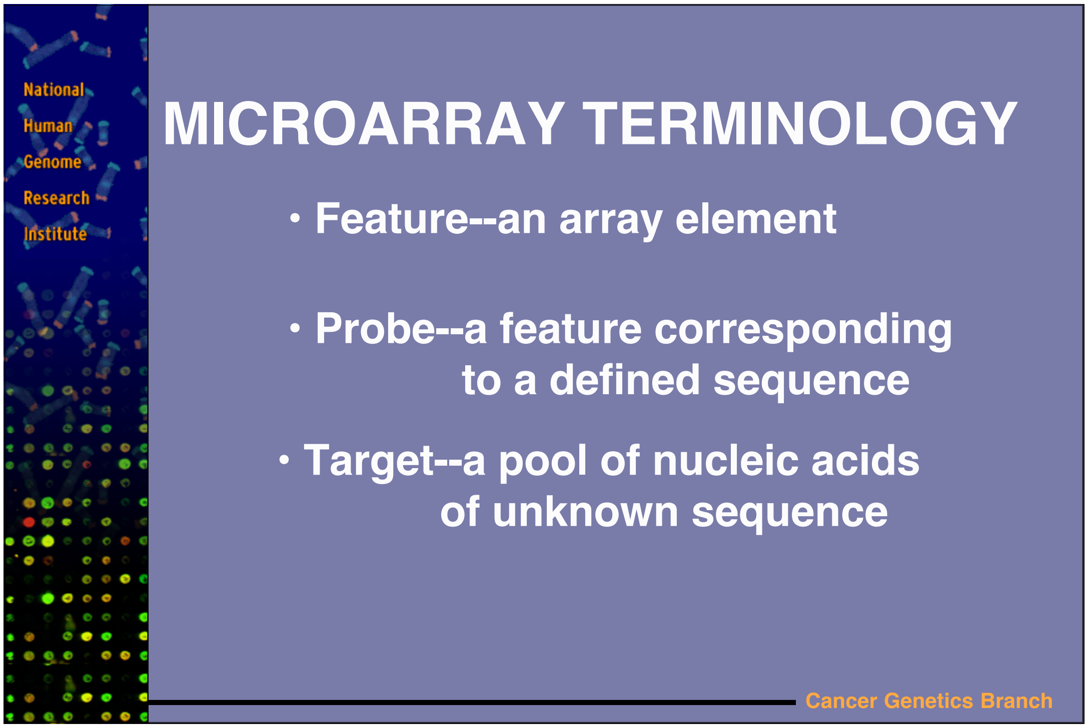
Development of Microarrays

1995 1996 1997 1998 1999 2000 2001 2002

Numerous publications addressing

- Class discovery and classification.
- Diagnostic classifiers.
- Biological/genetic correlations.
- Outcome correlations.
- Mathematical tools.
- Diverse biological systems.

Cancer Genetics Branch

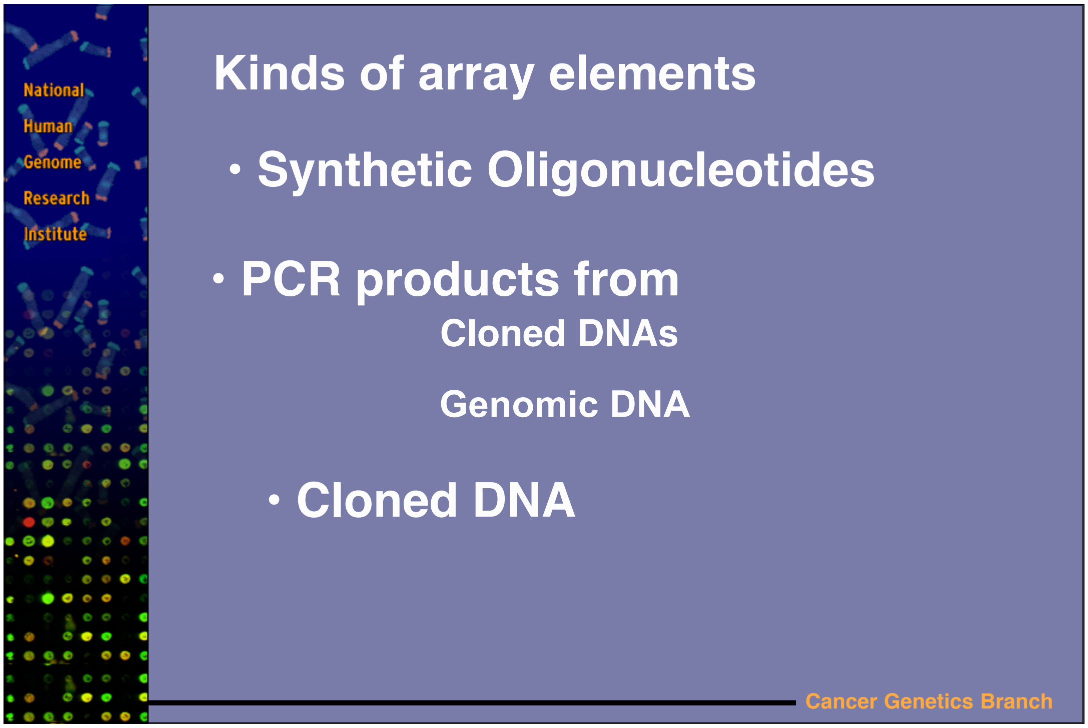


National
Human
Genome
Research
Institute

MICROARRAY TERMINOLOGY

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

Cancer Genetics Branch

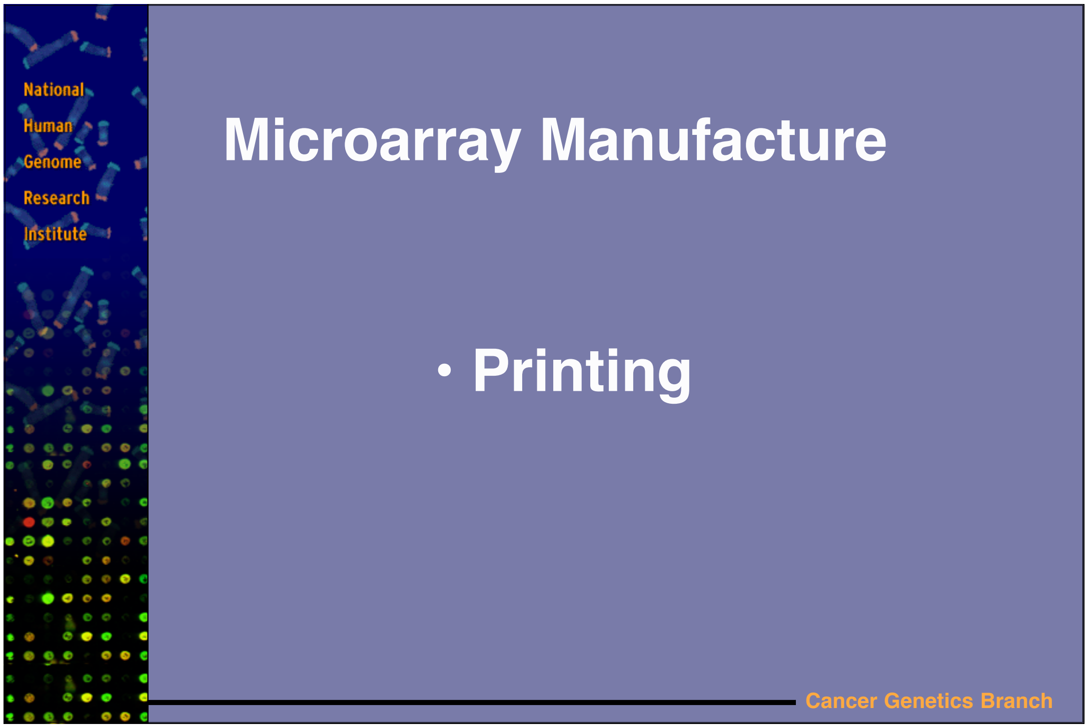


National
Human
Genome
Research
Institute

Kinds of array elements

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

Cancer Genetics Branch

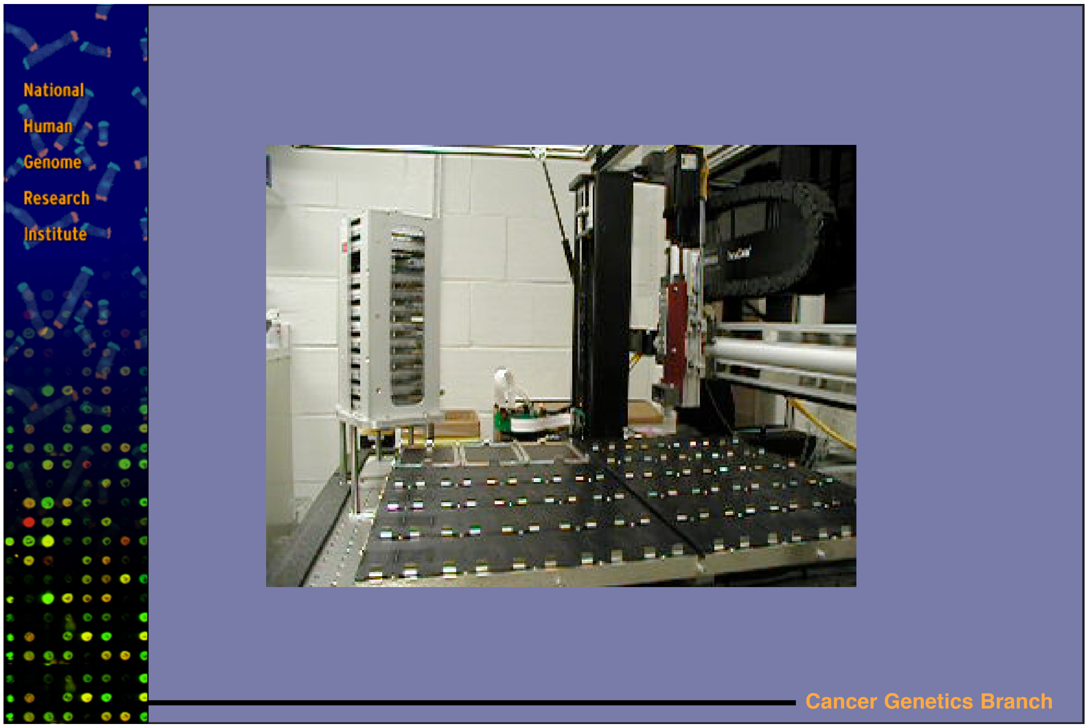


National
Human
Genome
Research
Institute

Microarray Manufacture

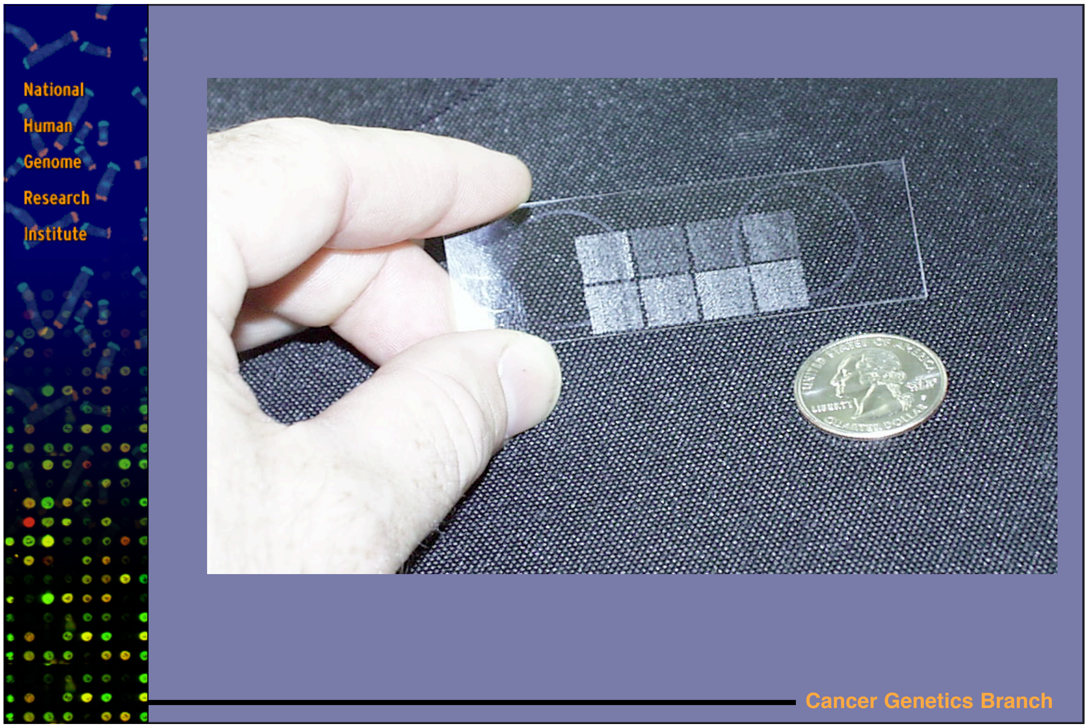
- Printing

Cancer Genetics Branch



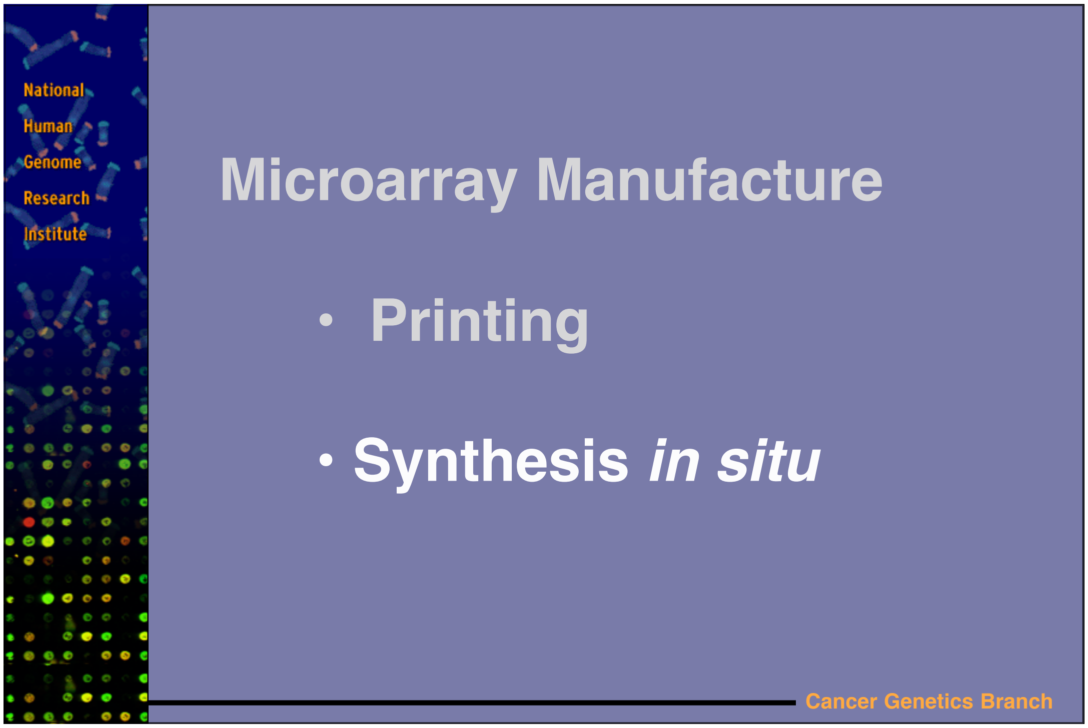
National
Human
Genome
Research
Institute

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Cancer Genetics Branch

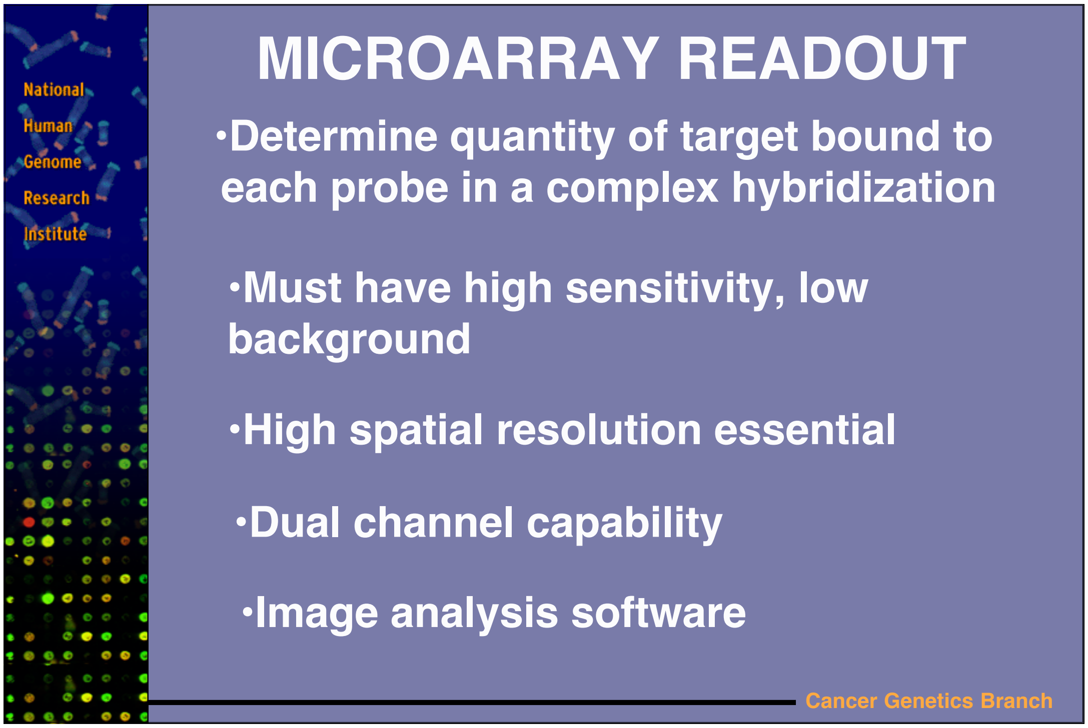


National
Human
Genome
Research
Institute

Microarray Manufacture

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

Cancer Genetics Branch

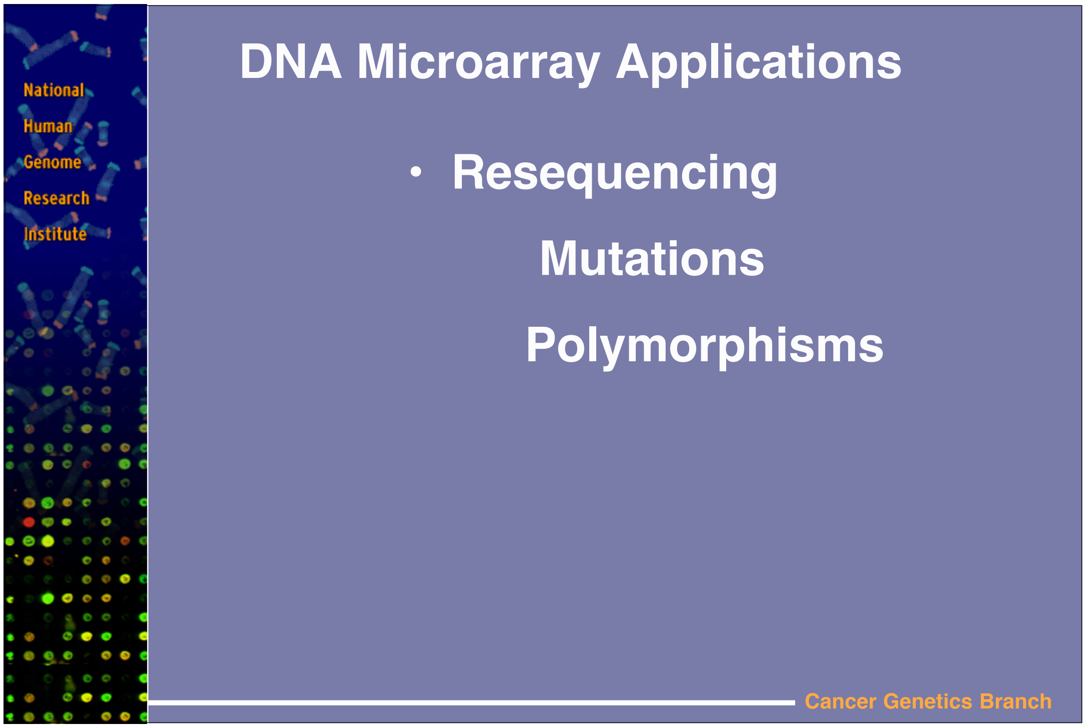


National
Human
Genome
Research
Institute

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**
- **Image analysis software**

Cancer Genetics Branch



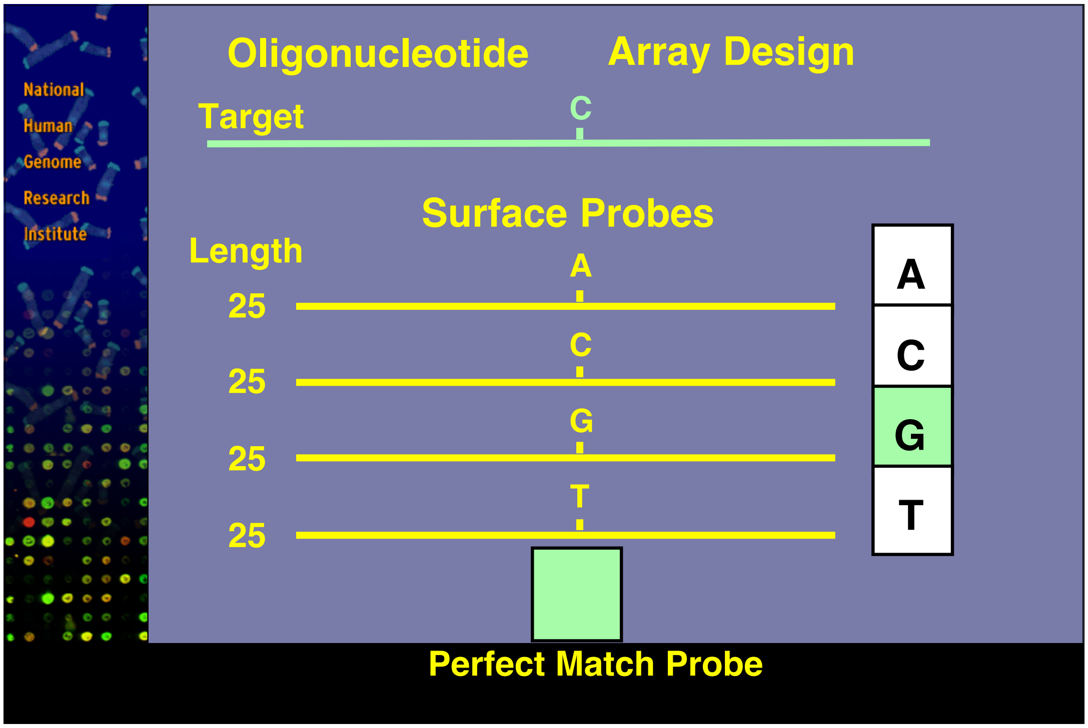
DNA Microarray Applications

- Resequencing

Mutations

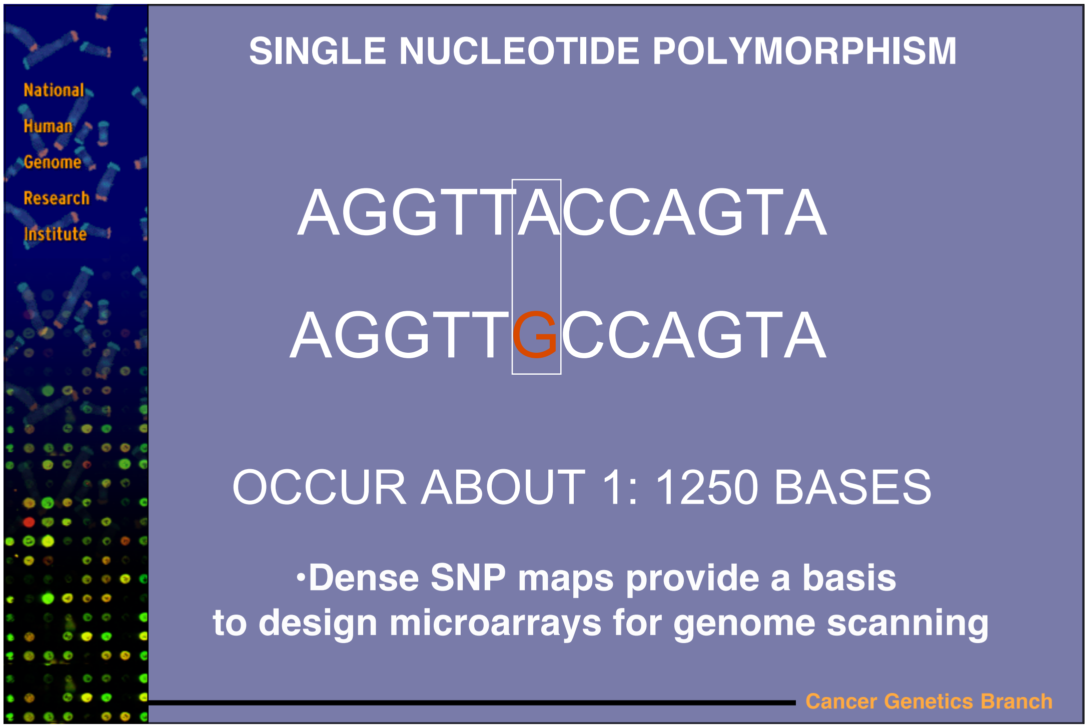
Polymorphisms

Cancer Genetics Branch



Oligonucleotide Array Design

Perfect Match Probe



SINGLE NUCLEOTIDE POLYMORPHISM

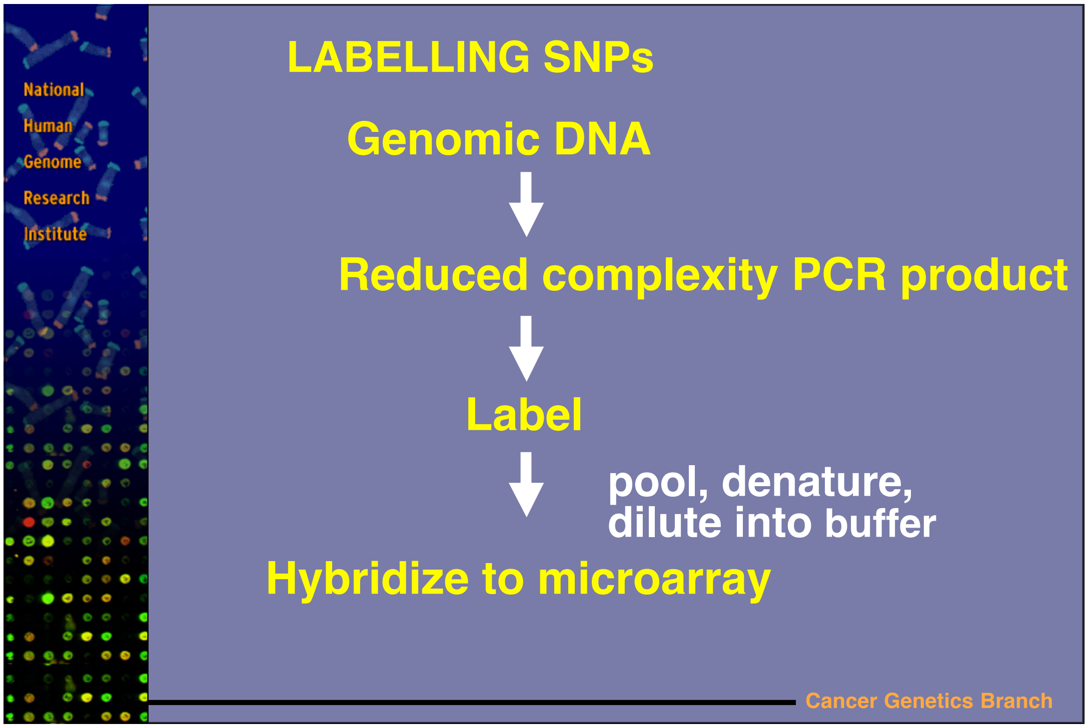
AGGTTACCAGTA

AGGTTGCCAGTA

OCCUR ABOUT 1: 1250 BASES

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

Cancer Genetics Branch



LABELLING SNPs

Genomic DNA

↓

Reduced complexity PCR product

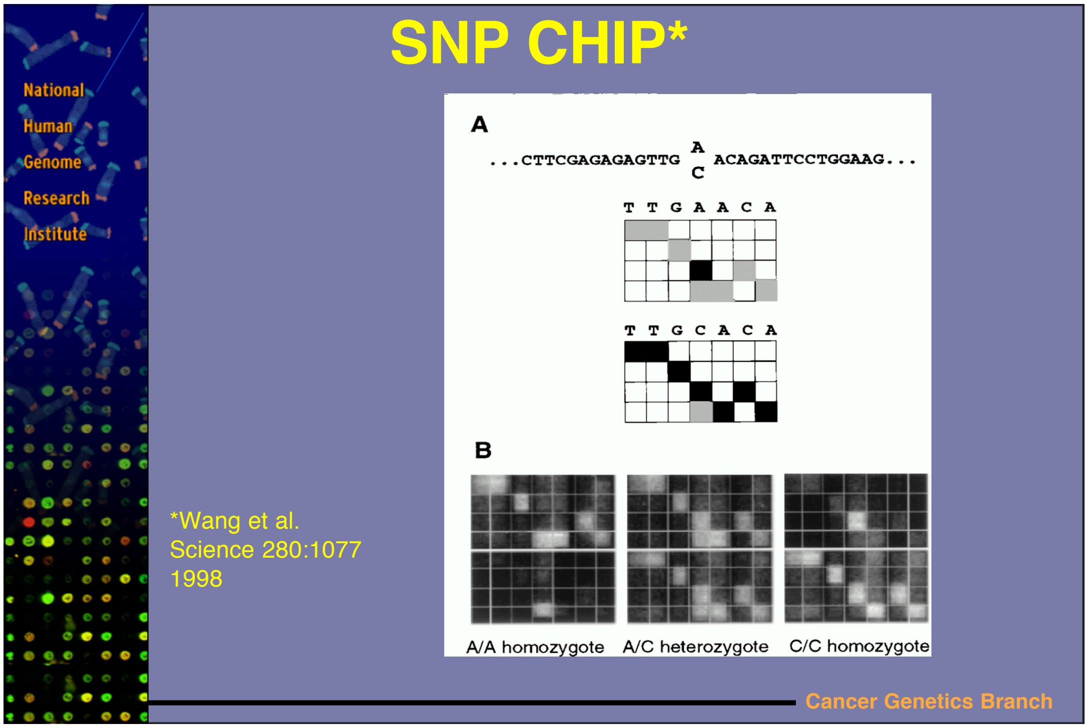
↓

Label

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

Hybridize to microarray

Cancer Genetics Branch



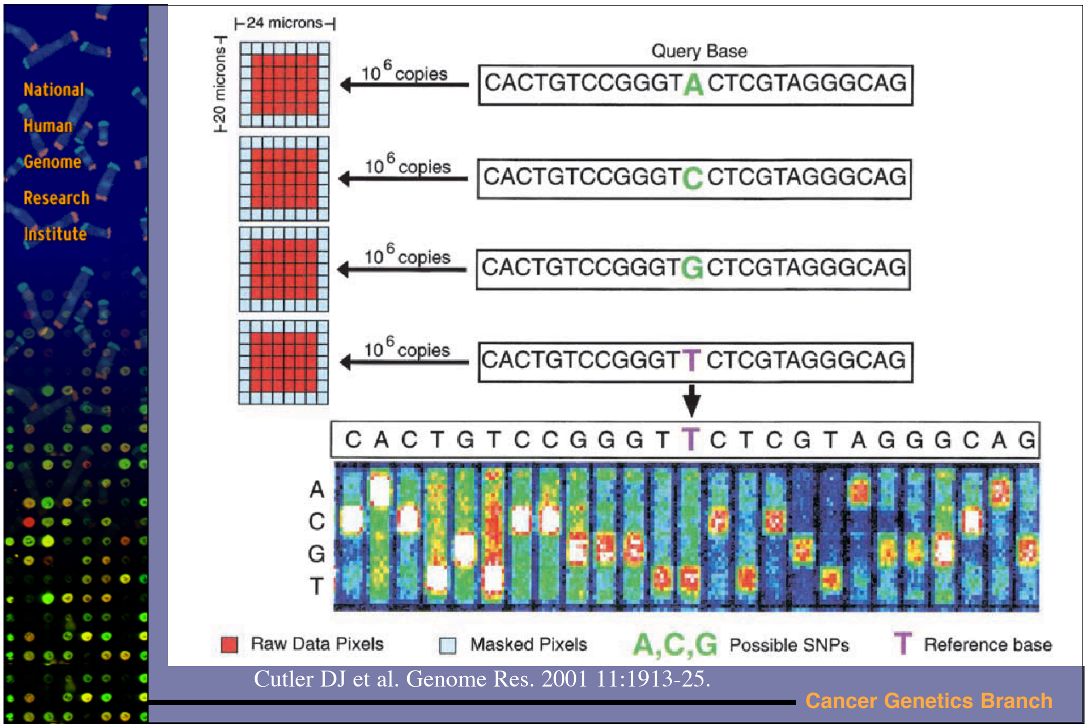
National
Human
Genome
Research
Institute

SNP CHIP*

A/A homozygote A/C heterozygote C/C homozygote

*Wang et al.
Science 280:1077
1998

Cancer Genetics Branch

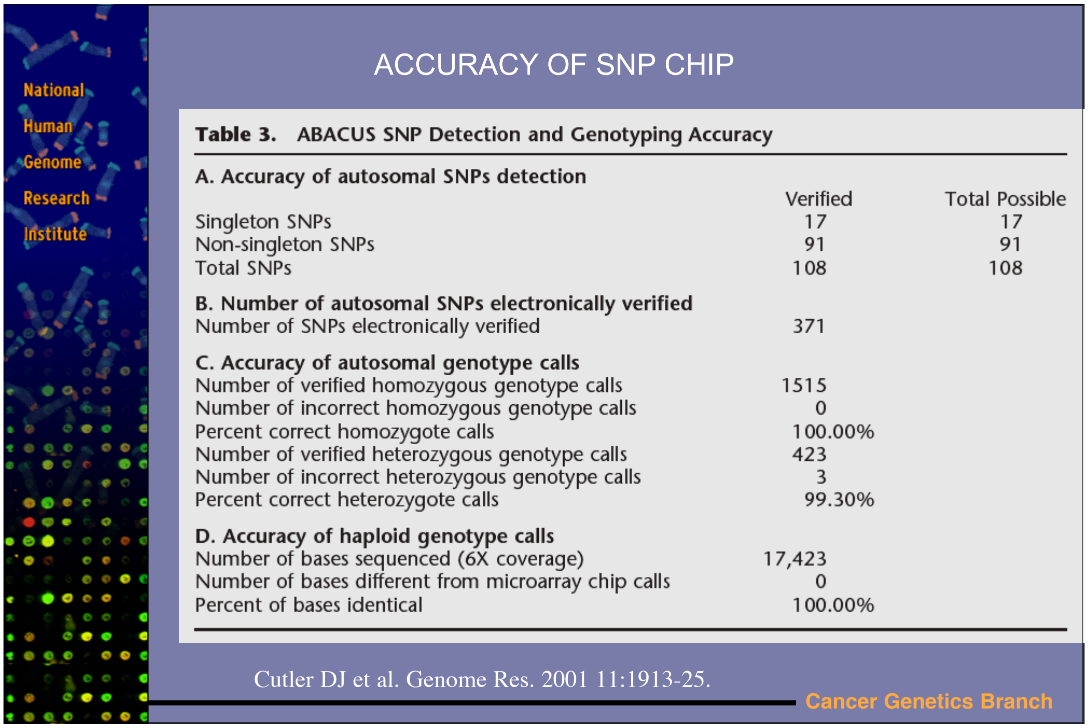


National
Human
Genome
Research
Institute

🟥 Raw Data Pixels ☐ Masked Pixels A,C,G Possible SNPs T Reference base

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

Cancer Genetics Branch



National
Human
Genome
Research
Institute

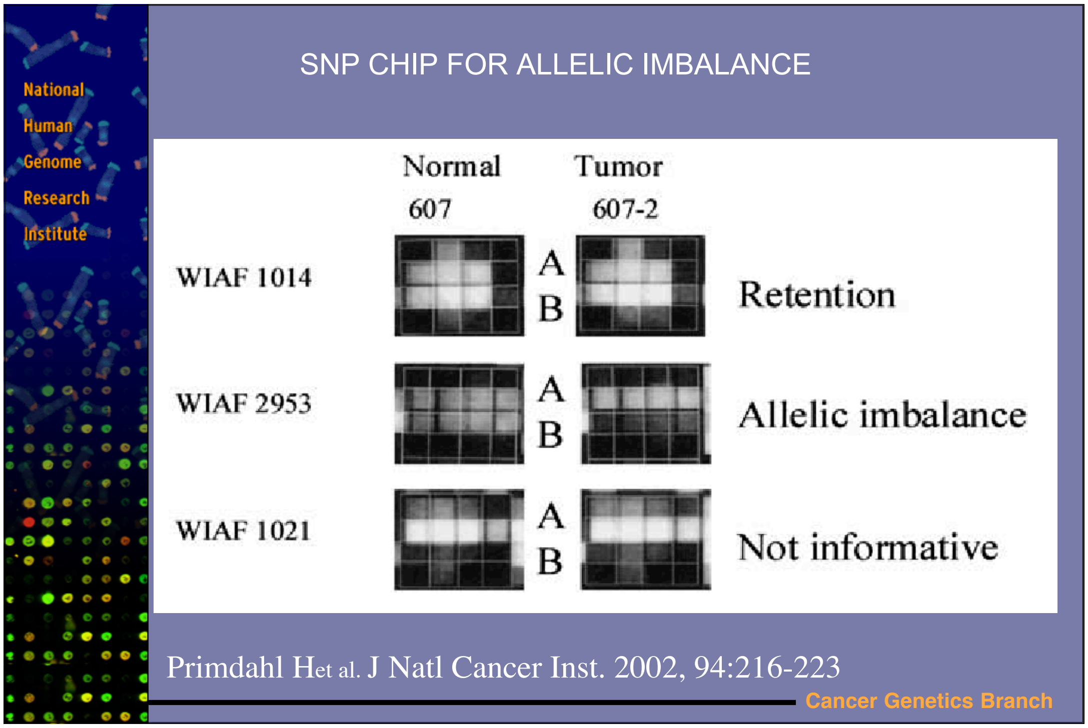
ACCURACY OF SNP CHIP

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

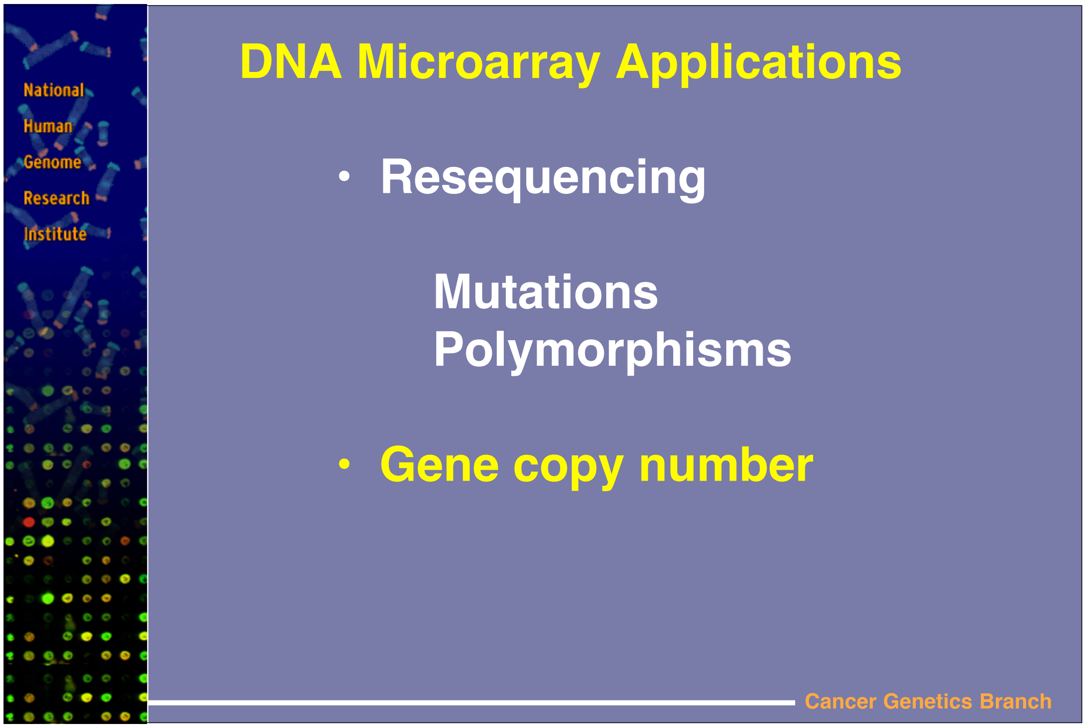


National
Human
Genome
Research
Institute

SNP CHIP FOR ALLELIC IMBALANCE

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

Cancer Genetics Branch



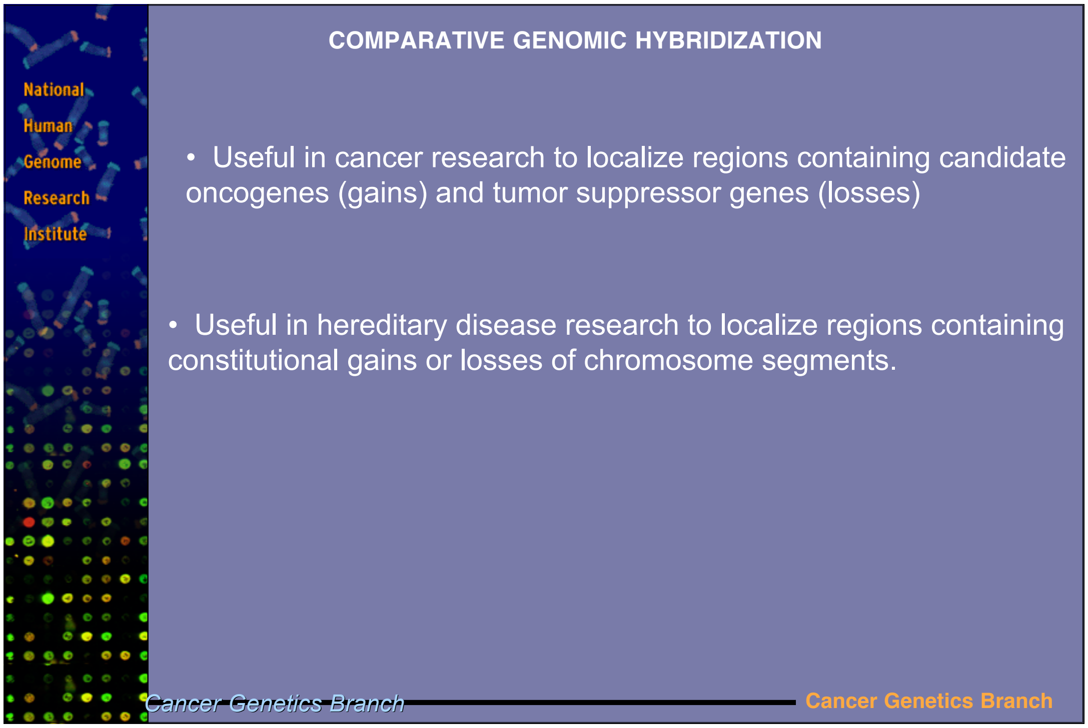
DNA Microarray Applications

- **Resequencing**

  Mutations
  Polymorphisms

- **Gene copy number**

Cancer Genetics Branch



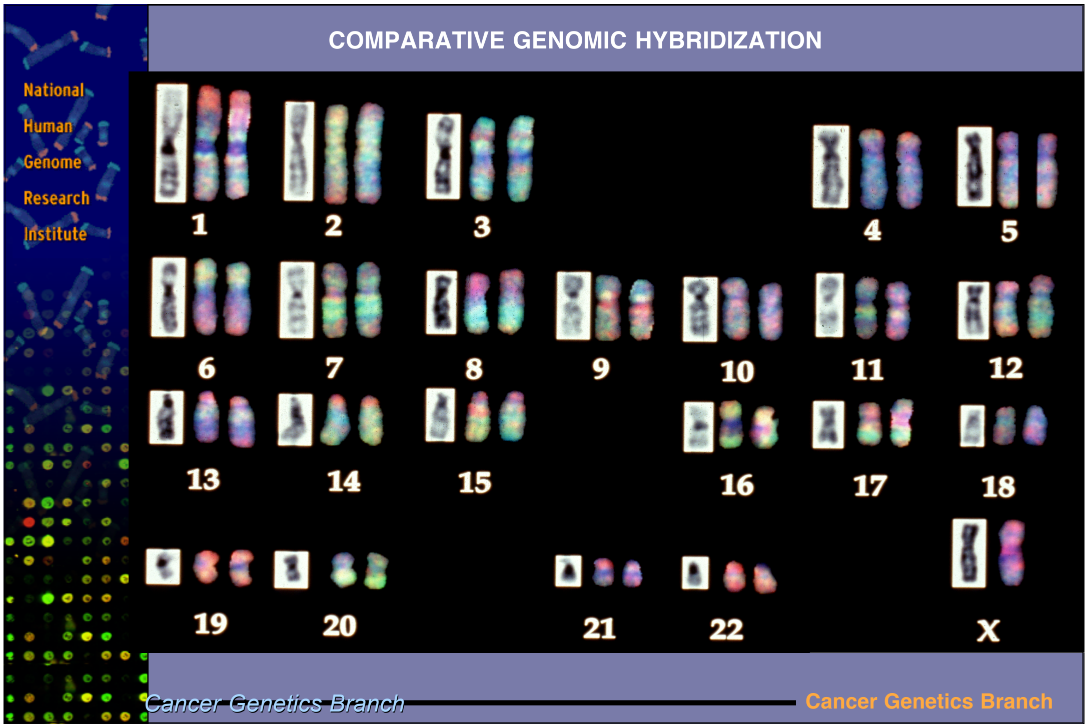
COMPARATIVE GENOMIC HYBRIDIZATION

- 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

Cancer Genetics Branch

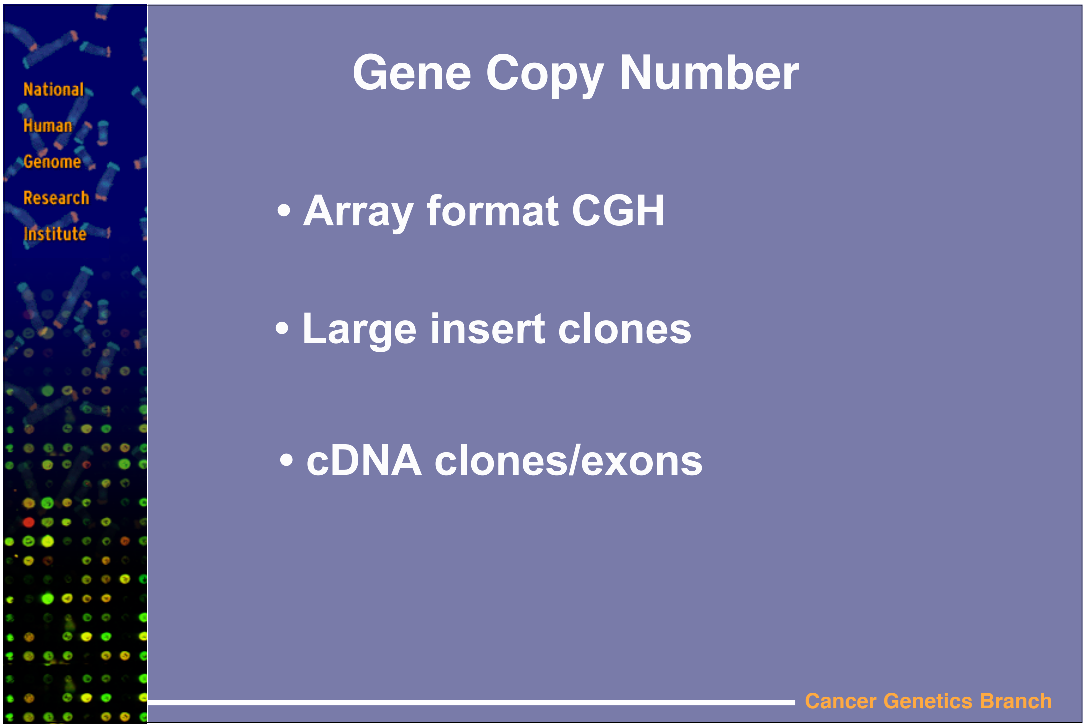


National
Human
Genome
Research
Institute

COMPARATIVE GENOMIC HYBRIDIZATION

Cancer Genetics Branch

Cancer Genetics Branch

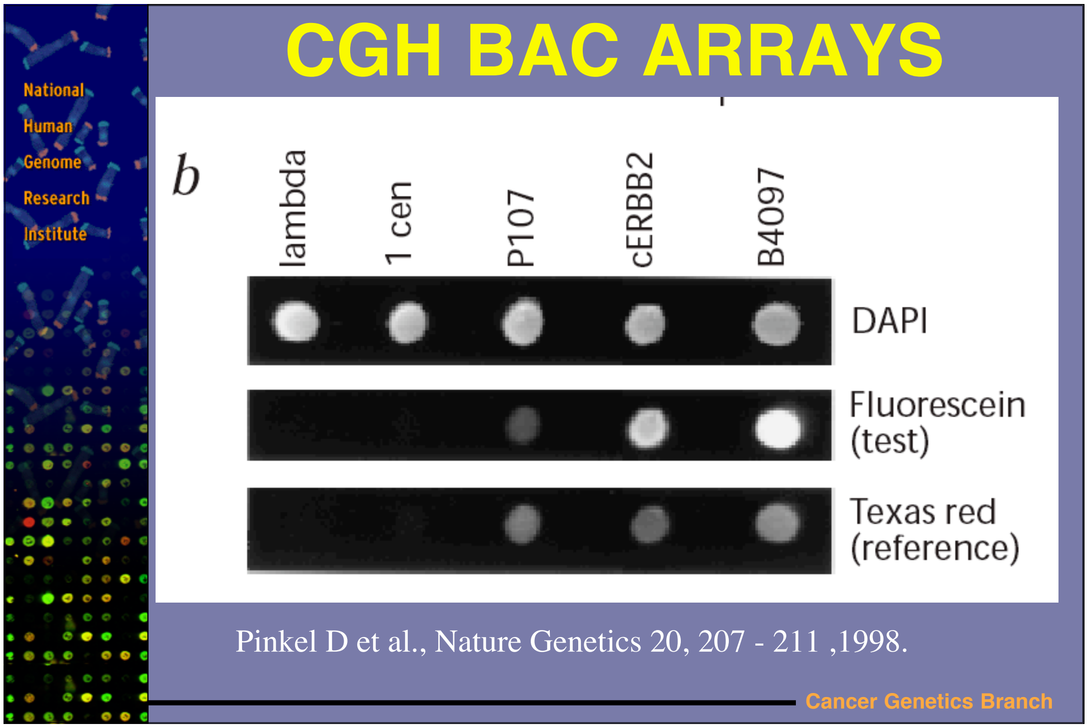


National
Human
Genome
Research
Institute

Gene Copy Number

- Array format CGH
- Large insert clones
- cDNA clones/exons

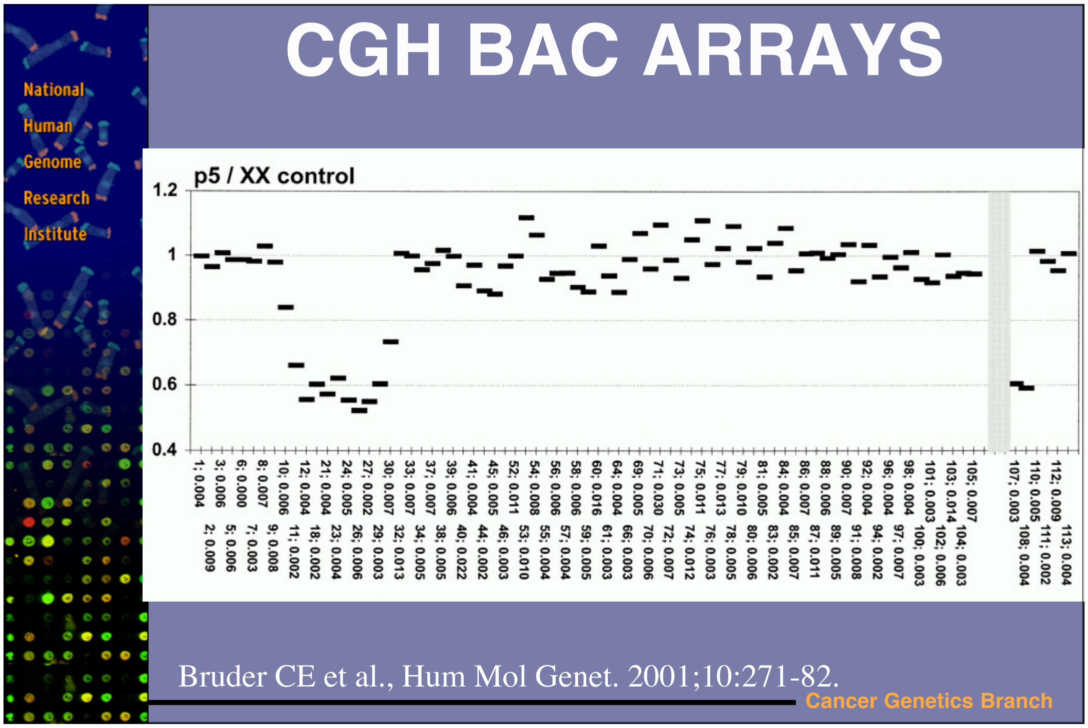
Cancer Genetics Branch



CGH BAC ARRAYS

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

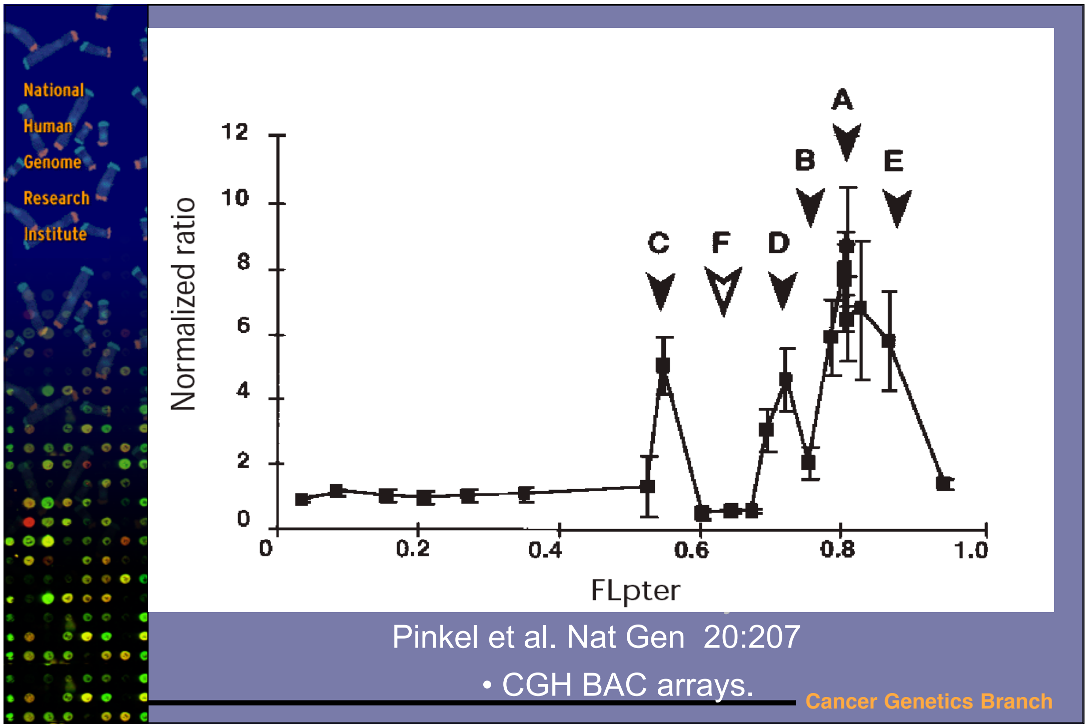
Cancer Genetics Branch



CGH BAC ARRAYS

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

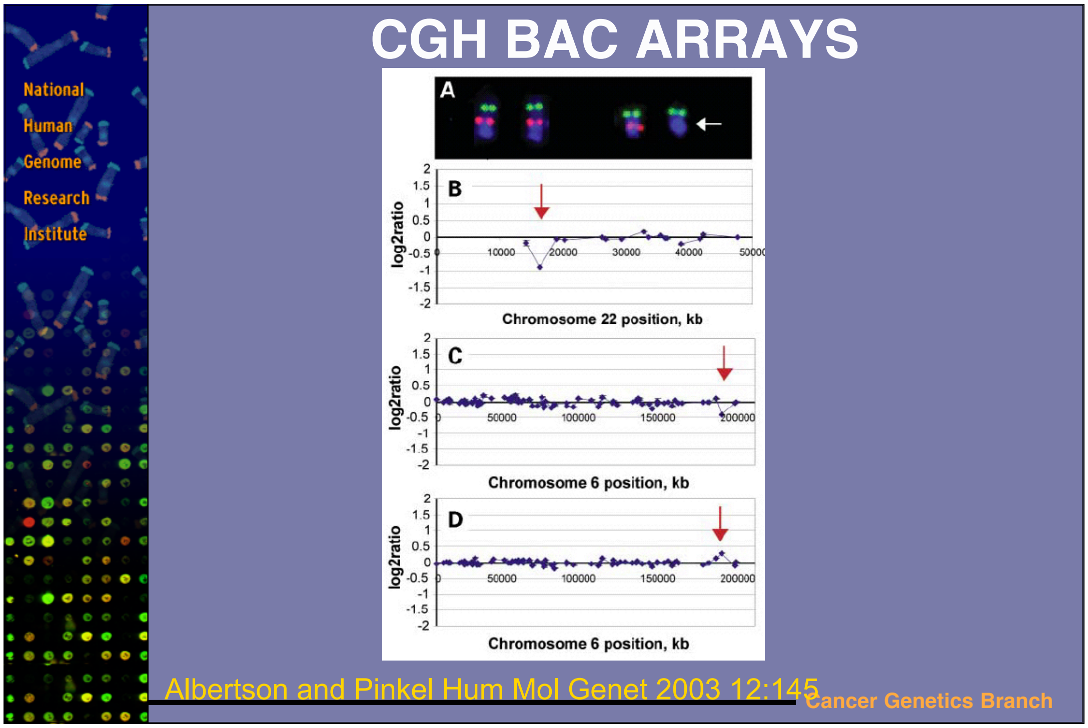
Cancer Genetics Branch



Pinkel et al. Nat Gen 20:207

- CGH BAC arrays.

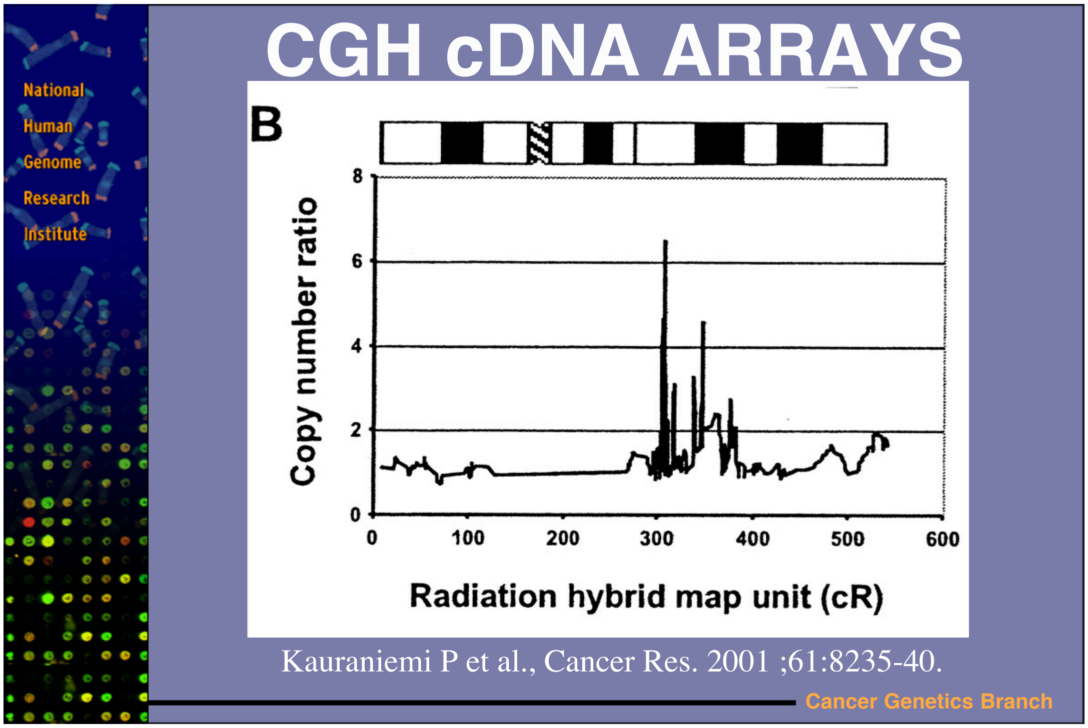
Cancer Genetics Branch



CGH BAC ARRAYS

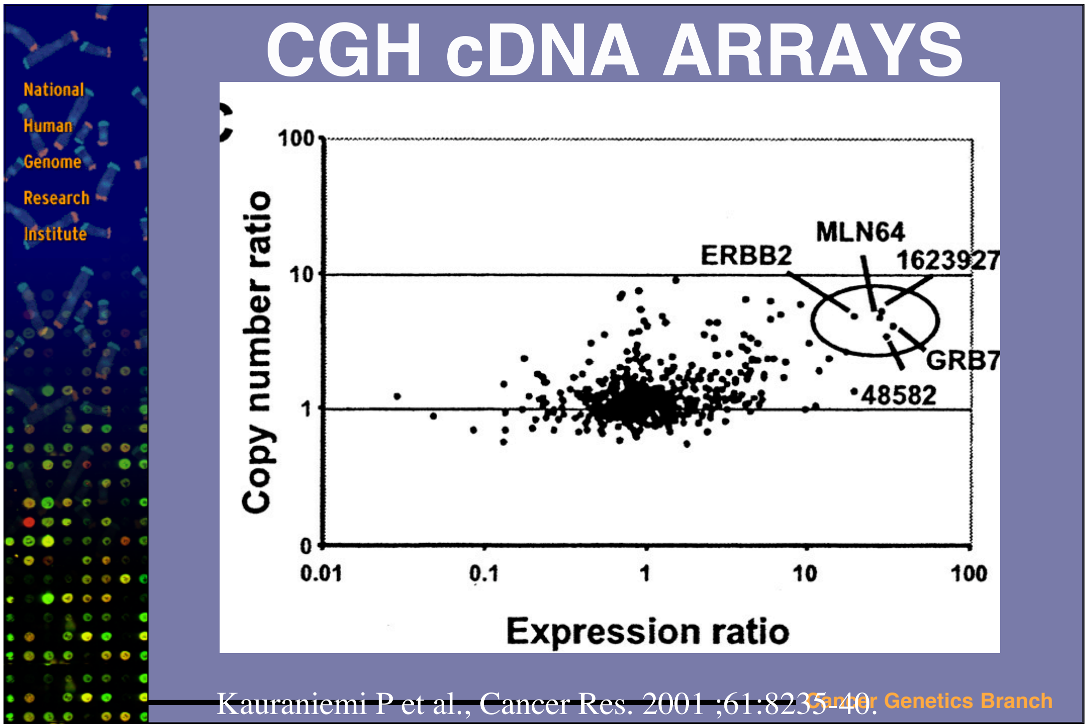
Albertson and Pinkel Hum Mol Genet 2003 12:145

Cancer Genetics Branch



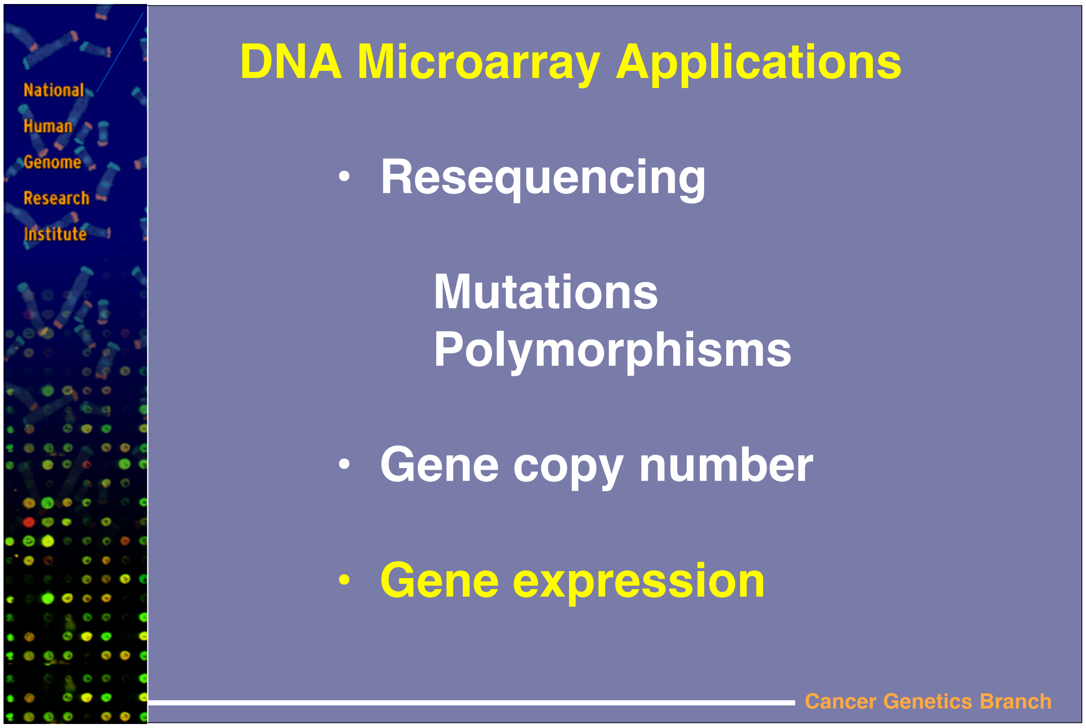
CGH cDNA ARRAYS

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



CGH cDNA ARRAYS

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



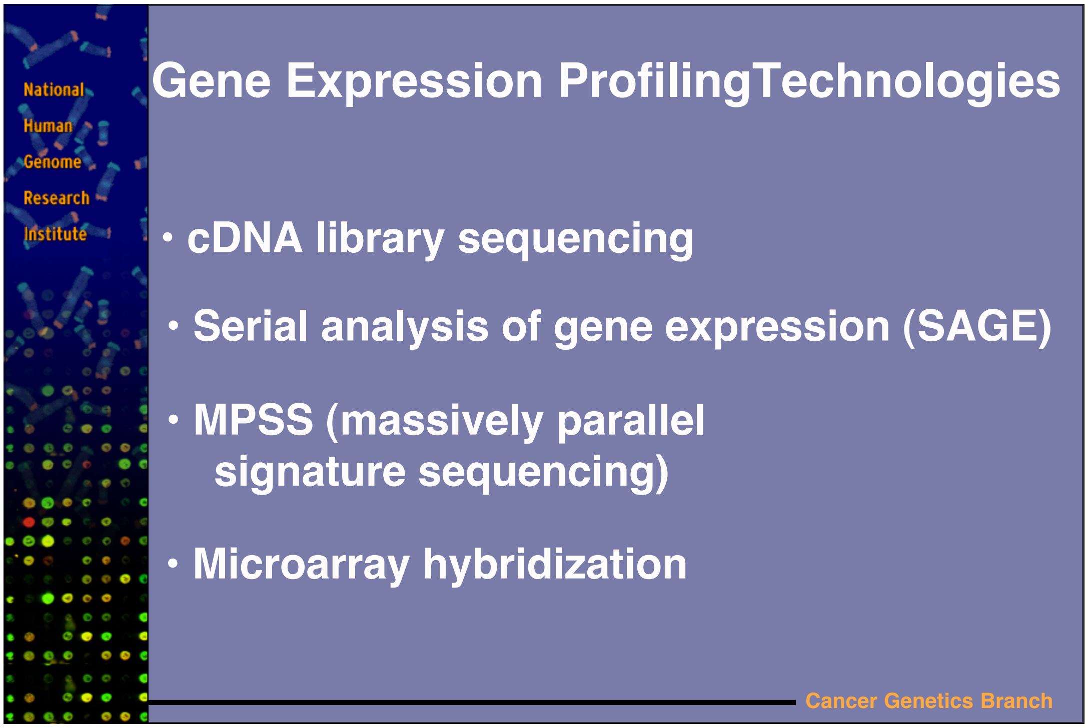
DNA Microarray Applications

- Resequencing

  Mutations
  Polymorphisms

- Gene copy number
- Gene expression

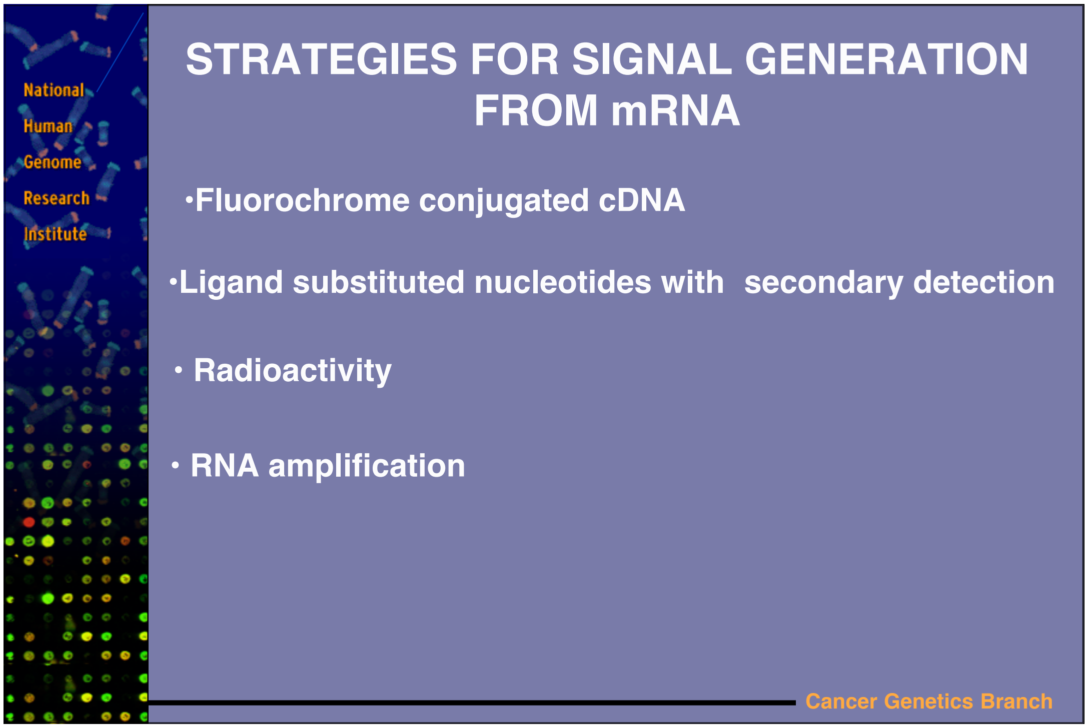
Cancer Genetics Branch



Gene Expression ProfilingTechnologies

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

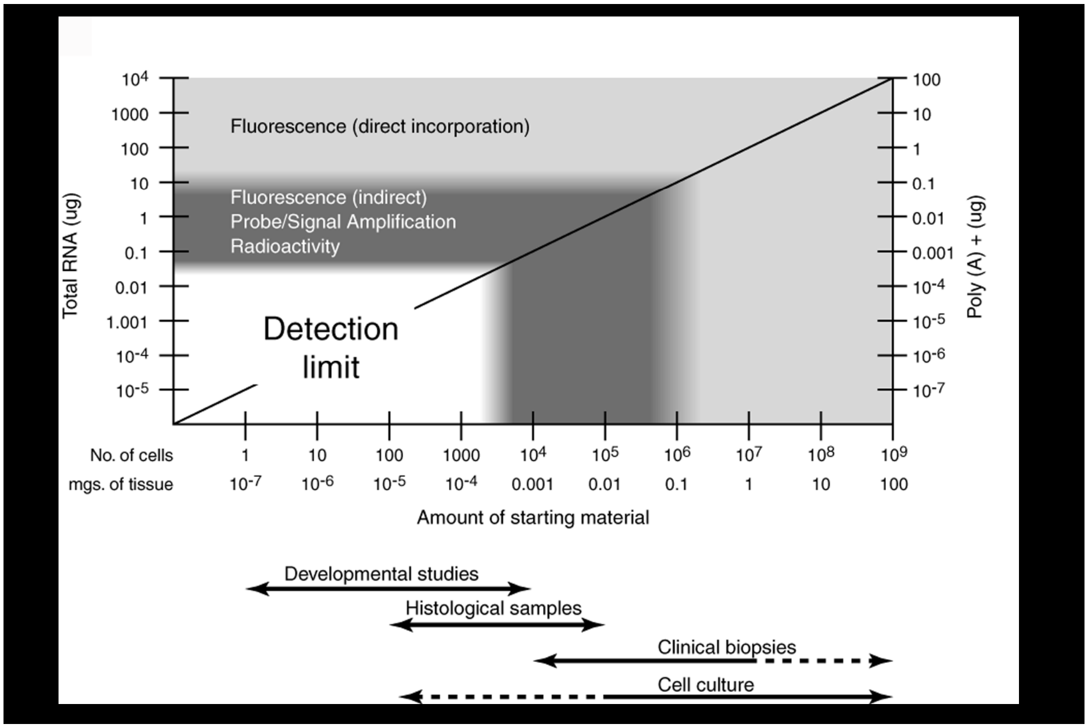
Cancer Genetics Branch

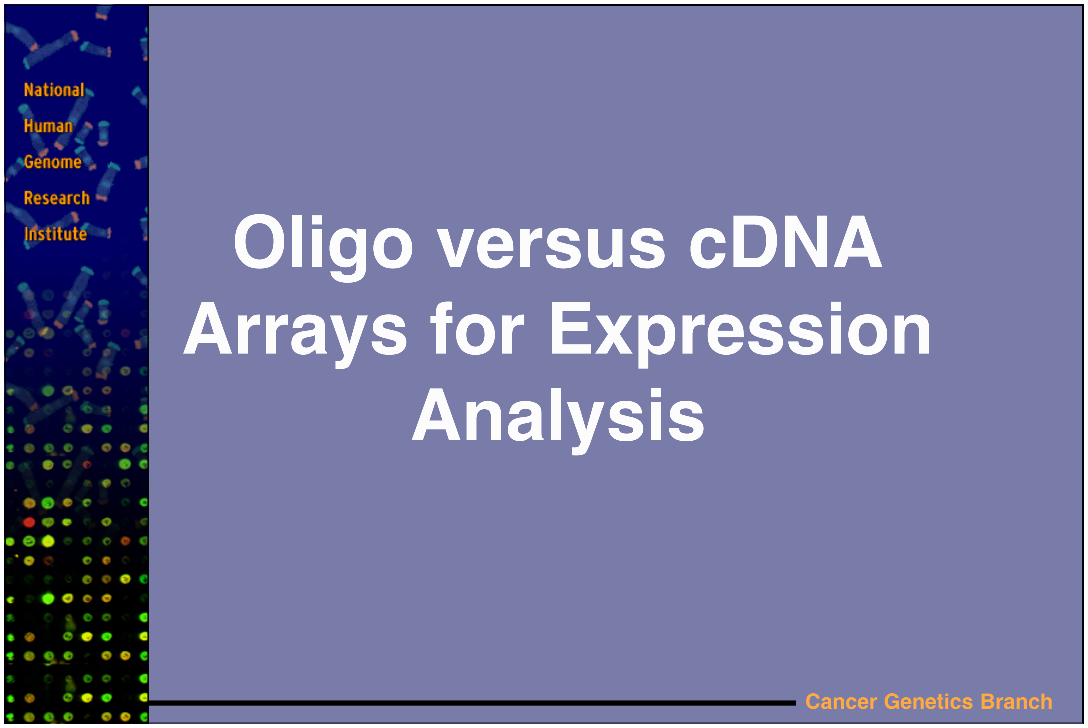


STRATEGIES FOR SIGNAL GENERATION FROM mRNA

- Fluorochrome conjugated cDNA
- Ligand substituted nucleotides with secondary detection
- Radioactivity
- RNA amplification

Cancer Genetics Branch

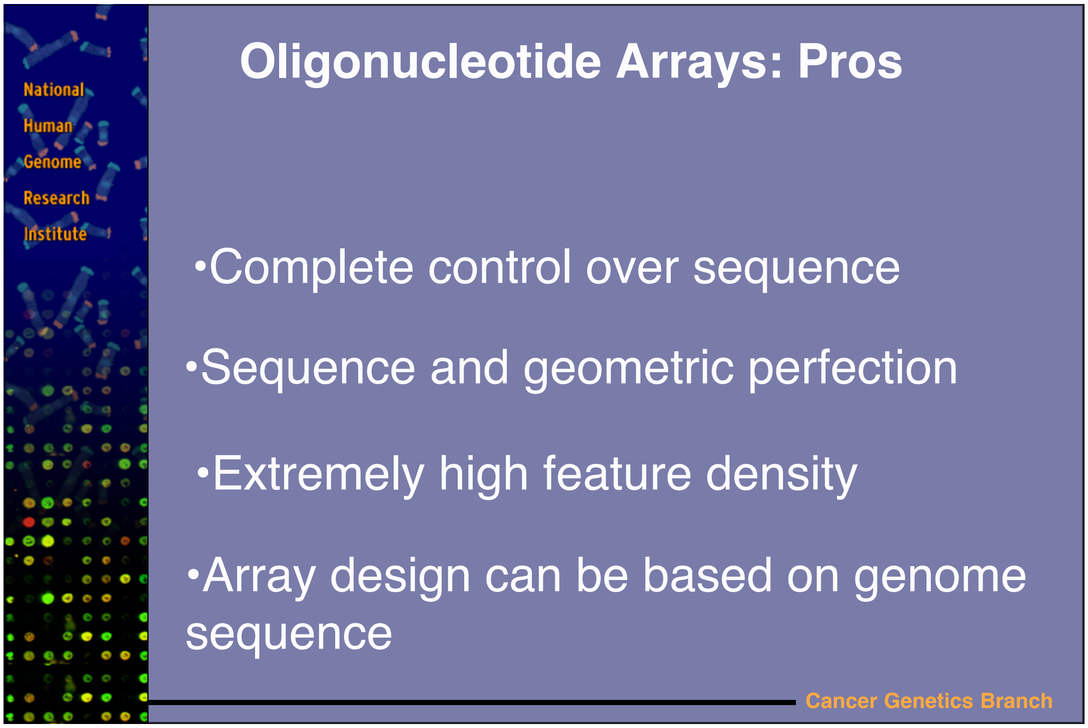




National
Human
Genome
Research
Institute

Oligo versus cDNA Arrays for Expression Analysis

Cancer Genetics Branch

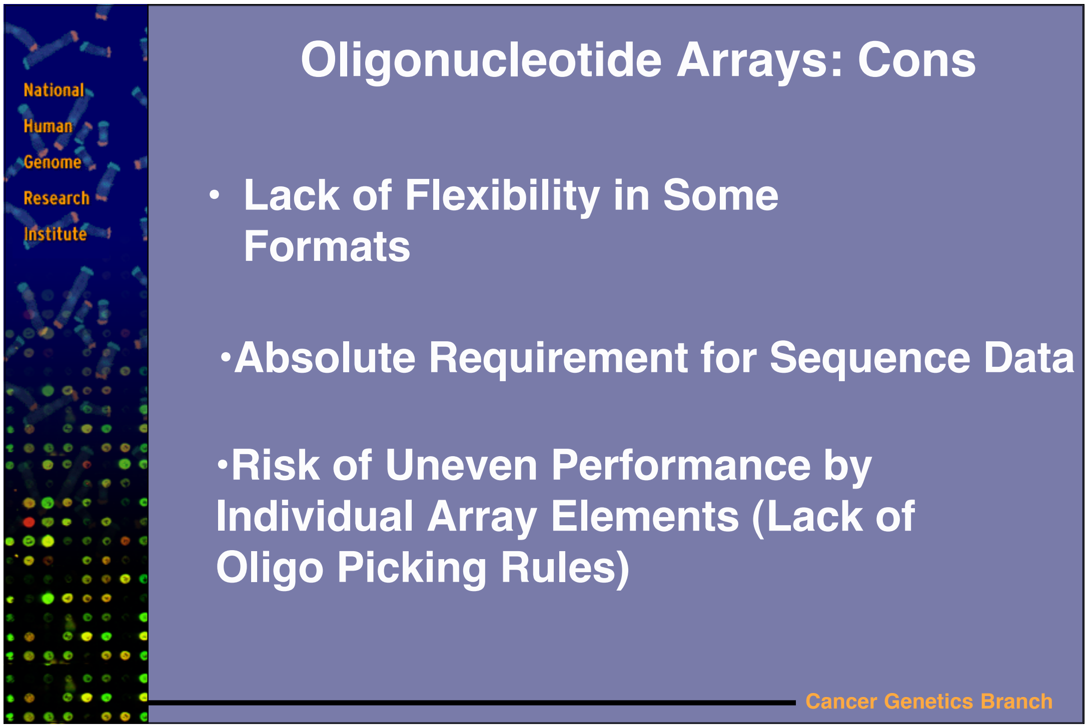


National
Human
Genome
Research
Institute

Oligonucleotide Arrays: Pros

- Complete control over sequence
- Sequence and geometric perfection
- Extremely high feature density
- Array design can be based on genome sequence

Cancer Genetics Branch

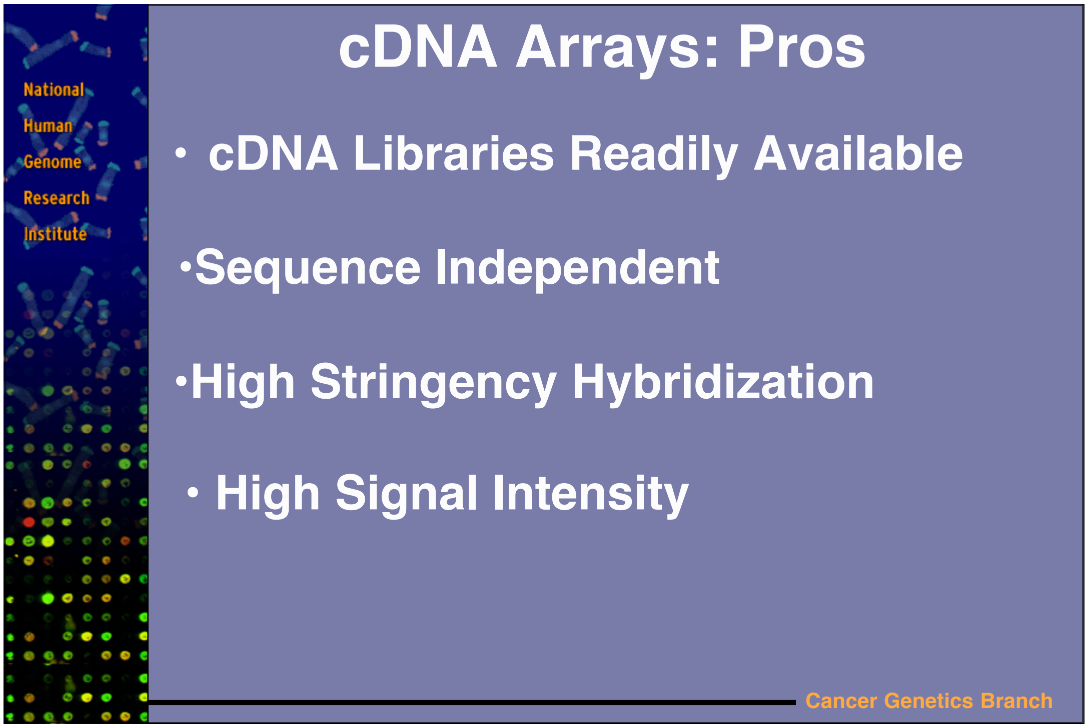


National
Human
Genome
Research
Institute

Oligonucleotide Arrays: Cons

- Lack of Flexibility in Some Formats
- Absolute Requirement for Sequence Data
- Risk of Uneven Performance by Individual Array Elements (Lack of Oligo Picking Rules)

Cancer Genetics Branch

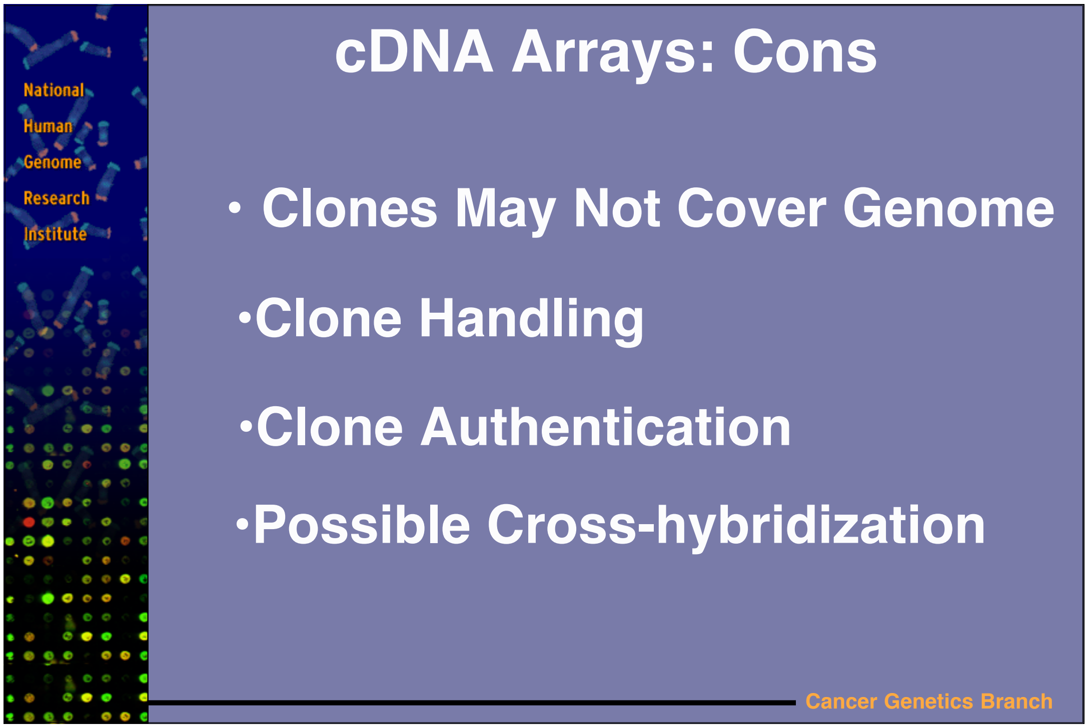


National
Human
Genome
Research
Institute

cDNA Arrays: Pros

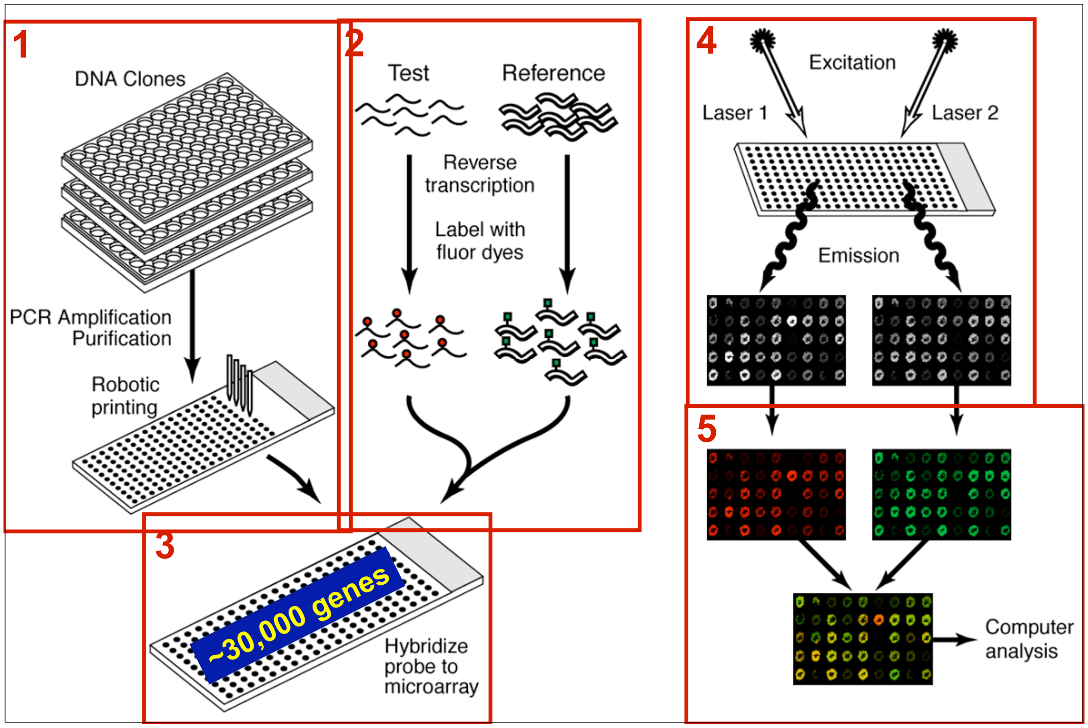
- cDNA Libraries Readily Available
- Sequence Independent
- High Stringency Hybridization
- High Signal Intensity

Cancer Genetics Branch



cDNA Arrays: Cons

- Clones May Not Cover Genome
- Clone Handling
- Clone Authentication
- Possible Cross-hybridization



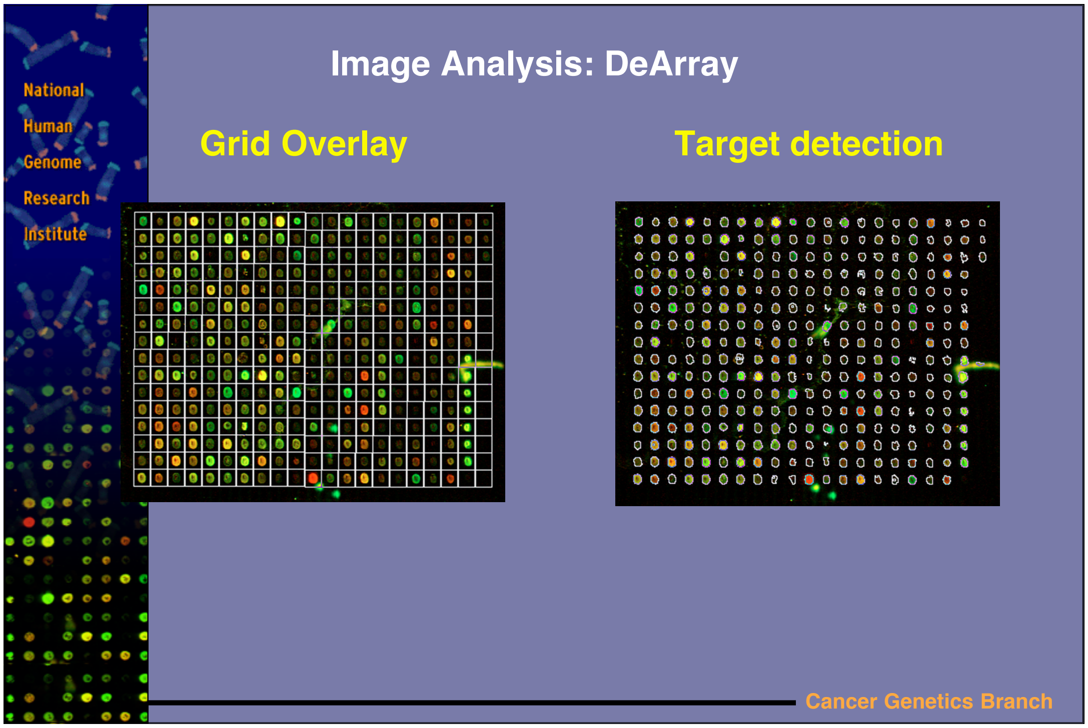


Image Analysis: DeArray

Grid Overlay

Target detection

Cancer Genetics Branch

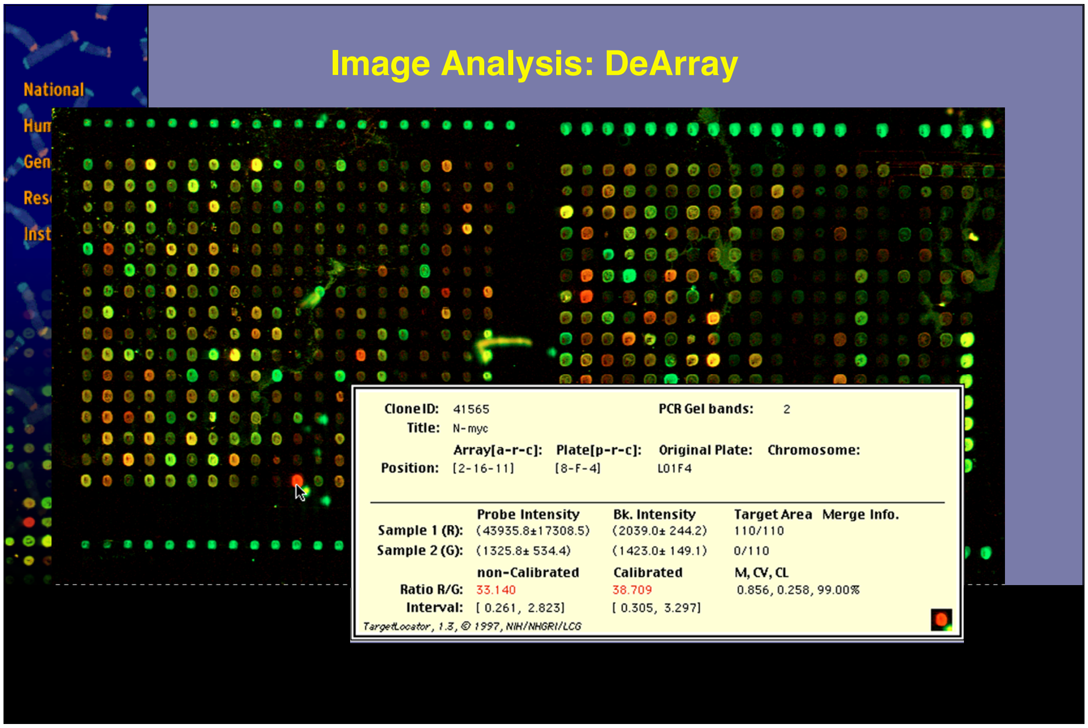


Image Analysis: DeArray

Clone ID: 41565 **PCR Gel bands:** 2

Title: N-myc

	Array[a-r-c]:	Plate[p-r-c]:	Original Plate:	Chromosome:
Position:	[2-16-11]	[8-F-4]	L01F4	

	Probe Intensity	Bk. Intensity	Target Area	Merge Info.
Sample 1 (R):	(43935.8±17308.5)	(2039.0± 244.2)	110/110	
Sample 2 (G):	(1325.8± 534.4)	(1423.0± 149.1)	0/110	

	non-Calibrated	Calibrated	M, CV, CL
Ratio R/G:	33.140	38.709	0.856, 0.258, 99.00%
Interval:	[0.261, 2.823]	[0.305, 3.297]	

TargetLocator, 1.3, © 1997, NIH/NHGRI/LCG

National
Human
Genome
Research
Institute

DATA QUALITY IS CRITICAL

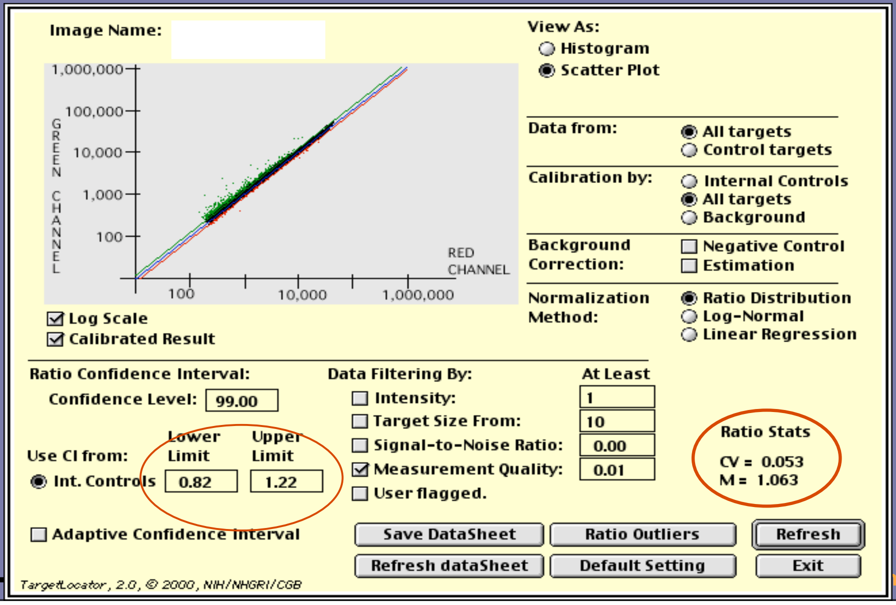


Image Name:

☑ Log Scale
☑ Calibrated Result

View As:
○ Histogram
● Scatter Plot

Data from: ● All targets ○ Control targets

Calibration by: ○ Internal Controls ● All targets ○ Background

Background Correction: ☐ Negative Control ☐ Estimation

Normalization Method: ● Ratio Distribution ○ Log-Normal ○ Linear Regression

Ratio Confidence Interval:
Confidence Level: 99.00

Use CI from:

	Lower Limit	Upper Limit
● Int. Controls	0.82	1.22

☐ Adaptive Confidence Interval

Data Filtering By:	At Least
☐ Intensity:	1
☐ Target Size From:	10
☐ Signal-to-Noise Ratio:	0.00
☑ Measurement Quality:	0.01
☐ User flagged.	

Ratio Stats
CV = 0.053
M = 1.063

TargetLocator, 2.0, © 2000, NIH/NHGRI/CGB

National
Human
Genome
Research
Institute

Output of cDNA microarrays: expression ratio

Output of oligonucleotide arrays: expression level or ratio

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

National
Human
Genome
Research
Institute

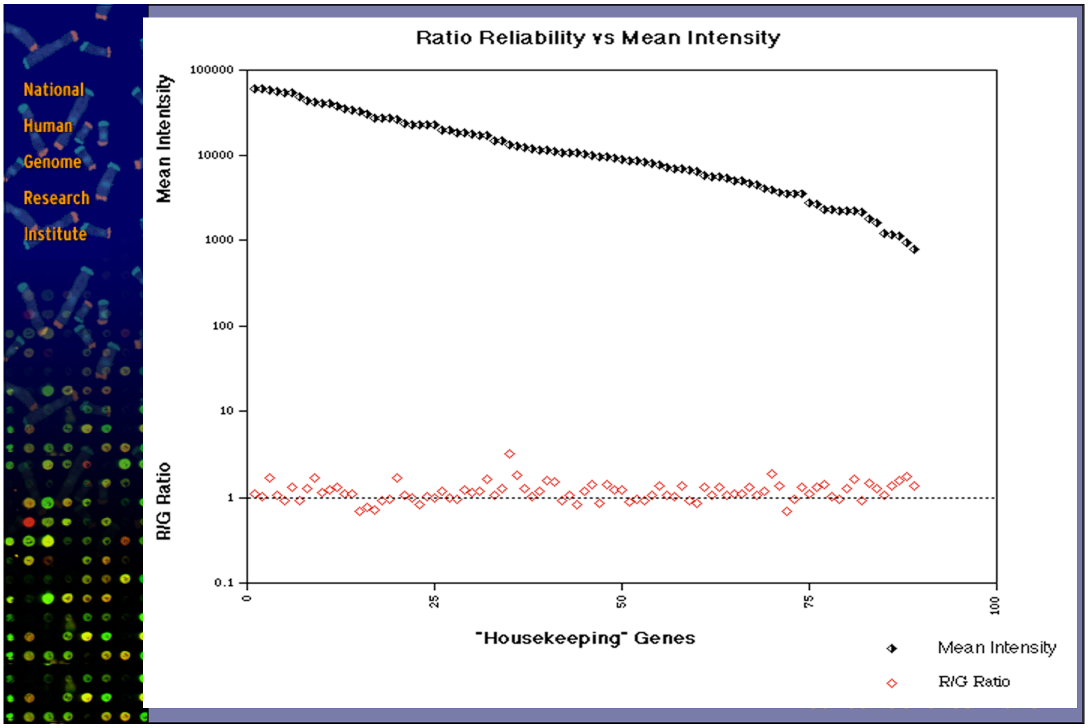
Normalization of cDNA microarrays:

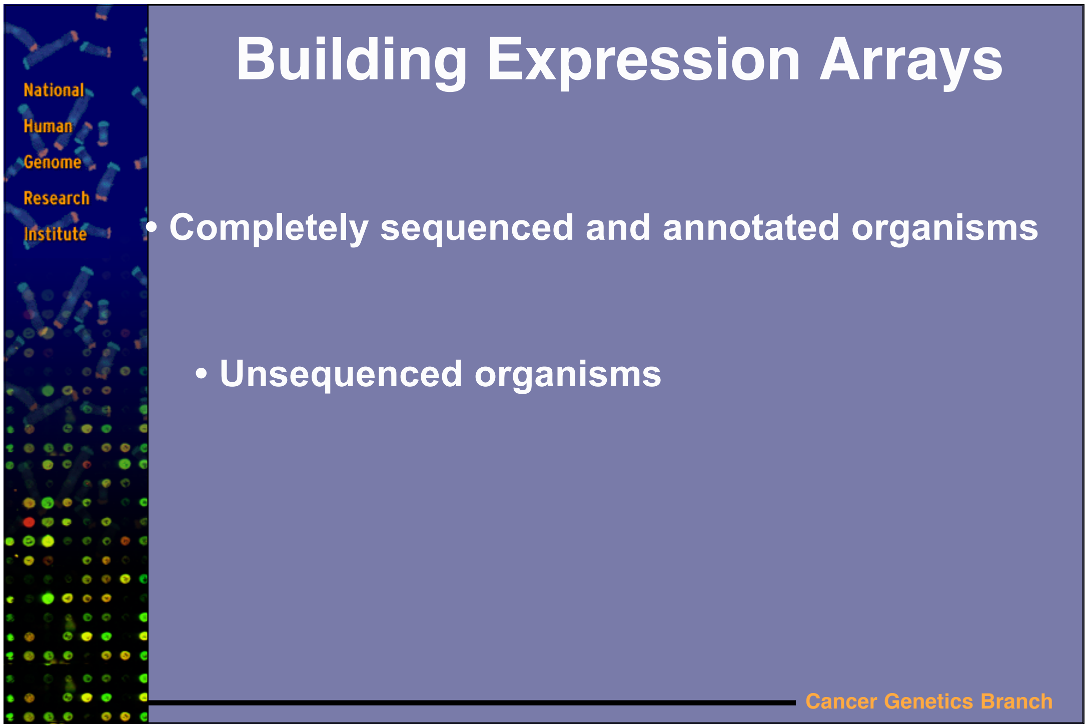
- global
- housekeeping
- spiked standards

Normalization of oligonucleotide arrays:

- global

Cancer Genetics Branch



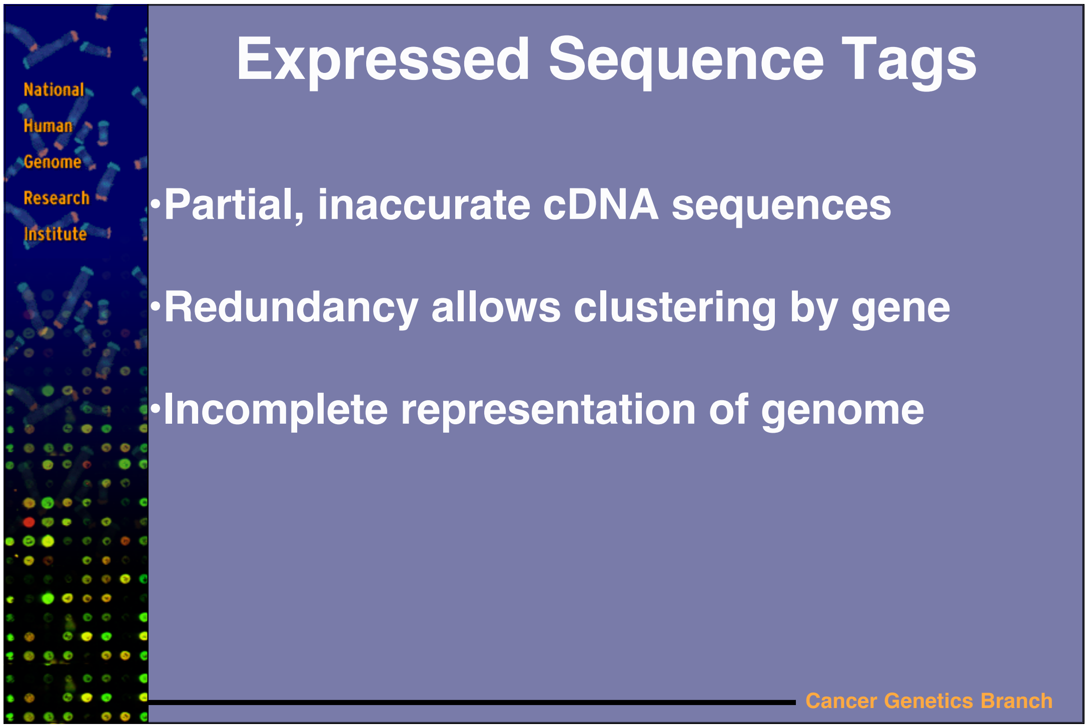


National
Human
Genome
Research
Institute

Building Expression Arrays

- Completely sequenced and annotated organisms
  - Unsequenced organisms

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Expressed Sequence Tags

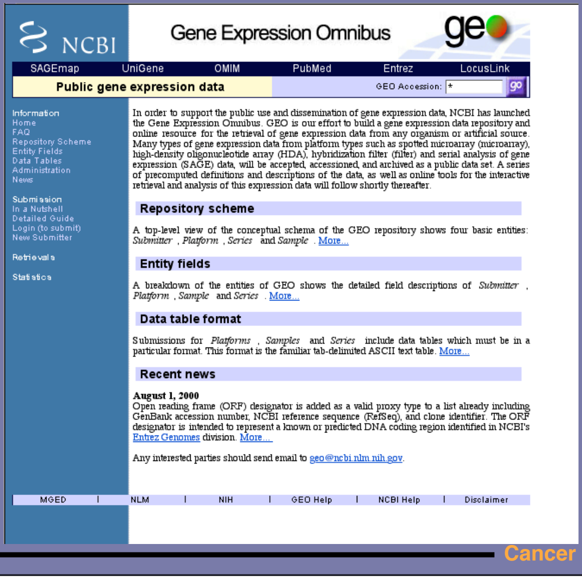
- Partial, inaccurate cDNA sequences
- Redundancy allows clustering by gene
- Incomplete representation of genome

Cancer Genetics Branch

National
Human
Genome
Research
Institute

Accessing Expression Data

- Individual Lab and Journal Sites



In order to support the public use and dissemination of gene expression data, NCBI has launched the Gene Expression Omnibus. GEO is our effort to build a gene expression data repository and online resource for the retrieval of gene expression data from any organism or artificial source. Many types of gene expression data from platform types such as spotted microarray (microarray), high-density oligonucleotide array (HDA), hybridization filter (filter) and serial analysis of gene expression (SAGE) data, will be accepted, accessioned, and archived as a public data set. A series of precomputed definitions and descriptions of the data, as well as online tools for the interactive retrieval and analysis of this expression data will follow shortly thereafter.

Repository scheme

A top-level view of the conceptual schema of the GEO repository shows four basic entities: *Submitter* , *Platform* , *Series* and *Sample* . More...

Entity fields

A breakdown of the entities of GEO shows the detailed field descriptions of *Submitter* , *Platform* , *Sample* and *Series* . More...

Data table format

Submissions for *Platforms* , *Samples* and *Series* include data tables which must be in a particular format. This format is the familiar tab-delimited ASCII text table. More...

Recent news

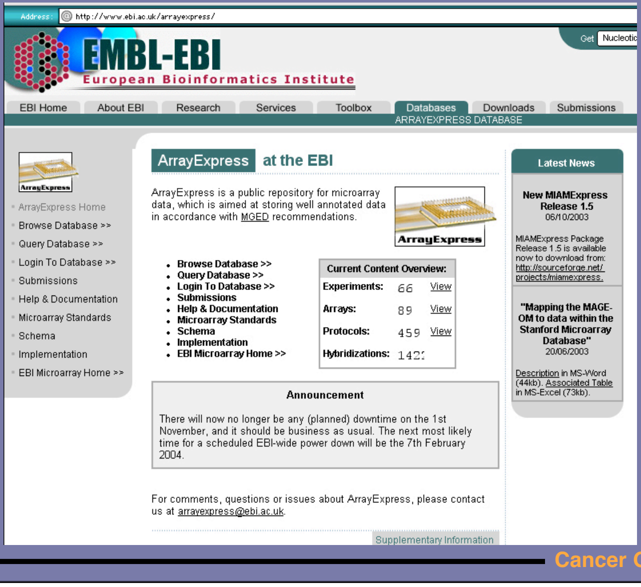
August 1, 2000
Open reading frame (ORF) designator is added as a valid proxy type to a list already including GenBank accession number, NCBI reference sequence (RefSeq), and clone identifier. The ORF designator is intended to represent a known or predicted DNA coding region identified in NCBI's Entrez Genomes division. More...

Any interested parties should send email to geo@ncbi.nlm.nih.gov.

Cancer Genetics Branch

National
Human
Genome
Research
Institute

Accessing Expression Data



ArrayExpress at the EBI

ArrayExpress is a public repository for microarray data, which is aimed at storing well annotated data in accordance with MGED recommendations.

- **Browse Database >>**
- **Query Database >>**
- **Login To Database >>**
- **Submissions**
- **Help & Documentation**
- **Microarray Standards**
- **Schema**
- **Implementation**
- **EBI Microarray Home >>**

Current Content Overview:		
Experiments:	66	View
Arrays:	89	View
Protocols:	459	View
Hybridizations:	142	

Announcement

There will now no longer be any (planned) downtime on the 1st November, and it should be business as usual. The next most likely time for a scheduled EBI-wide power down will be the 7th February 2004.

For comments, questions or issues about ArrayExpress, please contact us at arrayexpress@ebi.ac.uk.

Latest News

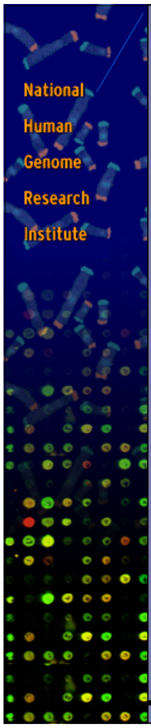
New MIAMExpress Release 1.5
06/10/2003

MIAMExpress Package Release 1.5 is available now to download from: http://sourceforge.net/projects/miamexpress.

"Mapping the MAGE-OM to data within the Stanford Microarray Database"
20/06/2003

Description in MS-Word (44kb). Associated Table in MS-Excel (73kb).

Cancer Genetics Branch



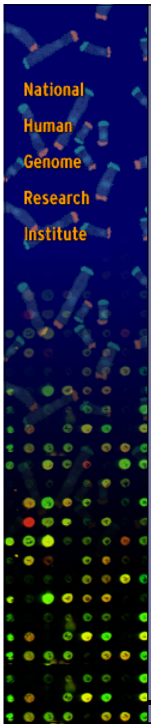
Publishing Expression Data

- MIAME standard

Minimum Information about a Microarray Experiment

Format required by some journals

Cancer Genetics Branch



APPLICATIONS OF EXPRESSION ARRAYS

- **Expression profiling**

  Requires statistical tools

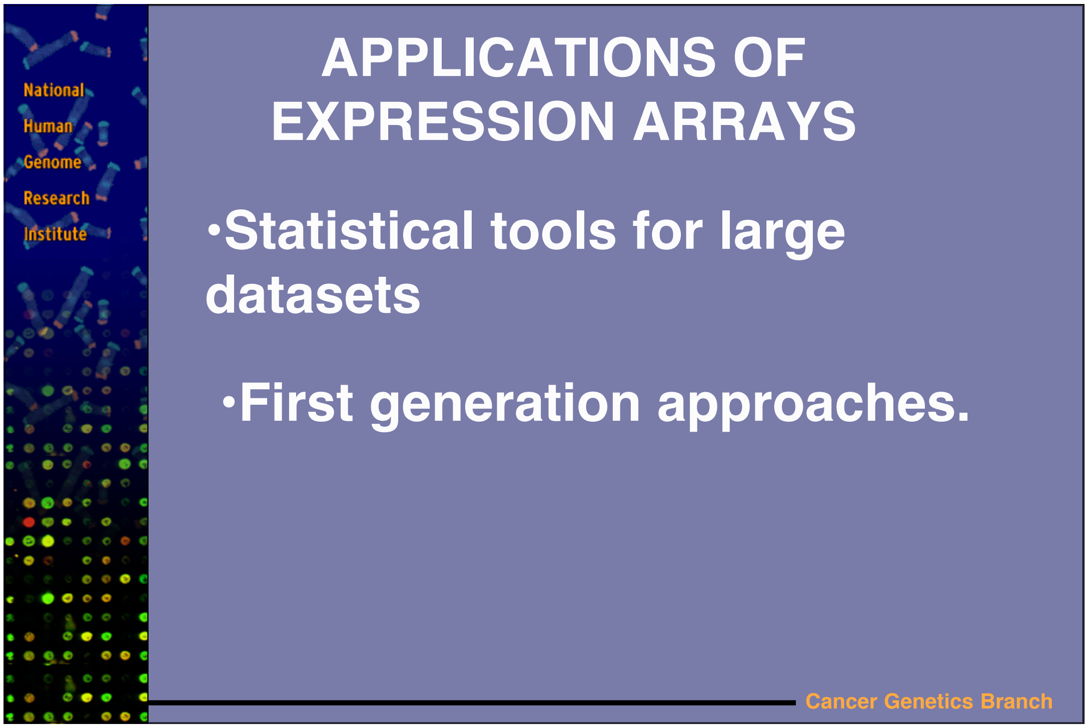
  Power arises from increasing sample number

- **Direct comparisons (Induction)**

  Biological system critical

- **Genome Annotation**

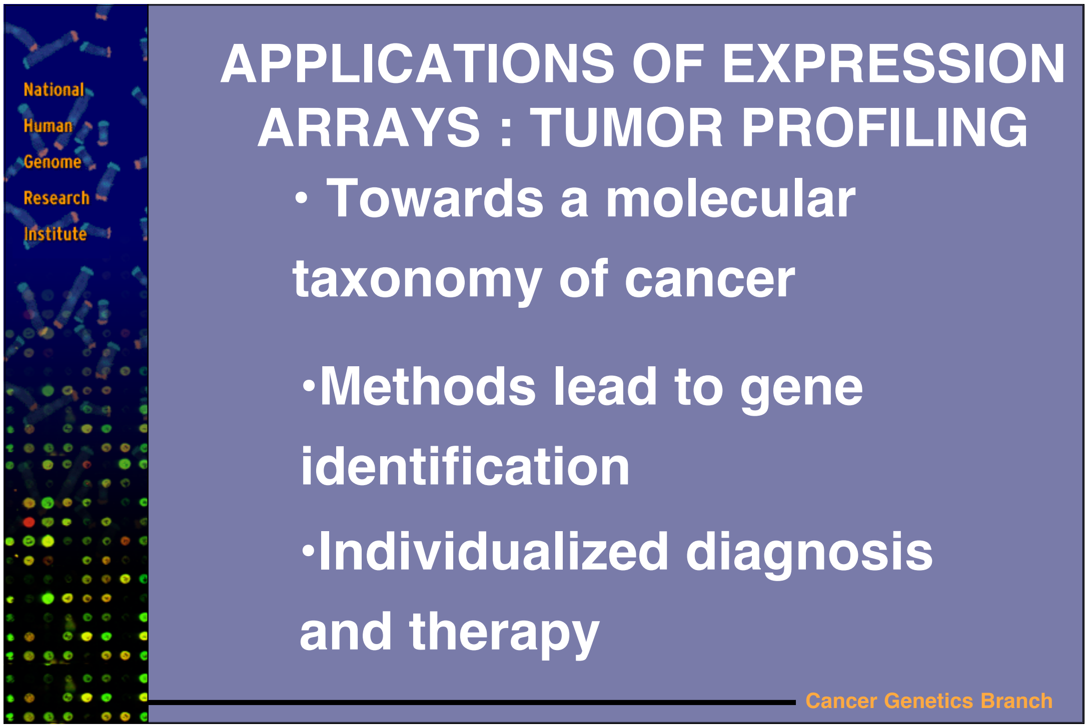
Cancer Genetics Branch



APPLICATIONS OF EXPRESSION ARRAYS

- Statistical tools for large datasets
- First generation approaches.

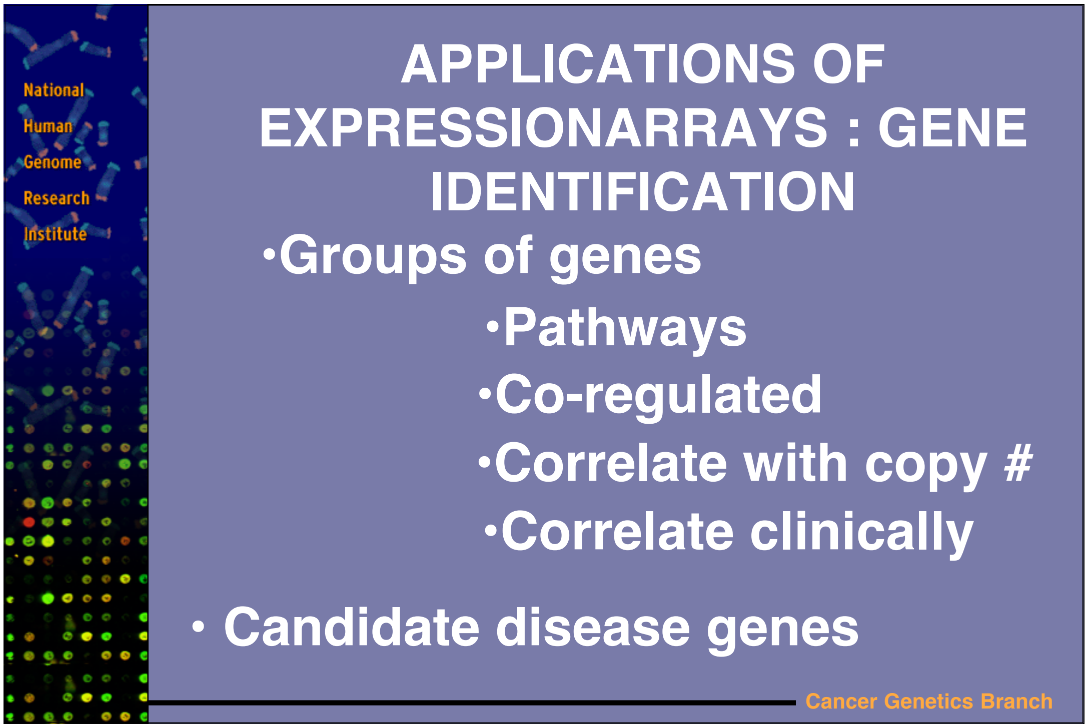
Cancer Genetics Branch



APPLICATIONS OF EXPRESSION ARRAYS : TUMOR PROFILING

- Towards a molecular taxonomy of cancer
- Methods lead to gene identification
- Individualized diagnosis and therapy

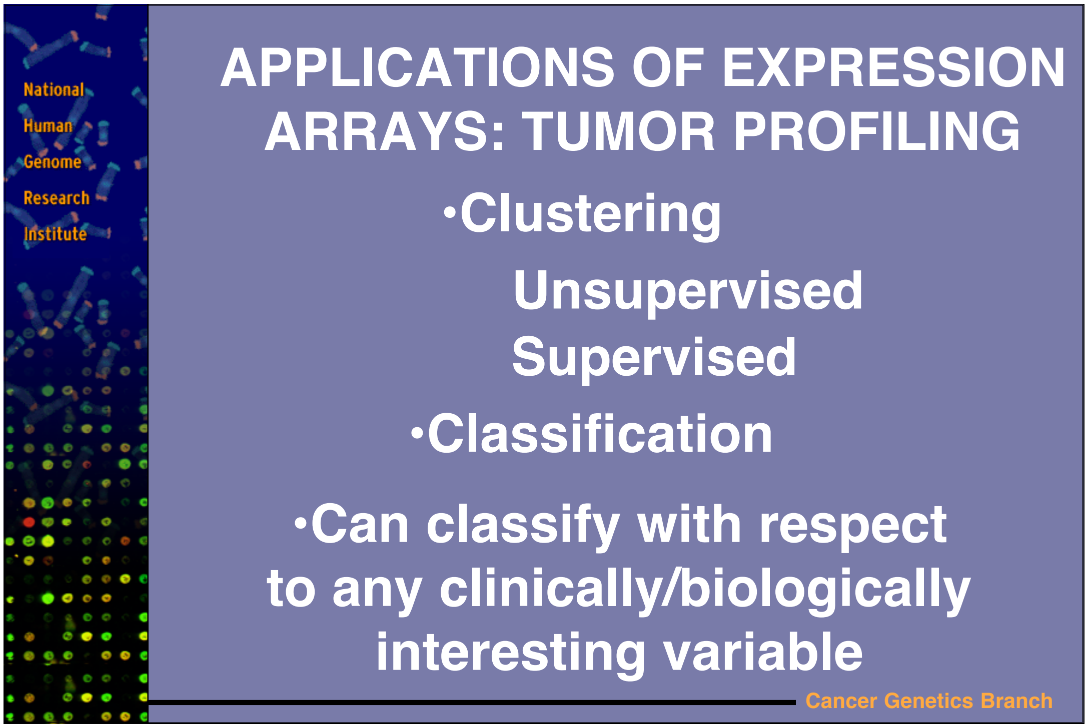
Cancer Genetics Branch



APPLICATIONS OF EXPRESSIONARRAYS : GENE IDENTIFICATION

- Groups of genes
  - Pathways
  - Co-regulated
  - Correlate with copy #
  - Correlate clinically
- Candidate disease genes

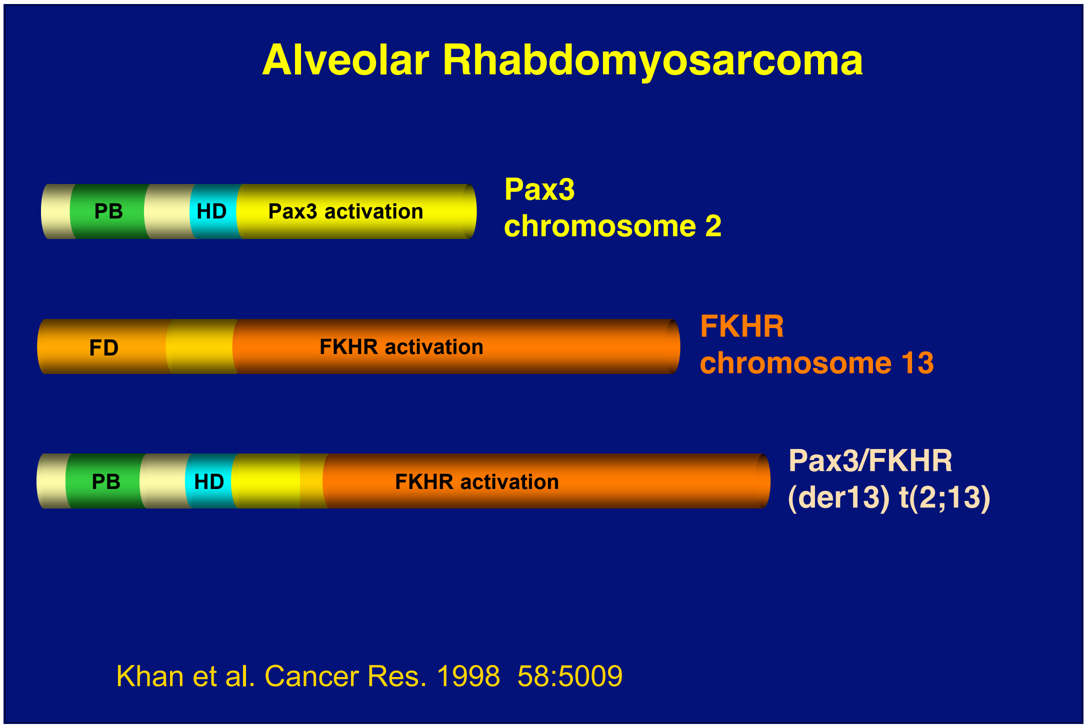
Cancer Genetics Branch



APPLICATIONS OF EXPRESSION ARRAYS: TUMOR PROFILING

- Clustering
  - Unsupervised
  - Supervised
- Classification
- Can classify with respect to any clinically/biologically interesting variable

Cancer Genetics Branch



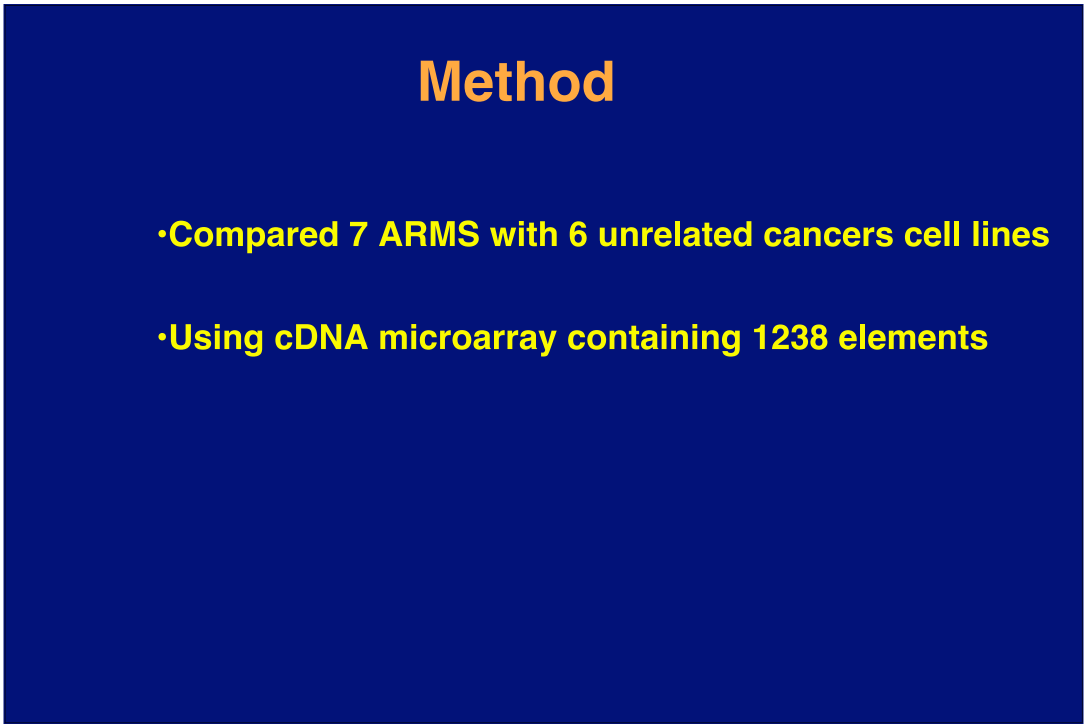
Alveolar Rhabdomyosarcoma

PB HD Pax3 activation
Pax3
chromosome 2

FD FKHR activation
FKHR
chromosome 13

PB HD FKHR activation
Pax3/FKHR
(der13) t(2;13)

Khan et al. Cancer Res. 1998 58:5009



Method

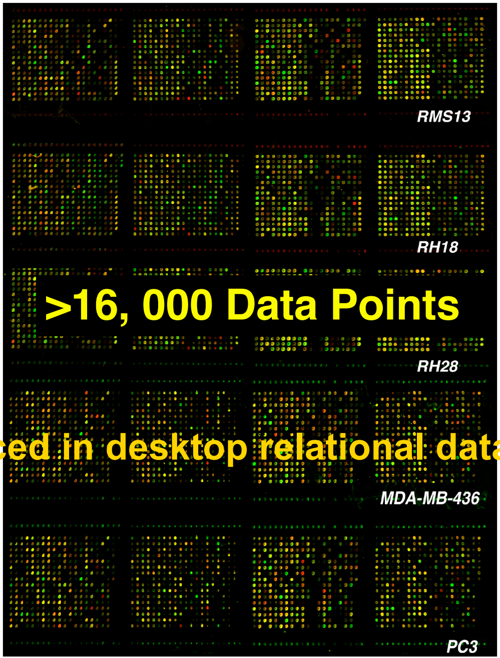
- **Compared 7 ARMS with 6 unrelated cancers cell lines**
- **Using cDNA microarray containing 1238 elements**

Cell Line Characteristics

Cell Line	Pax3-FKHR	Diagnosis
ARMS1	+	ARMS
RH3	+	ARMS
RH4	+	ARMS
RH5	+	ARMS
RMS18	+	ARMS
RMS13	+	ARMS
RH28	+	ARMS
A204	-	Undifferentiated Sarcoma
NGP127	-	Neuroblastoma
TC71	-	Ewing’s Sarcoma
UACC-903	-	Melanoma
PC3	-	Prostate Carcinoma
MDA-MB-436	-	Breast Carcinoma
Control		
NIL-C	-	Fibroblast

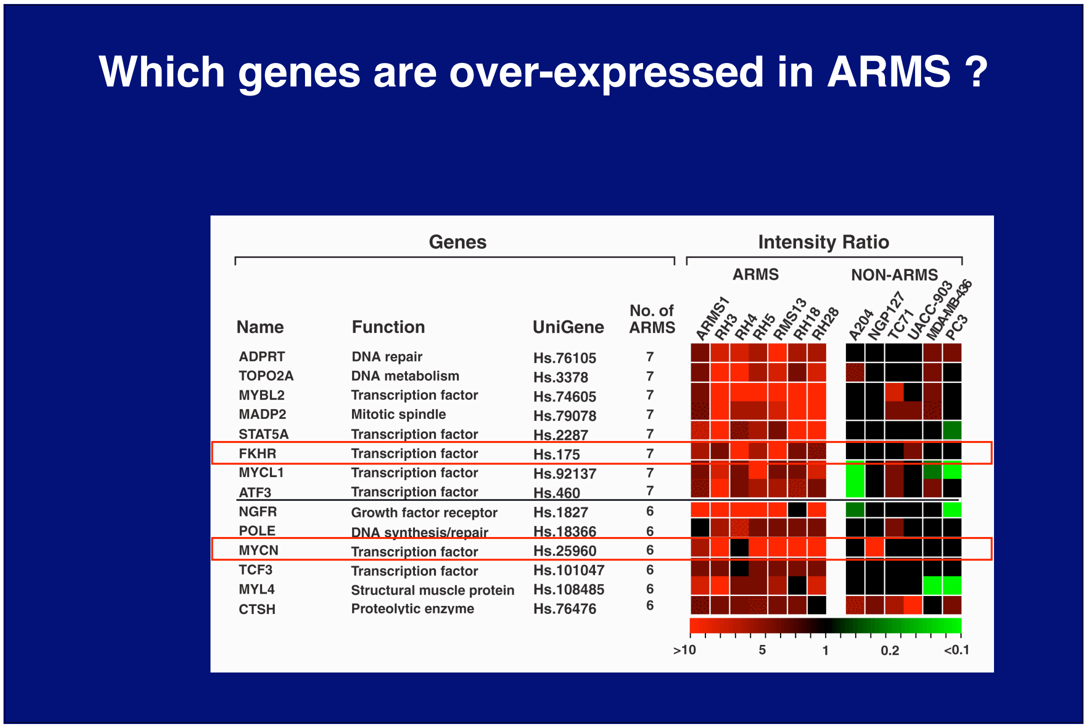
Results

- 13 Experiments
- 1238 genes



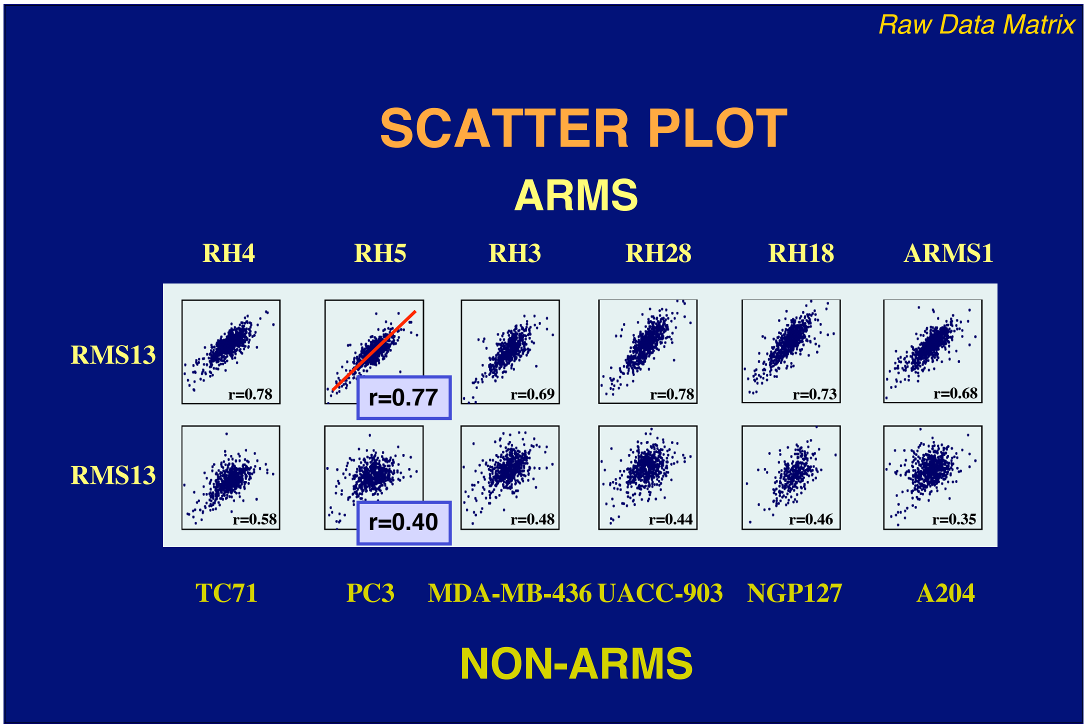
>16, 000 Data Points

Data placed in desktop relational database



Which genes are over-expressed in ARMS ?

Name	Function	UniGene	No. of ARMS
ADPRT	DNA repair	Hs.76105	7
TOPO2A	DNA metabolism	Hs.3378	7
MYBL2	Transcription factor	Hs.74605	7
MADP2	Mitotic spindle	Hs.79078	7
STAT5A	Transcription factor	Hs.2287	7
FKHR	Transcription factor	Hs.175	7
MYCL1	Transcription factor	Hs.92137	7
ATF3	Transcription factor	Hs.460	7
NGFR	Growth factor receptor	Hs.1827	6
POLE	DNA synthesis/repair	Hs.18366	6
MYCN	Transcription factor	Hs.25960	6
TCF3	Transcription factor	Hs.101047	6
MYL4	Structural muscle protein	Hs.108485	6
CTSH	Proteolytic enzyme	Hs.76476	6



Raw Data Matrix

SCATTER PLOT

ARMS

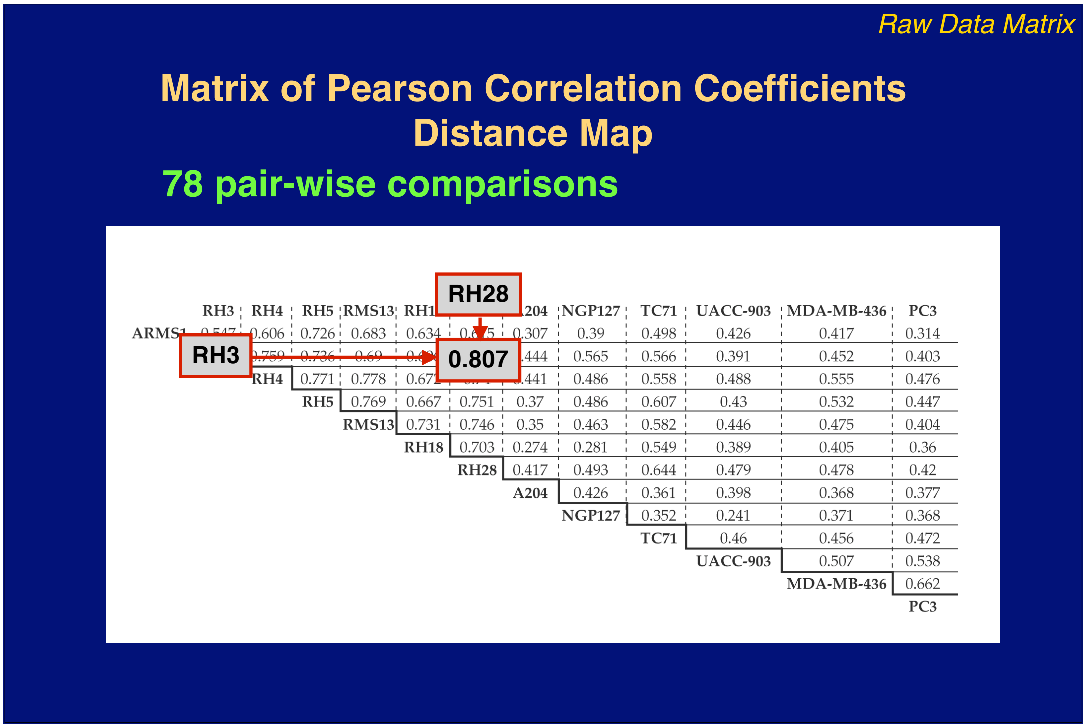
RH4 RH5 RH3 RH28 RH18 ARMS1

RMS13

RMS13

TC71 PC3 MDA-MB-436 UACC-903 NGP127 A204

NON-ARMS

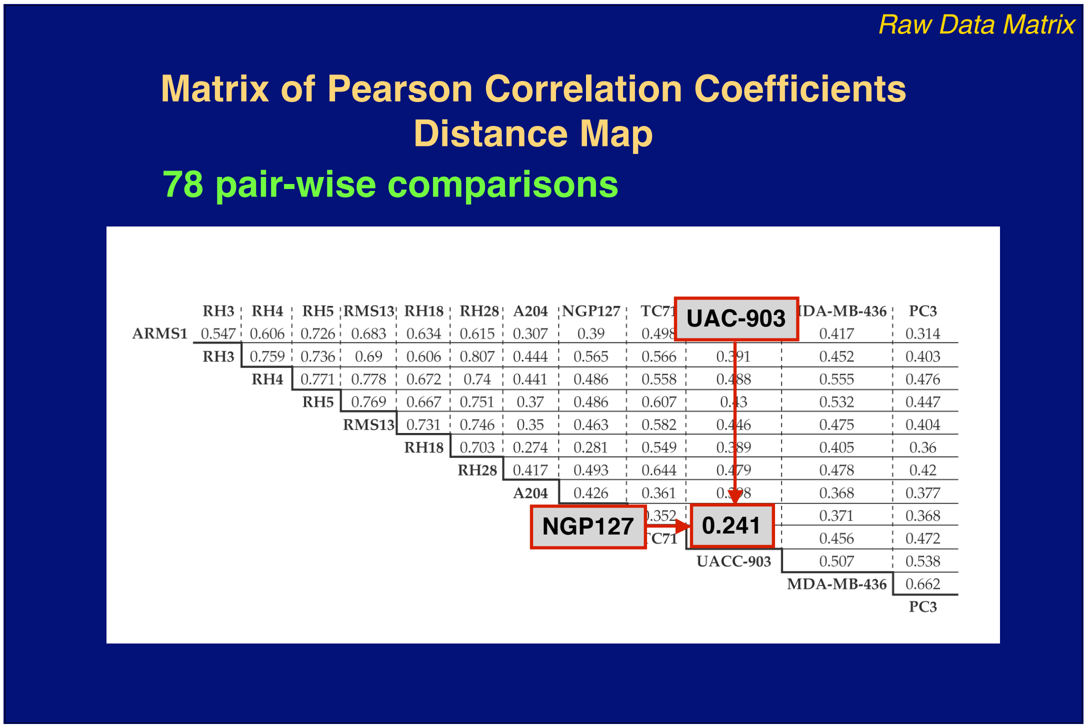


Raw Data Matrix

Matrix of Pearson Correlation Coefficients
Distance Map

78 pair-wise comparisons

	RH3	RH4	RH5	RMS13	RH1	RH28	204	NGP127	TC71	UACC-903	MDA-MB-436	PC3
ARMS1	0.547	0.606	0.726	0.683	0.634	[illegible]	0.307	0.39	0.498	0.426	0.417	0.314
RH3		0.759	0.736	0.69	[illegible]	0.807	[illegible]	0.565	0.566	0.391	0.452	0.403
		RH4	0.771	0.778	0.672	[illegible]	0.441	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

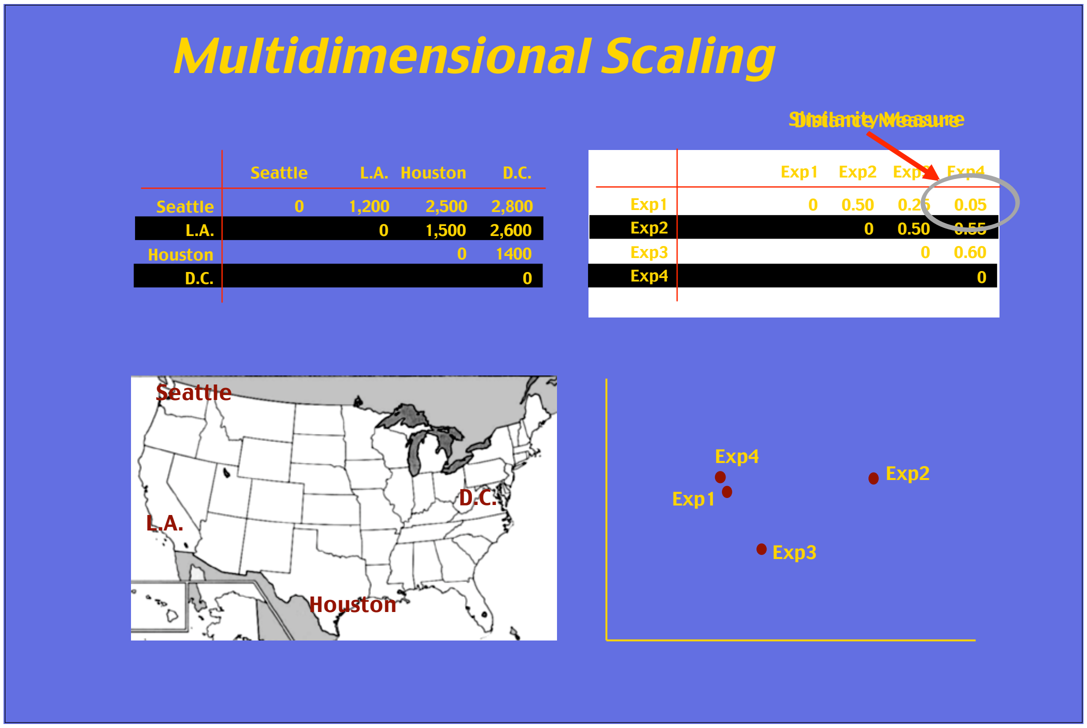


Raw Data Matrix

Matrix of Pearson Correlation Coefficients
Distance Map

78 pair-wise comparisons

	RH3	RH4	RH5	RMS13	RH18	RH28	A204	NGP127	TC7	UAC-903	DA-MB-436	PC3
ARMS1	0.547	0.606	0.726	0.683	0.634	0.615	0.307	0.39	0.498	[illegible]	0.417	0.314
RH3		0.759	0.736	0.69	0.606	0.807	0.444	0.565	0.566	0.391	0.452	0.403
		RH4	0.771	0.778	0.672	0.74	0.441	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	.352	0.241	0.371	0.368
									C71	[illegible]	0.456	0.472
										UACC-903	0.507	0.538
											MDA-MB-436	0.662
												PC3

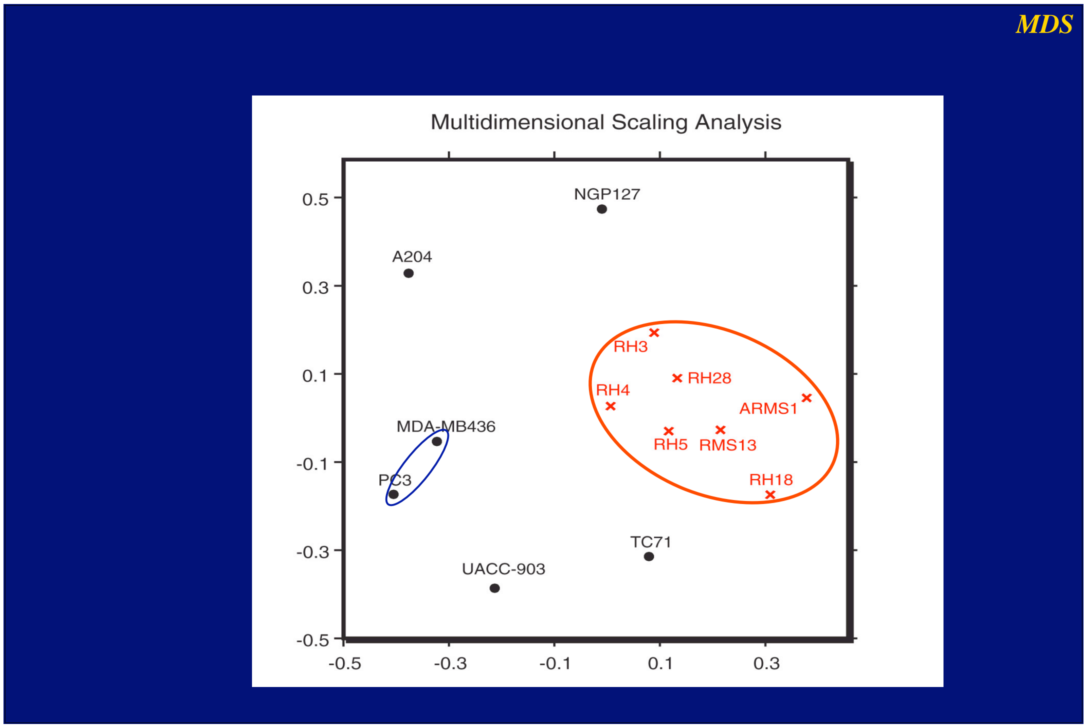


Multidimensional Scaling

Similarity Measure
Distance Measure

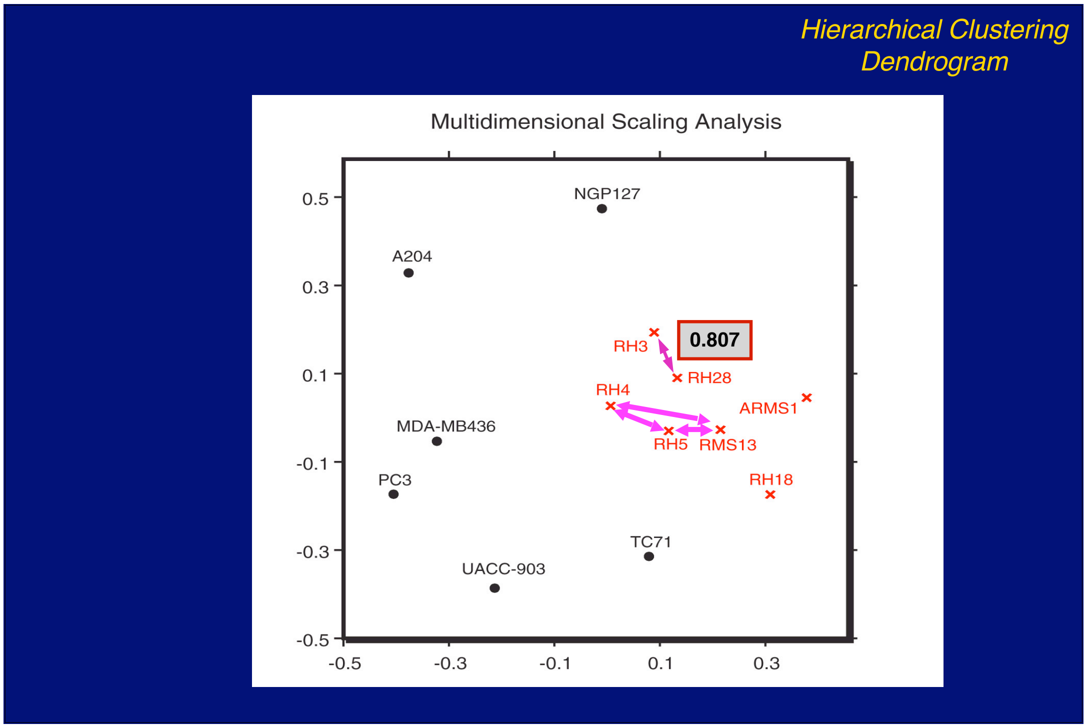
	Seattle	L.A.	Houston	D.C.
Seattle	0	1,200	2,500	2,800
L.A.		0	1,500	2,600
Houston			0	1400
D.C.				0

	Exp1	Exp2	Exp3	Exp4
Exp1	0	0.50	0.25	0.05
Exp2		0	0.50	0.55
Exp3			0	0.60
Exp4				0

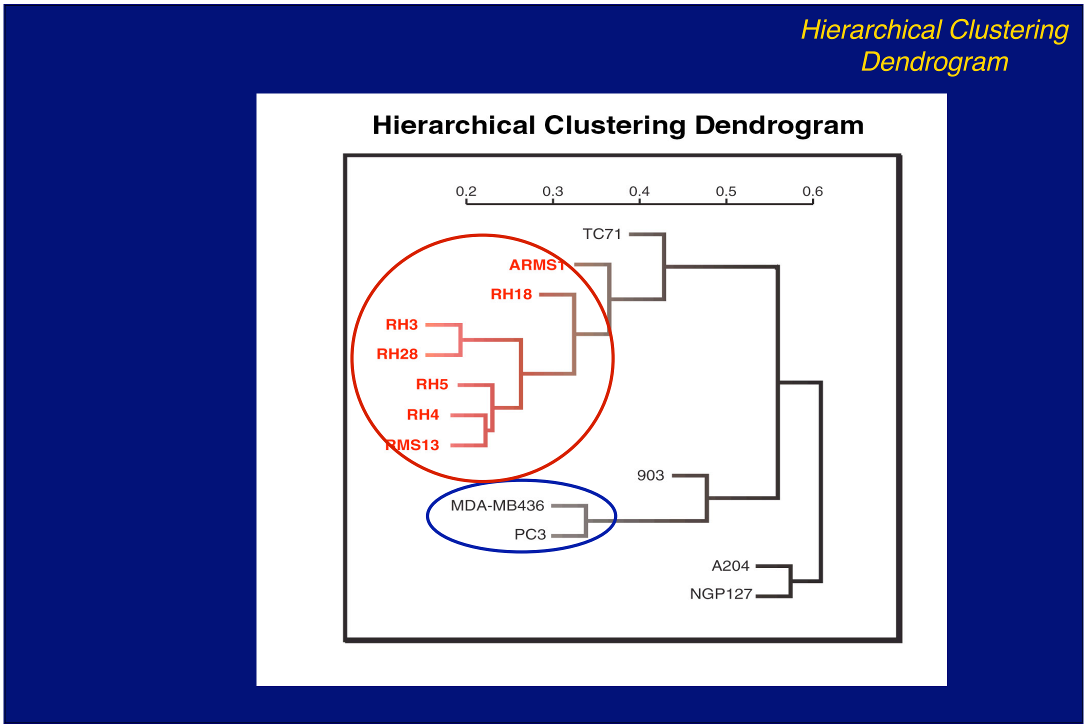


MDS

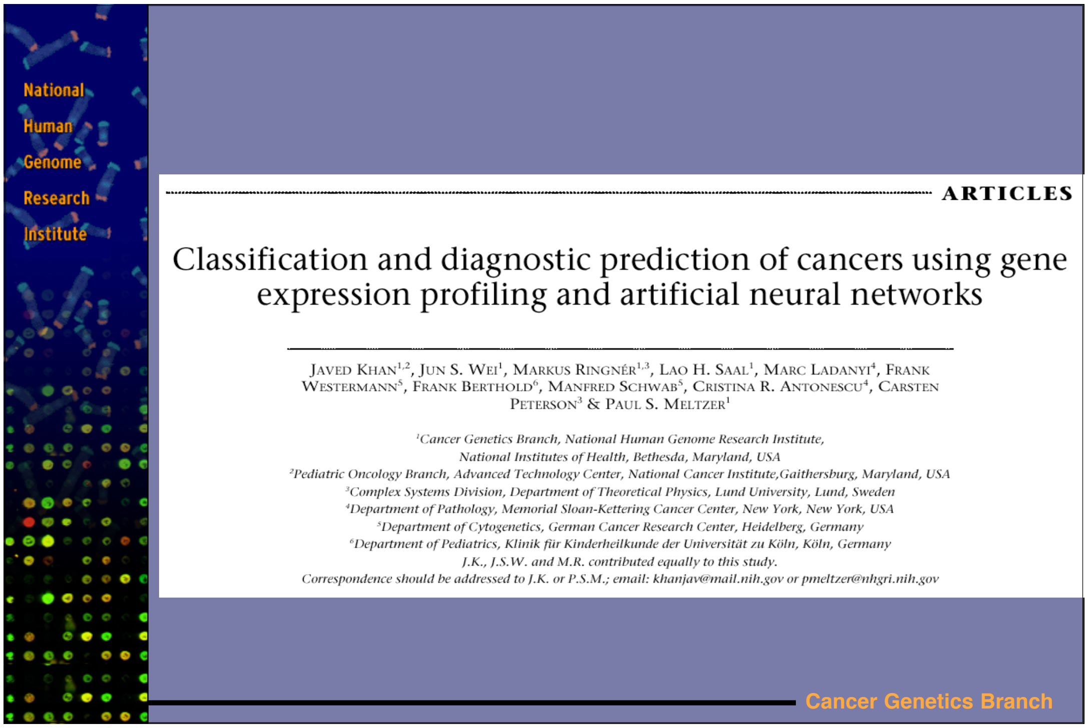
Multidimensional Scaling Analysis



Hierarchical Clustering Dendrogram



Hierarchical Clustering Dendrogram



ARTICLES

Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks

JAVED KHAN[1,2], JUN S. WEI[1], MARKUS RINGNÉR[1,3], LAO H. SAAL[1], MARC LADANYI[4], FRANK WESTERMANN[5], FRANK BERTHOLD[6], MANFRED SCHWAB[5], CRISTINA R. ANTONESCU[4], CARSTEN PETERSON[3] & PAUL S. MELTZER[1]

[1]*Cancer Genetics Branch, National Human Genome Research Institute,*
National Institutes of Health, Bethesda, Maryland, USA
[2]*Pediatric Oncology Branch, Advanced Technology Center, National Cancer Institute,Gaithersburg, Maryland, USA*
[3]*Complex Systems Division, Department of Theoretical Physics, Lund University, Lund, Sweden*
[4]*Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA*
[5]*Department of Cytogenetics, German Cancer Research Center, Heidelberg, Germany*
[6]*Department of Pediatrics, Klinik für Kinderheilkunde der Universität zu Köln, Köln, Germany*
J.K., J.S.W. and M.R. contributed equally to this study.
Correspondence should be addressed to J.K. or P.S.M.; email: khanjav@mail.nih.gov or pmeltzer@nhgri.nih.gov

Cancer Genetics Branch

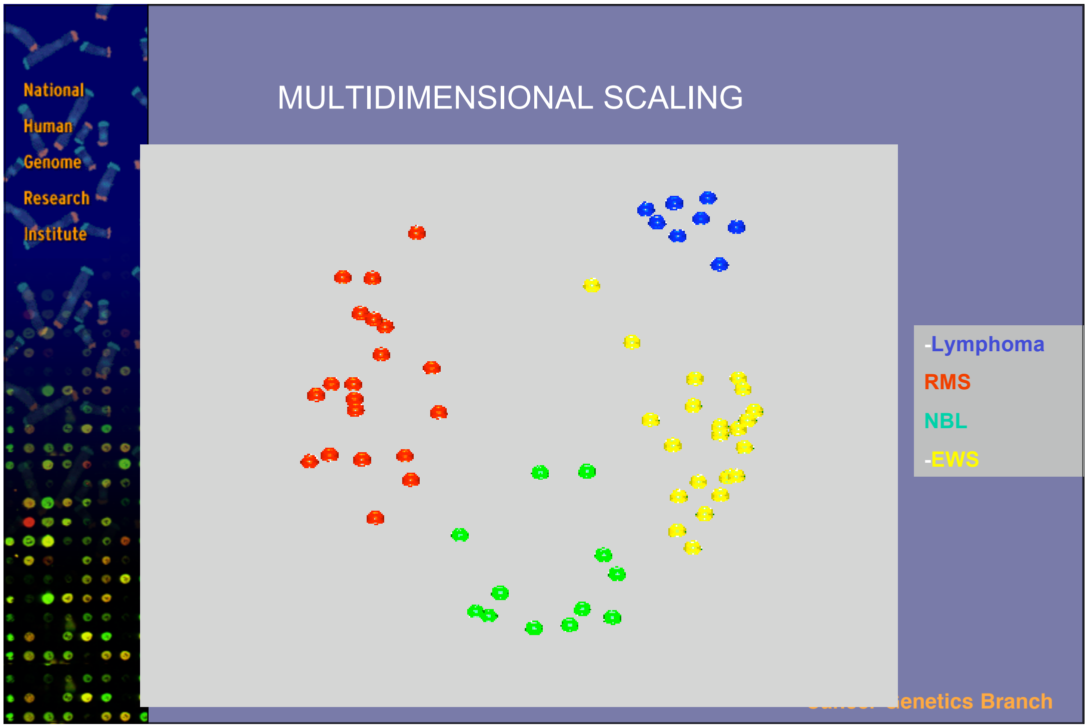
Model: Small Blue Round Cell Tumors

Ewing's

Neuroblastoma

Rhabdomyosarcoma

Lymphoma

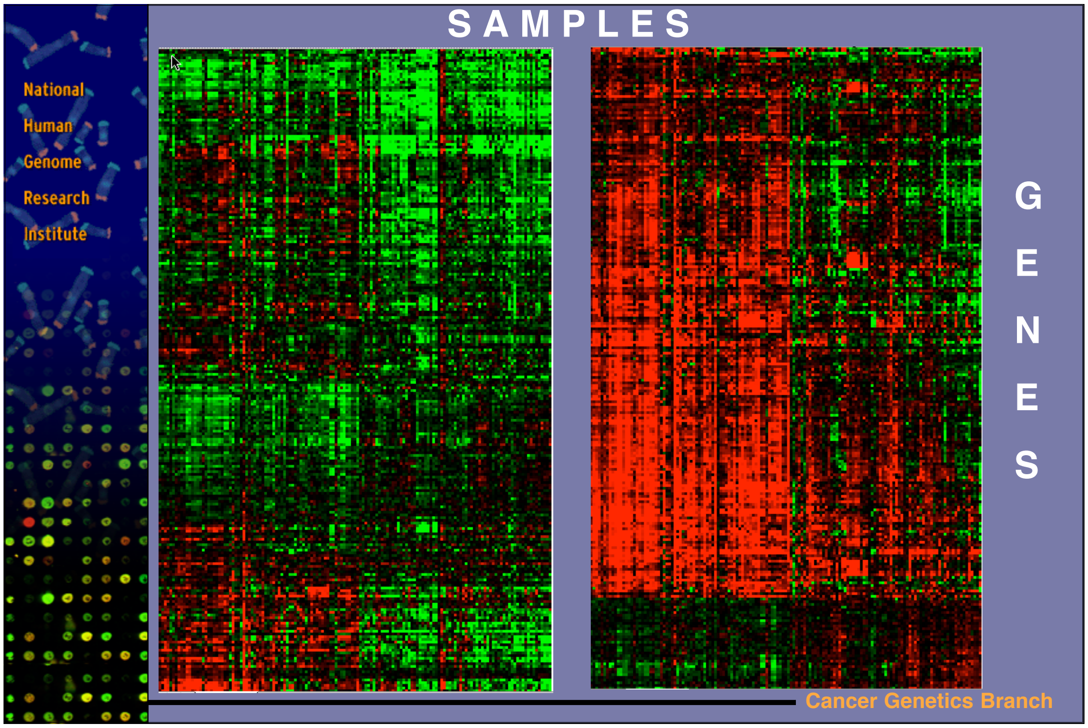


National
Human
Genome
Research
Institute

MULTIDIMENSIONAL SCALING

-Lymphoma
RMS
NBL
-EWS

Cancer Genetics Branch

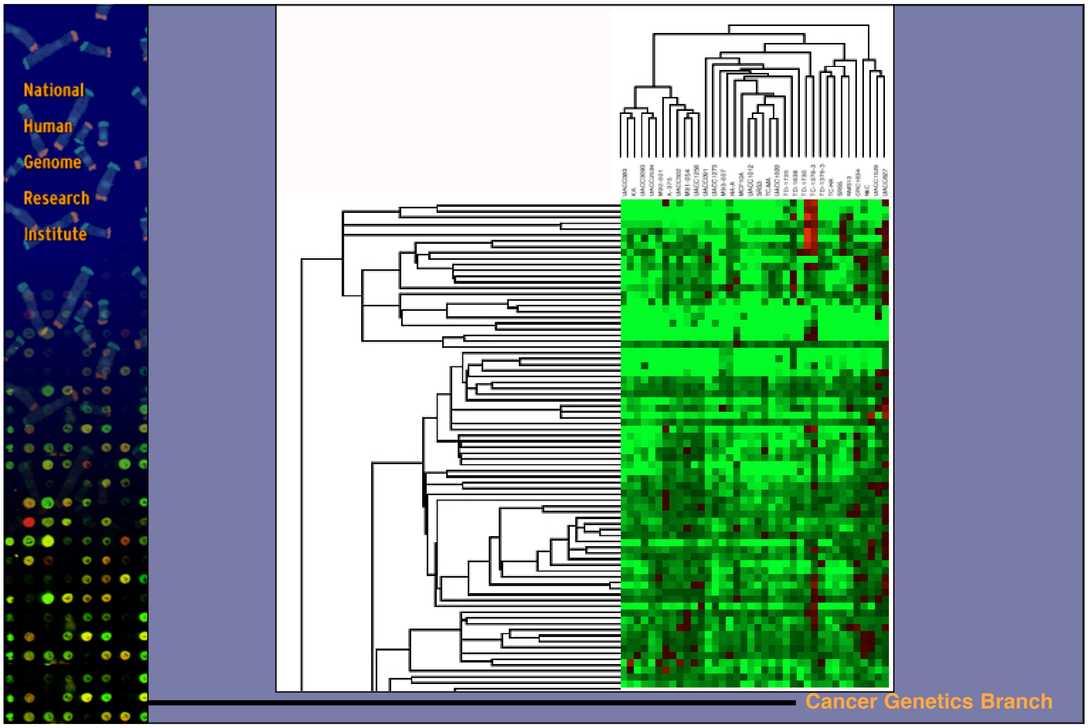


National
Human
Genome
Research
Institute

SAMPLES

GENES

Cancer Genetics Branch



National Human Genome Research Institute

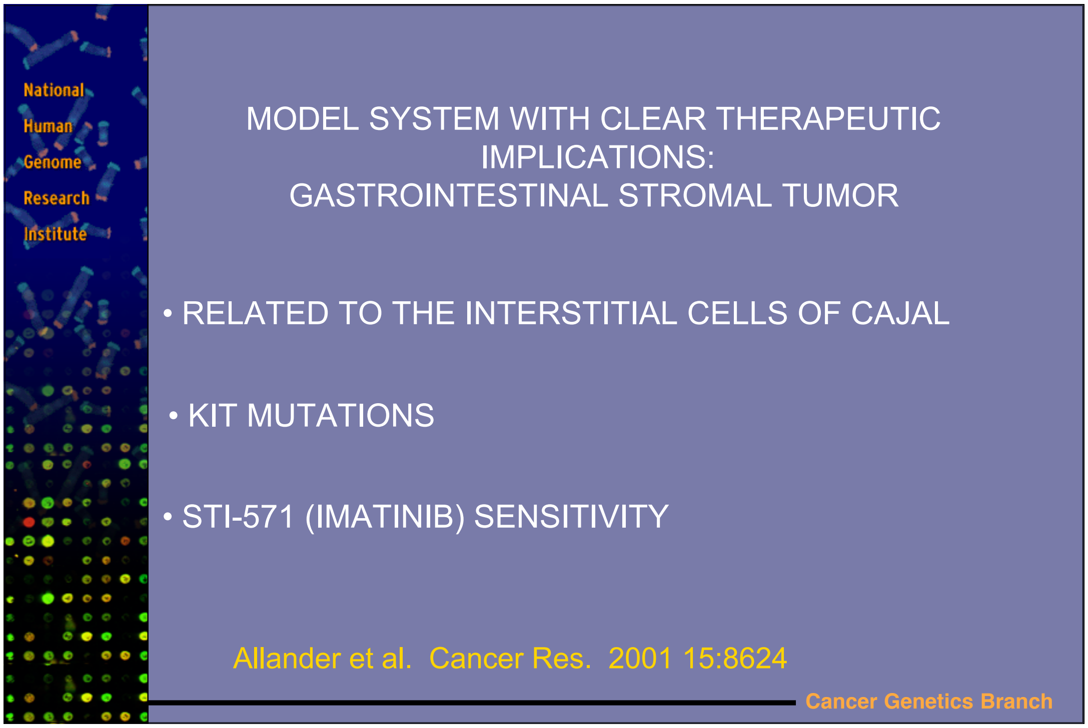
Cancer Genetics Branch

National Human Genome Research Institute

CLUSTERING METHODS

- Data display/visualization tools
- Require the addition of a statistical test.
  - Leave one out cross validation.
  - Random permutation test.
  - Test by introduction of noise.

Cancer Genetics Branch



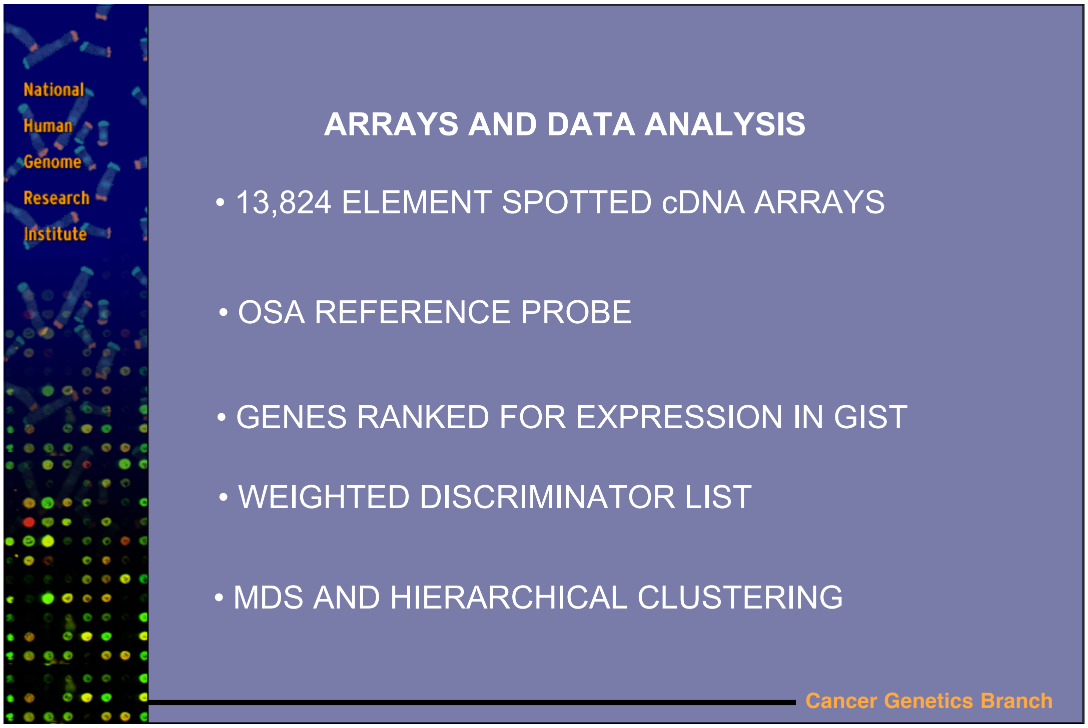
National
Human
Genome
Research
Institute

MODEL SYSTEM WITH CLEAR THERAPEUTIC IMPLICATIONS:
GASTROINTESTINAL STROMAL TUMOR

- RELATED TO THE INTERSTITIAL CELLS OF CAJAL
- KIT MUTATIONS
- STI-571 (IMATINIB) SENSITIVITY

Allander et al. Cancer Res. 2001 15:8624

Cancer Genetics Branch



National
Human
Genome
Research
Institute

ARRAYS AND DATA ANALYSIS

- 13,824 ELEMENT SPOTTED cDNA ARRAYS
- OSA REFERENCE PROBE
- GENES RANKED FOR EXPRESSION IN GIST
- WEIGHTED DISCRIMINATOR LIST
- MDS AND HIERARCHICAL CLUSTERING

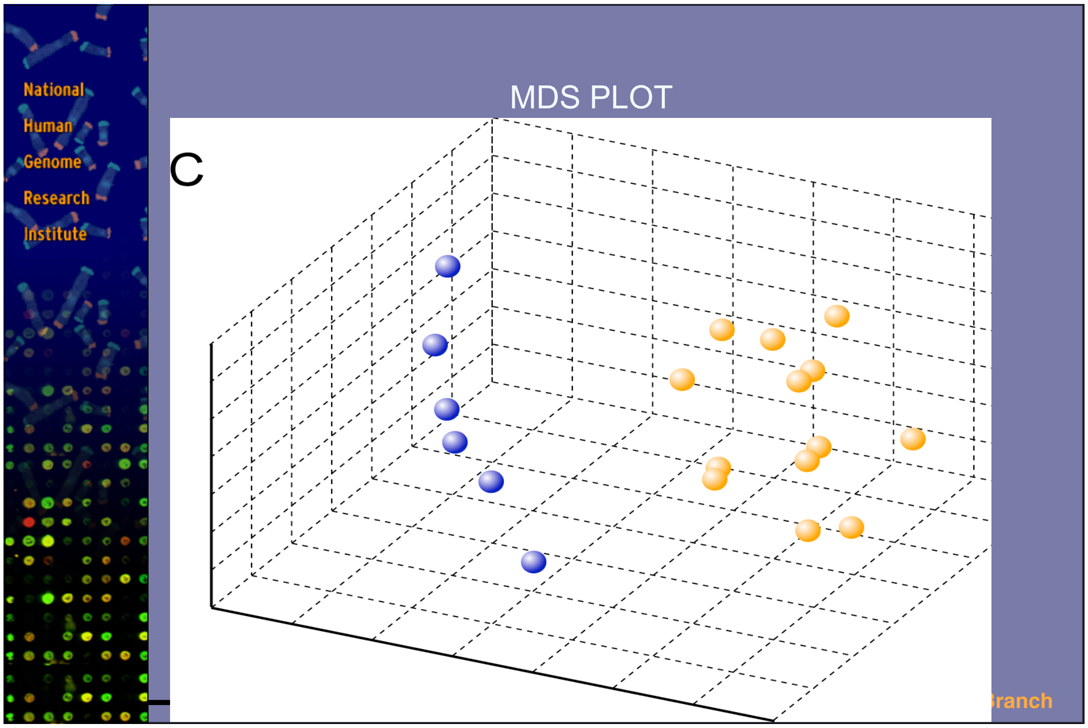
Cancer Genetics Branch

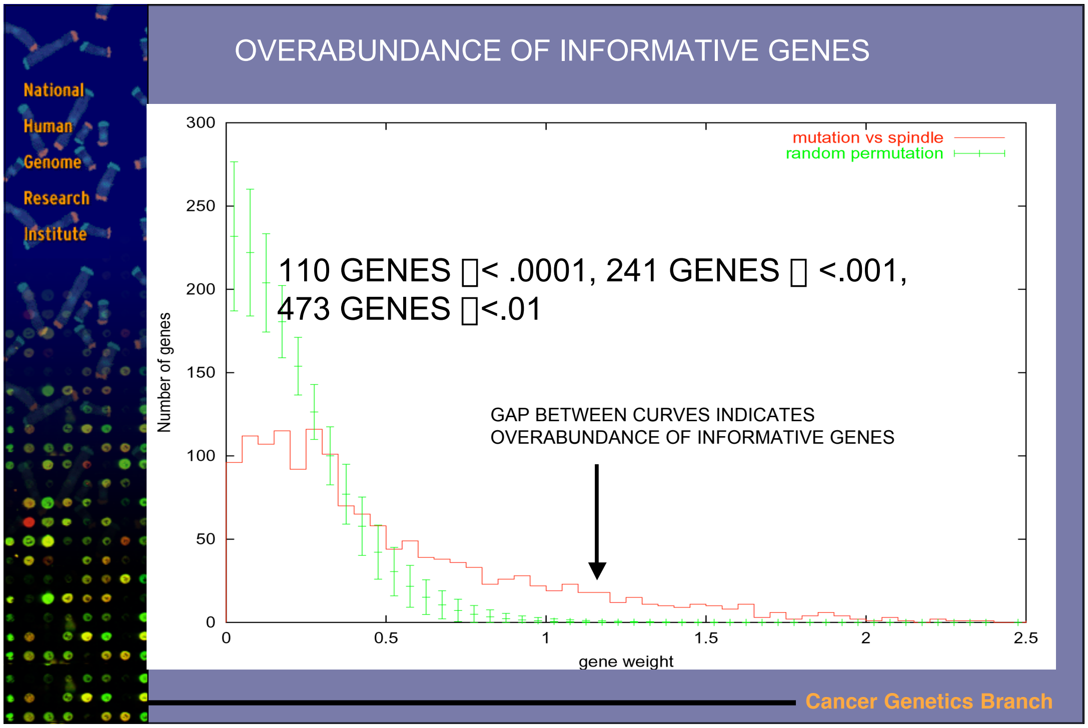
National
Human
Genome
Research
Institute

DATA ANALYSIS

- PREFILTER FOR QUALITY AND IDENTIFY GENES HIGHLY EXPRESSED IN GIST: 1987 GENES
- RANK BY WEIGHTED DISCRIMINATOR METHOD
  $w(g, \pm) = |\mu_+(g) - \mu_-(g)| / [\sigma_+(g) + \sigma_-(g)]$
- RANDOM PERMUTATION TEST (10^5 trials)
- CLUSTER ANALYSIS USING DISCRIMINATORS

Cancer Genetics Branch



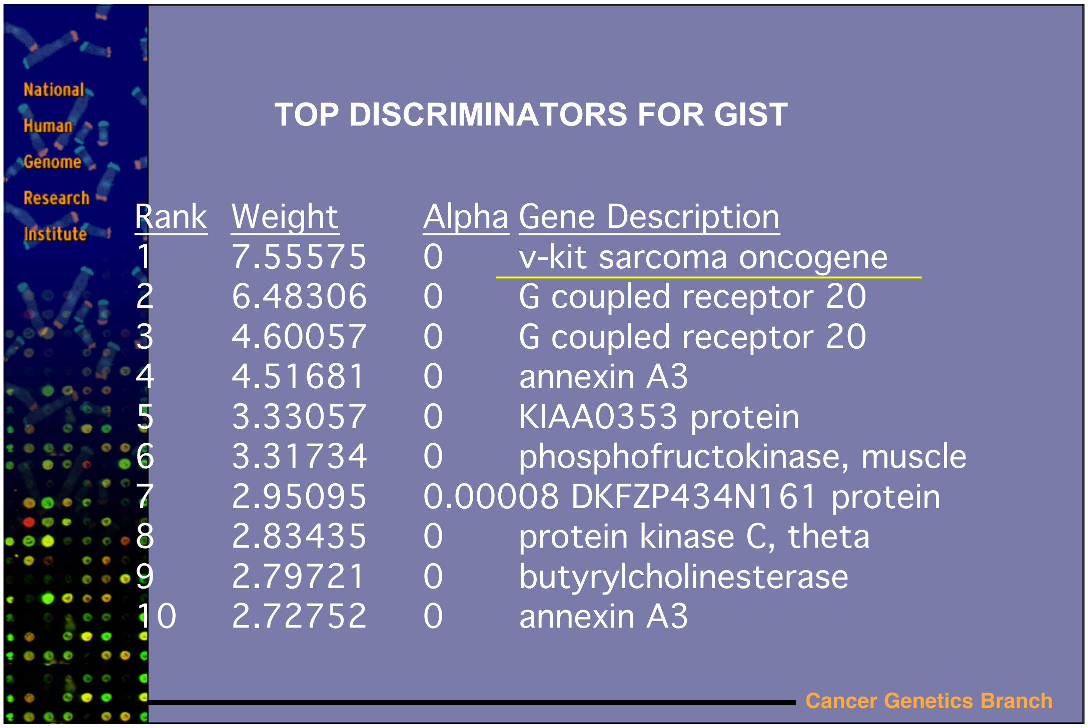


OVERABUNDANCE OF INFORMATIVE GENES

110 GENES α< .0001, 241 GENES α <.001,
473 GENES α<.01

GAP BETWEEN CURVES INDICATES
OVERABUNDANCE OF INFORMATIVE GENES

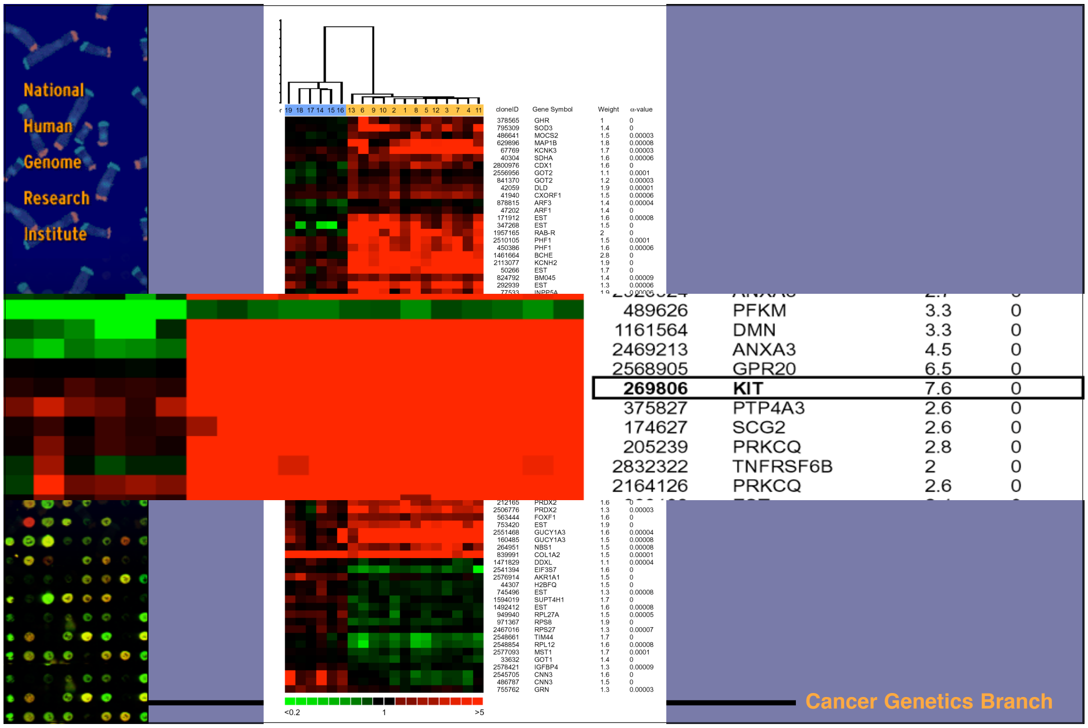
Cancer Genetics Branch



TOP DISCRIMINATORS FOR GIST

Rank	Weight	Alpha	Gene Description
1	7.55575	0	v-kit sarcoma oncogene
2	6.48306	0	G coupled receptor 20
3	4.60057	0	G coupled receptor 20
4	4.51681	0	annexin A3
5	3.33057	0	KIAA0353 protein
6	3.31734	0	phosphofructokinase, muscle
7	2.95095	0.00008	DKFZP434N161 protein
8	2.83435	0	protein kinase C, theta
9	2.79721	0	butyrylcholinesterase
10	2.72752	0	annexin A3

Cancer Genetics Branch



cloneID	Gene Symbol	Weight	α-value
378565	GHR	1	0
795309	SOD3	1.4	0
486641	MOCS2	1.5	0.00003
629896	MAP1B	1.8	0.00008
67769	KCNK3	1.7	0.00003
40304	SDHA	1.6	0.00006
2800976	CDX1	1.6	0
2556956	GOT2	1.1	0.0001
841370	GOT2	1.2	0.00003
42059	DLD	1.9	0.00001
41940	CXORF1	1.5	0.00006
878815	ARF3	1.4	0.00004
47202	ARF1	1.4	0
171912	EST	1.6	0.00008
347268	EST	1.5	0
1957165	RAB-R	2	0
2510105	PHF1	1.5	0.0001
450386	PHF1	1.6	0.00006
1461664	BCHE	2.8	0
2113077	KCNH2	1.9	0
50266	EST	1.7	0
824792	BM045	1.4	0.00009
292939	EST	1.3	0.00006
77533	INPP5A	1.9	0.00006
489626	PFKM	3.3	0
1161564	DMN	3.3	0
2469213	ANXA3	4.5	0
2568905	GPR20	6.5	0
269806	**KIT**	**7.6**	**0**
375827	PTP4A3	2.6	0
174627	SCG2	2.6	0
205239	PRKCQ	2.8	0
2832322	TNFRSF6B	2	0
2164126	PRKCQ	2.6	0
212165	PRDX2	1.6	0
2506776	PRDX2	1.3	0.00003
563444	FOXF1	1.6	0
753420	EST	1.9	0
2551468	GUCY1A3	1.6	0.00004
160485	GUCY1A3	1.5	0.00008
264951	NBS1	1.5	0.00008
839991	COL1A2	1.5	0.00001
1471829	DDXL	1.1	0.00004
2541394	EIF3S7	1.6	0
2576914	AKR1A1	1.5	0
44307	H2BFQ	1.5	0
745496	EST	1.3	0.00008
1594019	SUPT4H1	1.7	0
1492412	EST	1.6	0.00008
949940	RPL27A	1.5	0.00005
971367	RPS8	1.9	0
2467016	RPS27	1.3	0.00007
2548661	TIM44	1.7	0
2548854	RPL12	1.6	0.00008
2577093	MST1	1.7	0.0001
33632	GOT1	1.4	0
2578421	IGFBP4	1.3	0.00009
2545705	CNN3	1.6	0
486787	CNN3	1.5	0
755762	GRN	1.3	0.00003

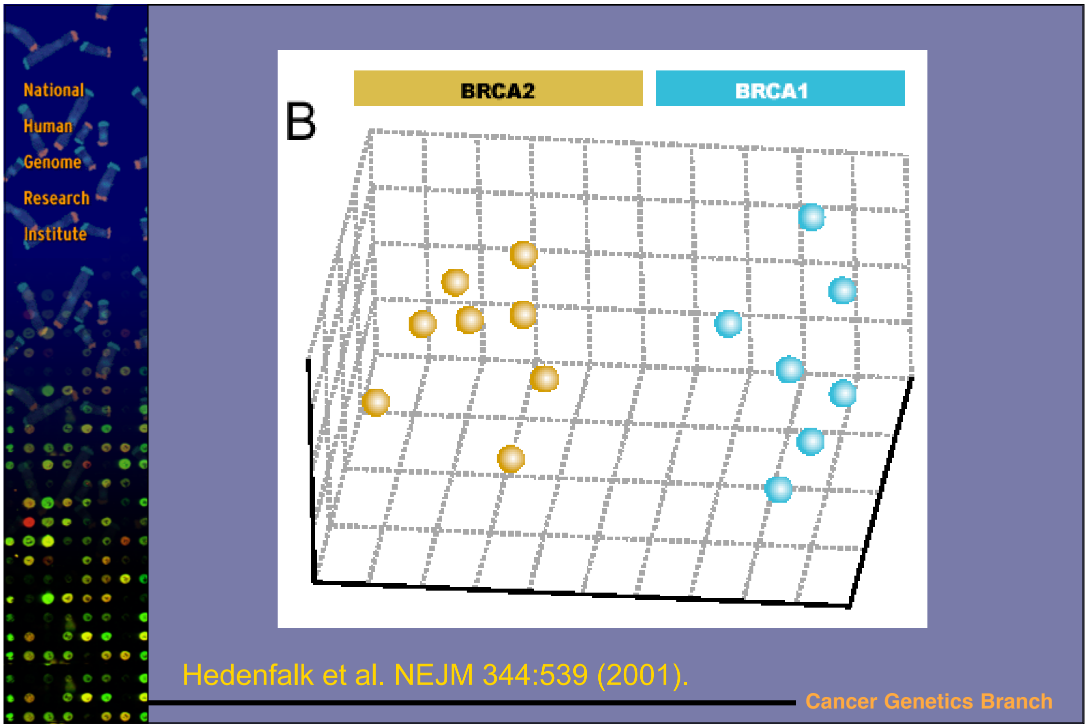
Cancer Genetics Branch

National
Human
Genome
Research
Institute

HEREDITARY CANCER:

DO INHERITED MUTATIONS IN CANCER SUSCEPTIBILITY GENES LEAVE A DISTINCT FOOTPRINT ON THE PATTERN OF TUMOR GENE EXPRESSION?

Cancer Genetics Branch



National
Human
Genome
Research
Institute

Hedenfalk et al. NEJM 344:539 (2001).

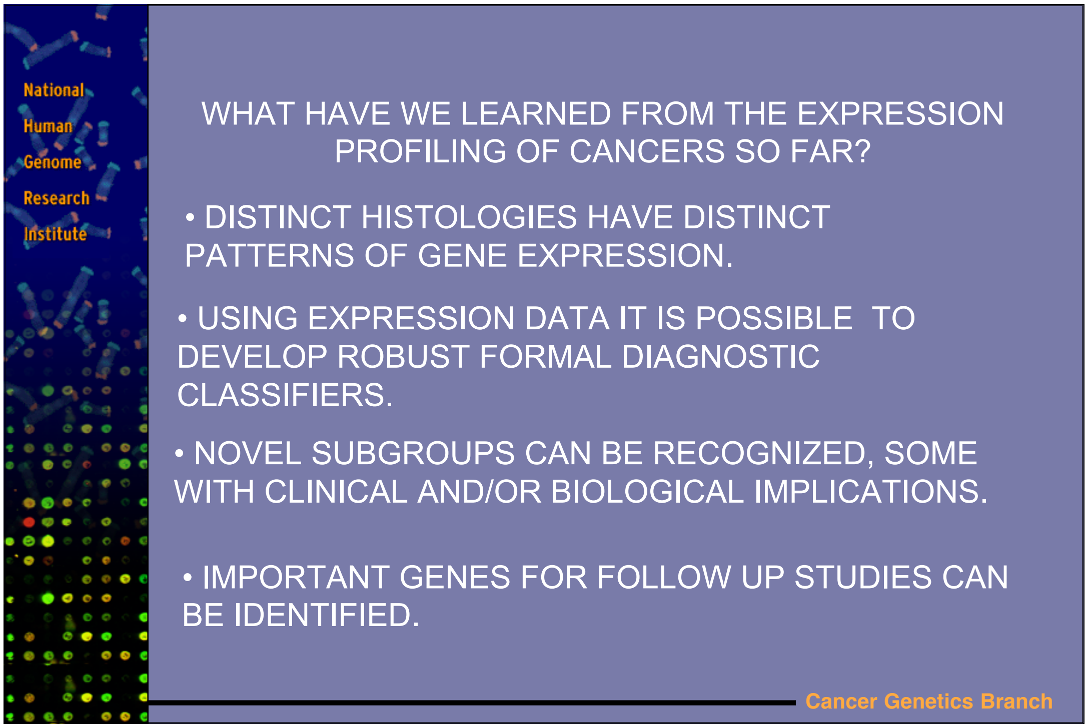
Cancer Genetics Branch

National
Human
Genome
Research
Institute

WHAT SHOULD YOU LOOK FOR IN A CLINICAL MICROARRAY STUDY?

ARE MICROARRAY TECHNOLOGIES READY TO BE IMPLEMENTED IN CLINICAL PRACTICE?

Cancer Genetics Branch

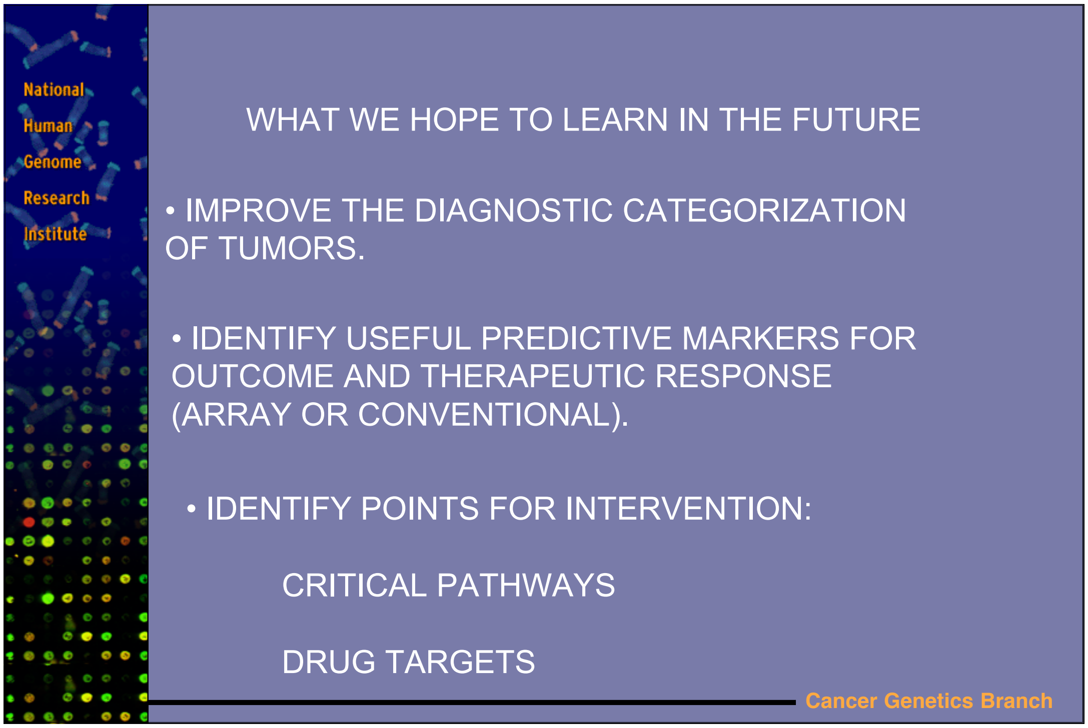


National
Human
Genome
Research
Institute

WHAT HAVE WE LEARNED FROM THE EXPRESSION PROFILING OF CANCERS SO FAR?

- DISTINCT HISTOLOGIES HAVE DISTINCT PATTERNS OF GENE EXPRESSION.
- USING EXPRESSION DATA IT IS POSSIBLE TO DEVELOP ROBUST FORMAL DIAGNOSTIC CLASSIFIERS.
- NOVEL SUBGROUPS CAN BE RECOGNIZED, SOME WITH CLINICAL AND/OR BIOLOGICAL IMPLICATIONS.
- IMPORTANT GENES FOR FOLLOW UP STUDIES CAN BE IDENTIFIED.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

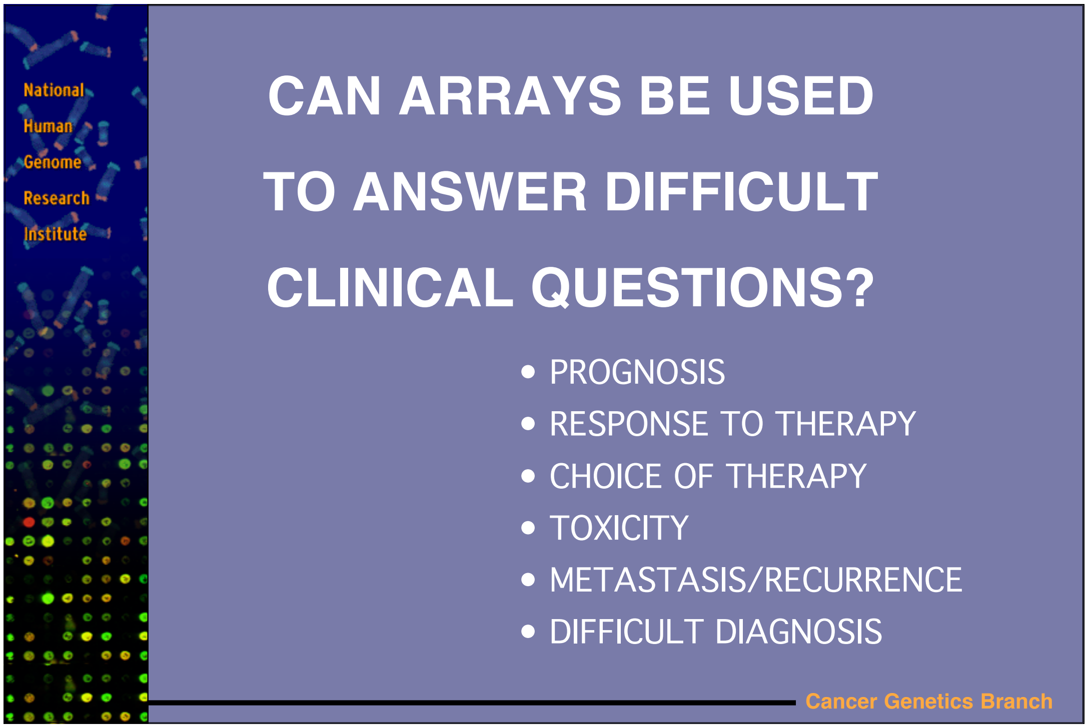
WHAT WE HOPE TO LEARN IN THE FUTURE

- IMPROVE THE DIAGNOSTIC CATEGORIZATION OF TUMORS.
- IDENTIFY USEFUL PREDICTIVE MARKERS FOR OUTCOME AND THERAPEUTIC RESPONSE (ARRAY OR CONVENTIONAL).
- IDENTIFY POINTS FOR INTERVENTION:

  CRITICAL PATHWAYS

  DRUG TARGETS

Cancer Genetics Branch

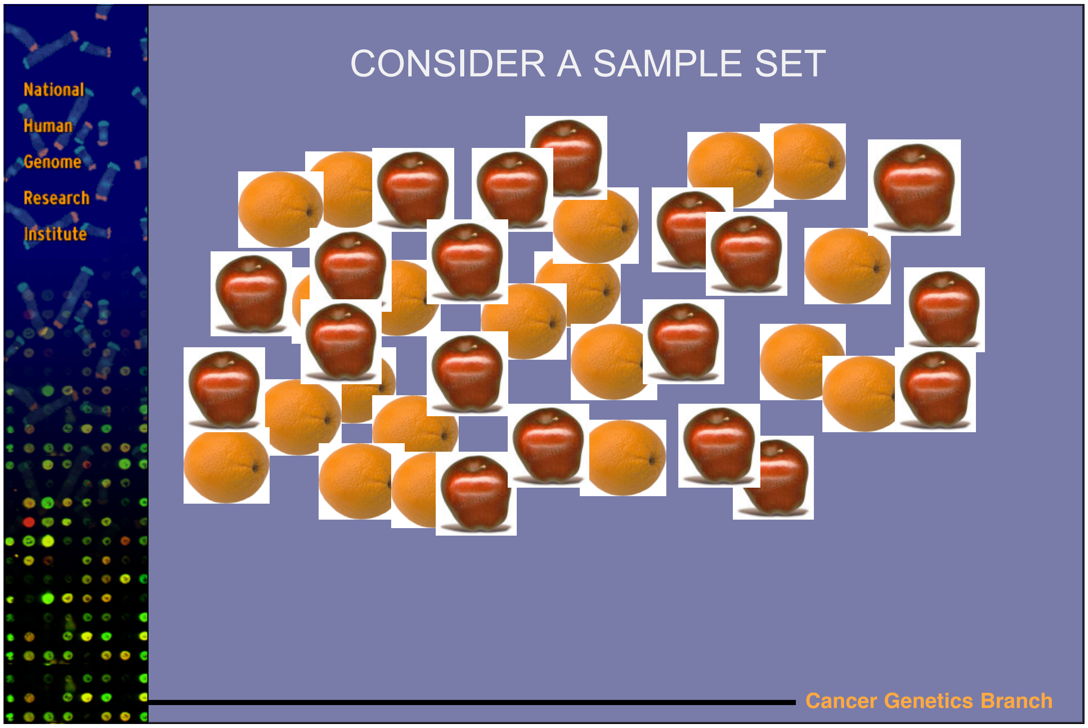


National
Human
Genome
Research
Institute

CAN ARRAYS BE USED TO ANSWER DIFFICULT CLINICAL QUESTIONS?

- PROGNOSIS
- RESPONSE TO THERAPY
- CHOICE OF THERAPY
- TOXICITY
- METASTASIS/RECURRENCE
- DIFFICULT DIAGNOSIS

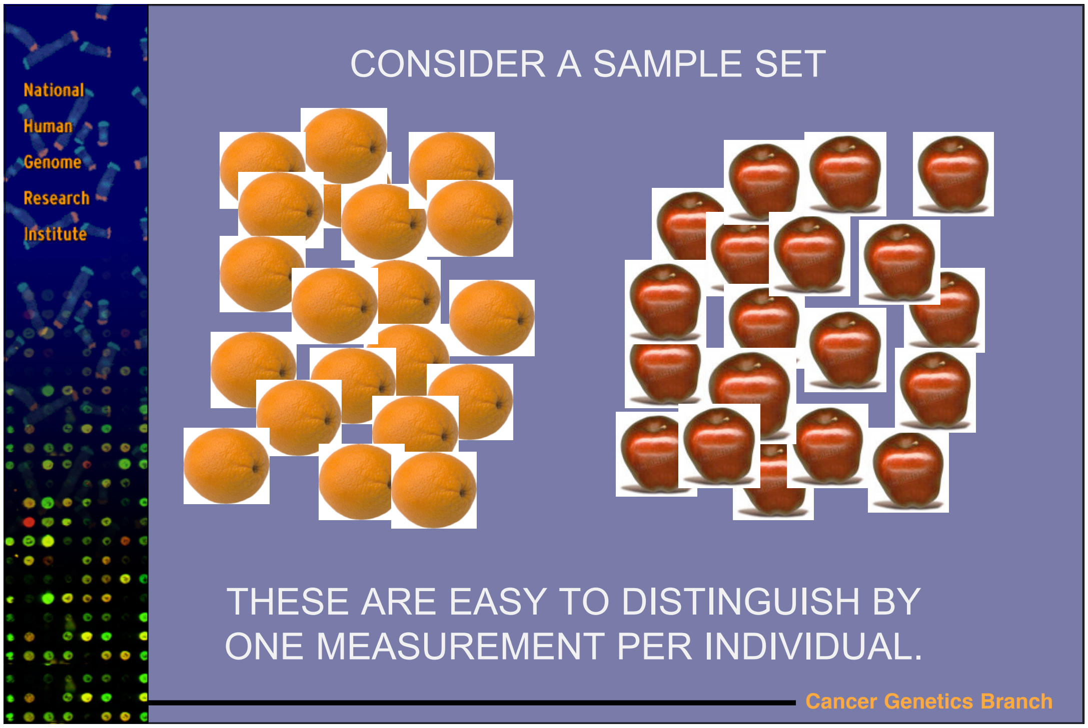
Cancer Genetics Branch



National
Human
Genome
Research
Institute

CONSIDER A SAMPLE SET

Cancer Genetics Branch

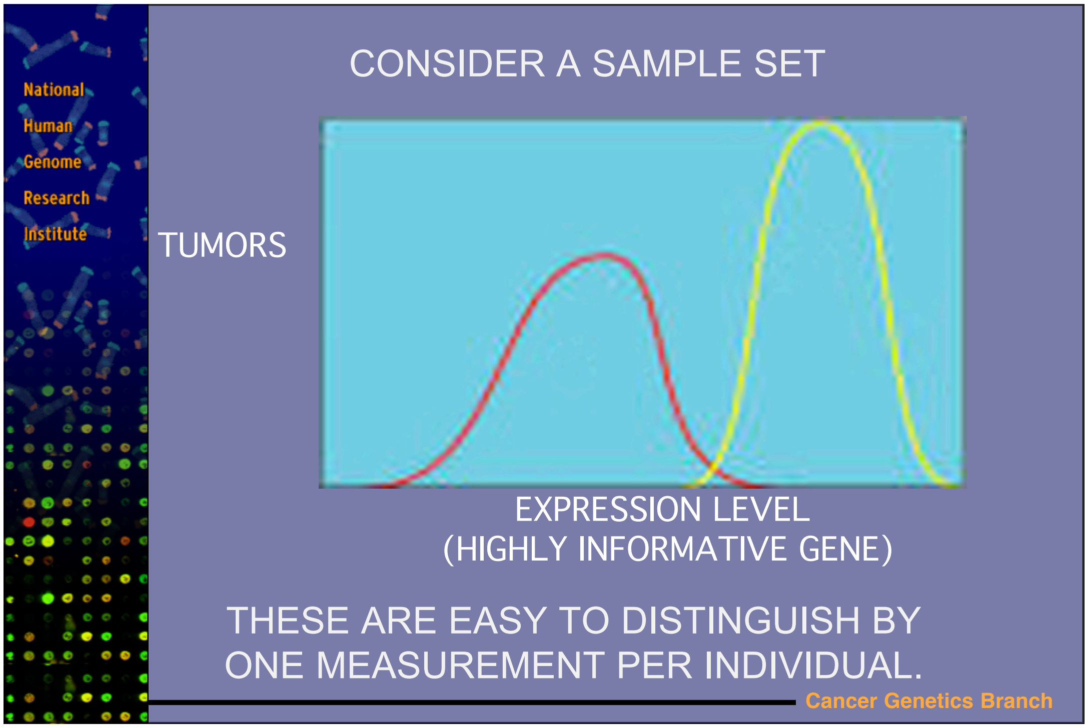


National
Human
Genome
Research
Institute

CONSIDER A SAMPLE SET

THESE ARE EASY TO DISTINGUISH BY ONE MEASUREMENT PER INDIVIDUAL.

Cancer Genetics Branch



National
Human
Genome
Research
Institute

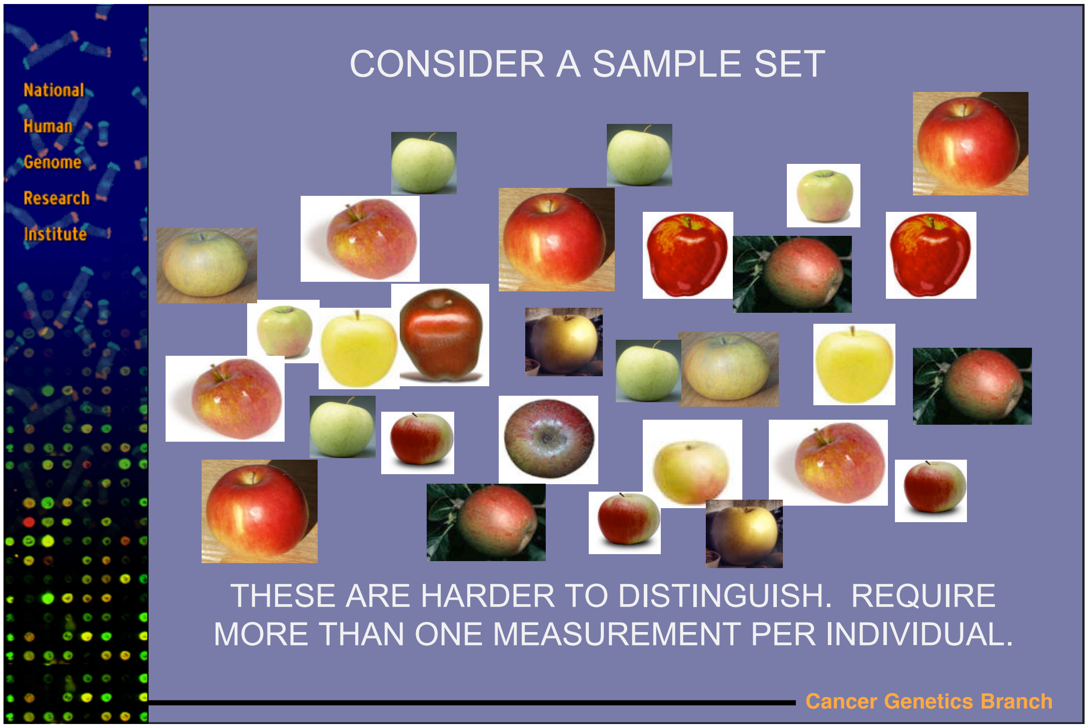
CONSIDER A SAMPLE SET

TUMORS

EXPRESSION LEVEL
(HIGHLY INFORMATIVE GENE)

THESE ARE EASY TO DISTINGUISH BY ONE MEASUREMENT PER INDIVIDUAL.

Cancer Genetics Branch

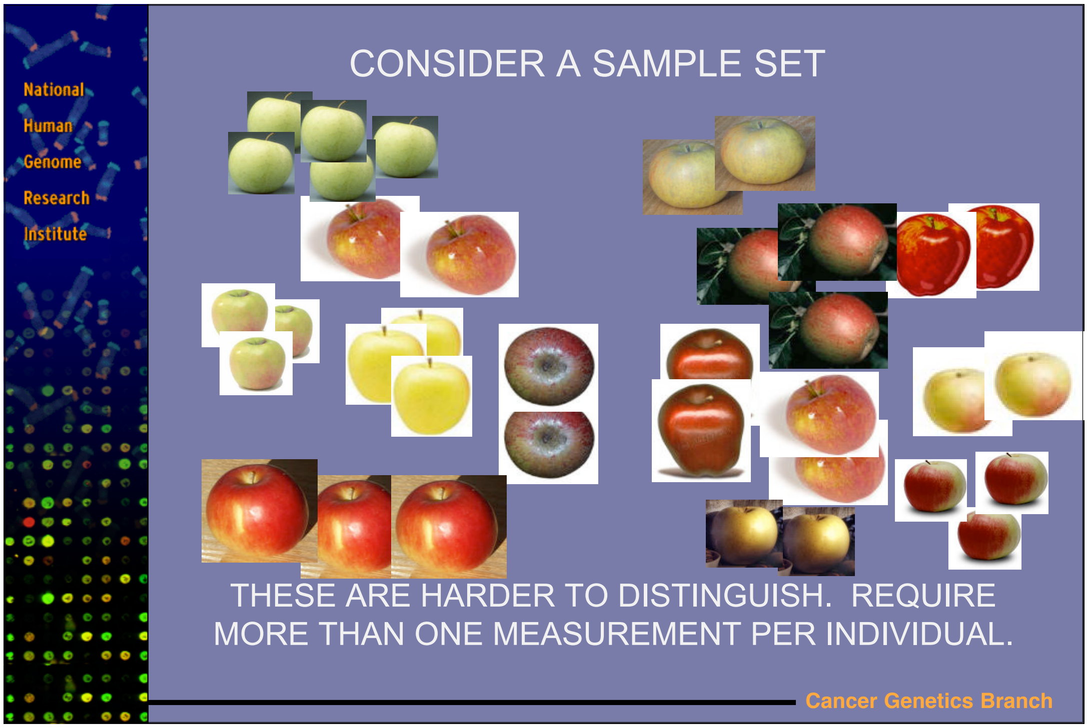


National
Human
Genome
Research
Institute

CONSIDER A SAMPLE SET

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

Cancer Genetics Branch

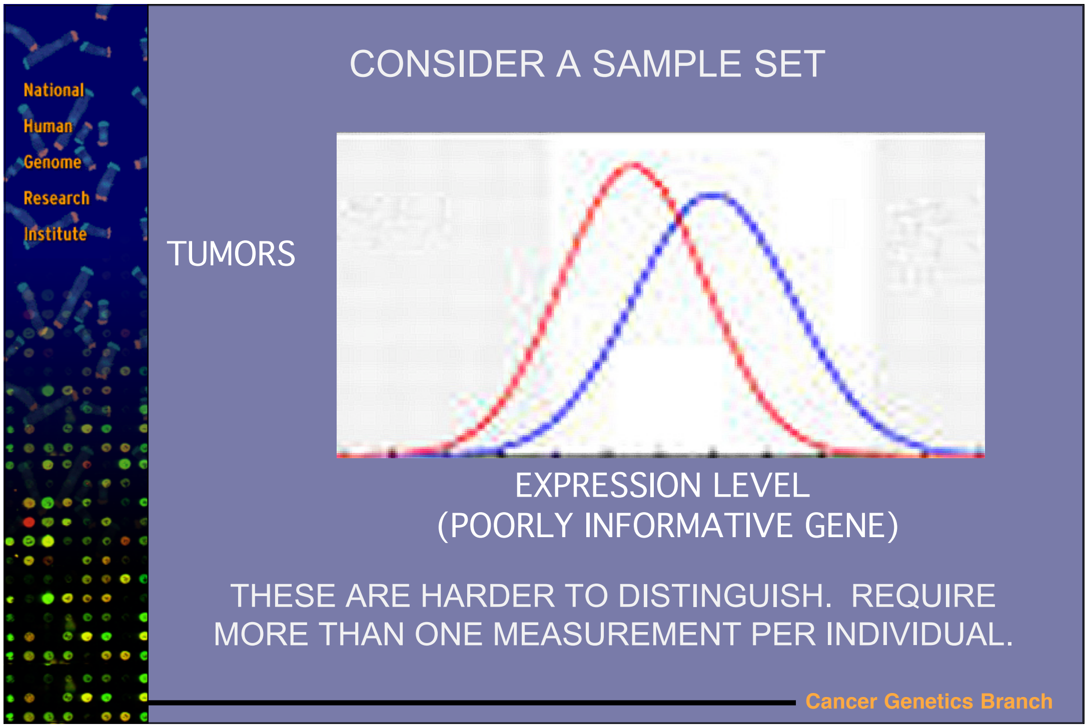


National
Human
Genome
Research
Institute

CONSIDER A SAMPLE SET

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

Cancer Genetics Branch



National
Human
Genome
Research
Institute

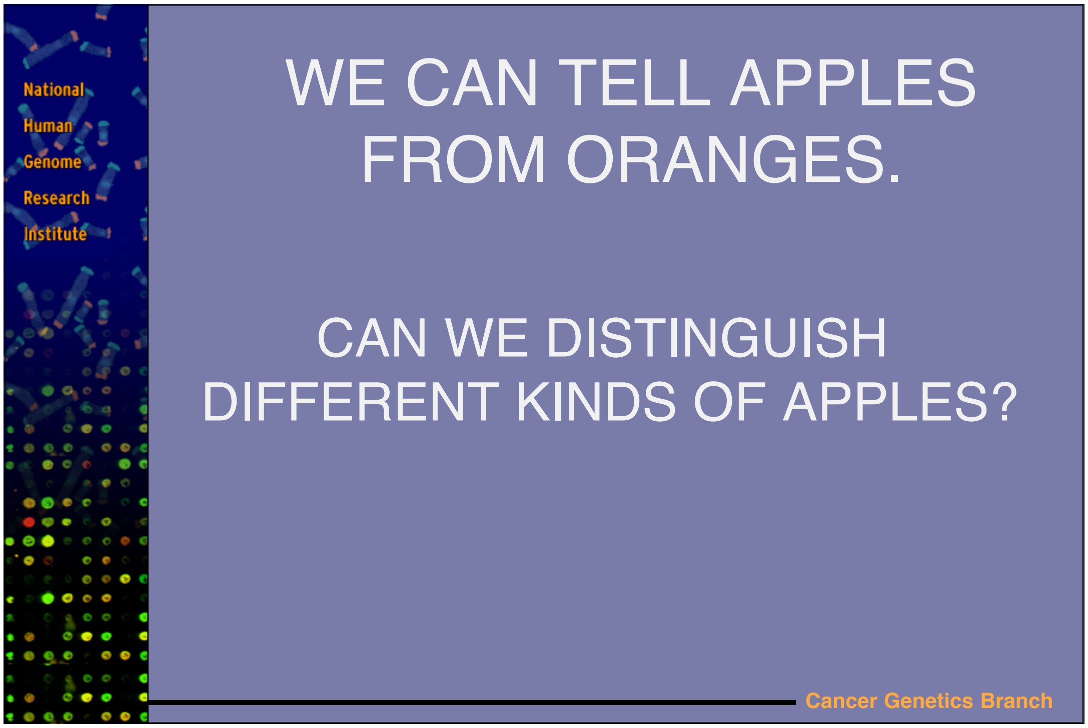
CONSIDER A SAMPLE SET

TUMORS

EXPRESSION LEVEL
(POORLY INFORMATIVE GENE)

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

Cancer Genetics Branch

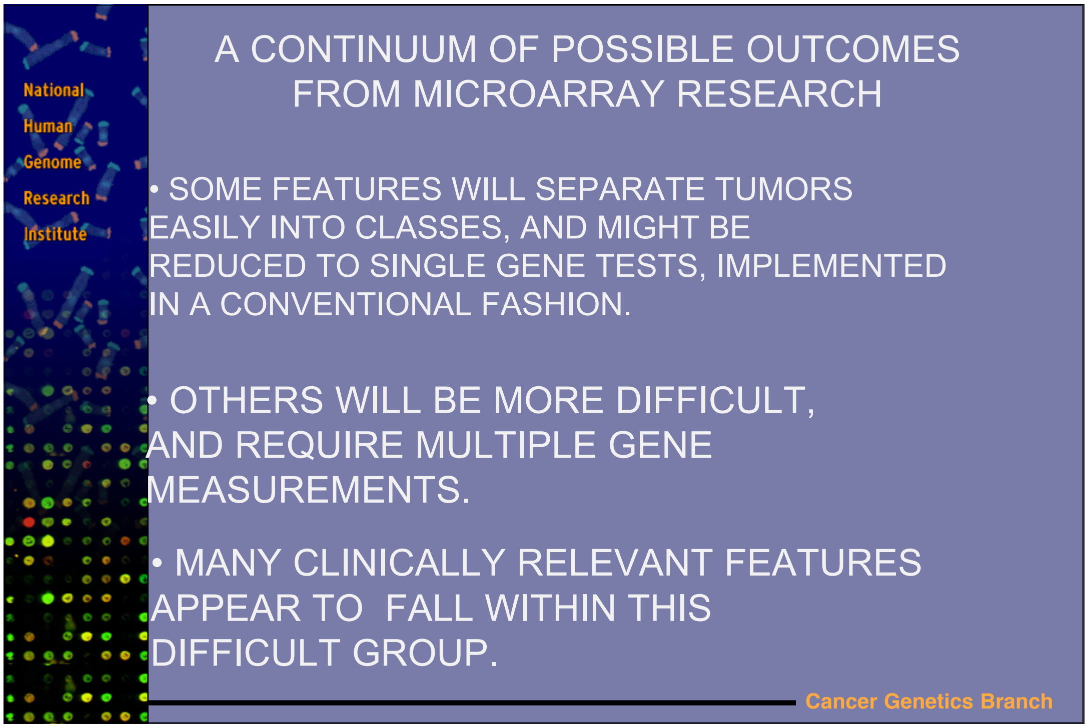


National
Human
Genome
Research
Institute

WE CAN TELL APPLES FROM ORANGES.

CAN WE DISTINGUISH DIFFERENT KINDS OF APPLES?

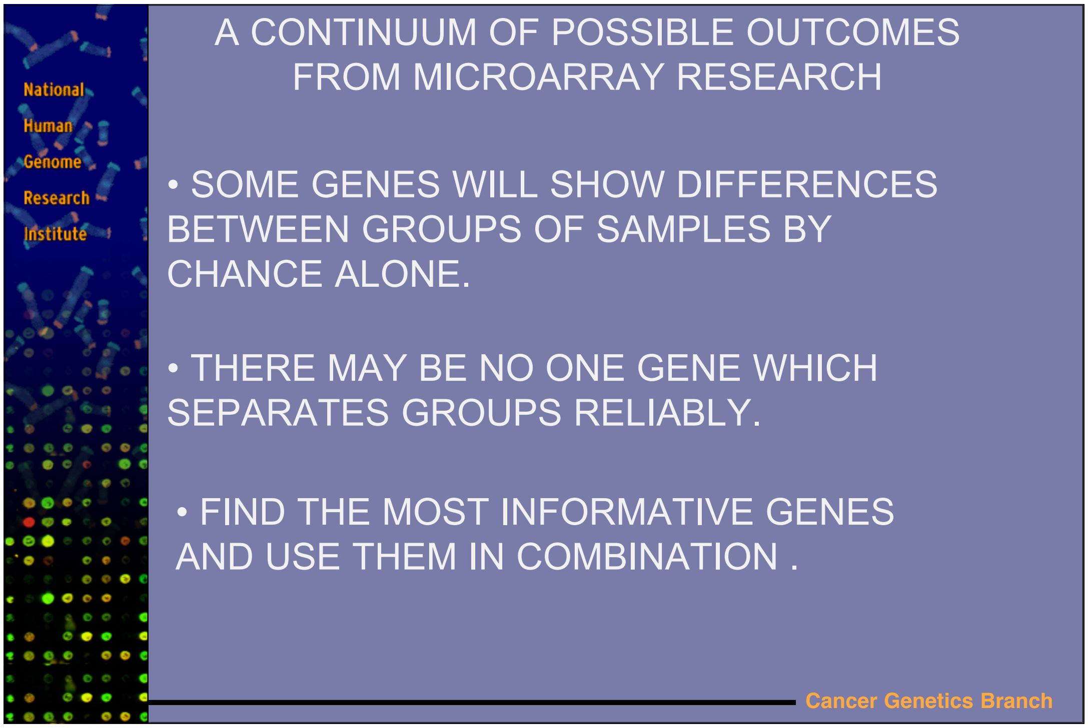
Cancer Genetics Branch



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.

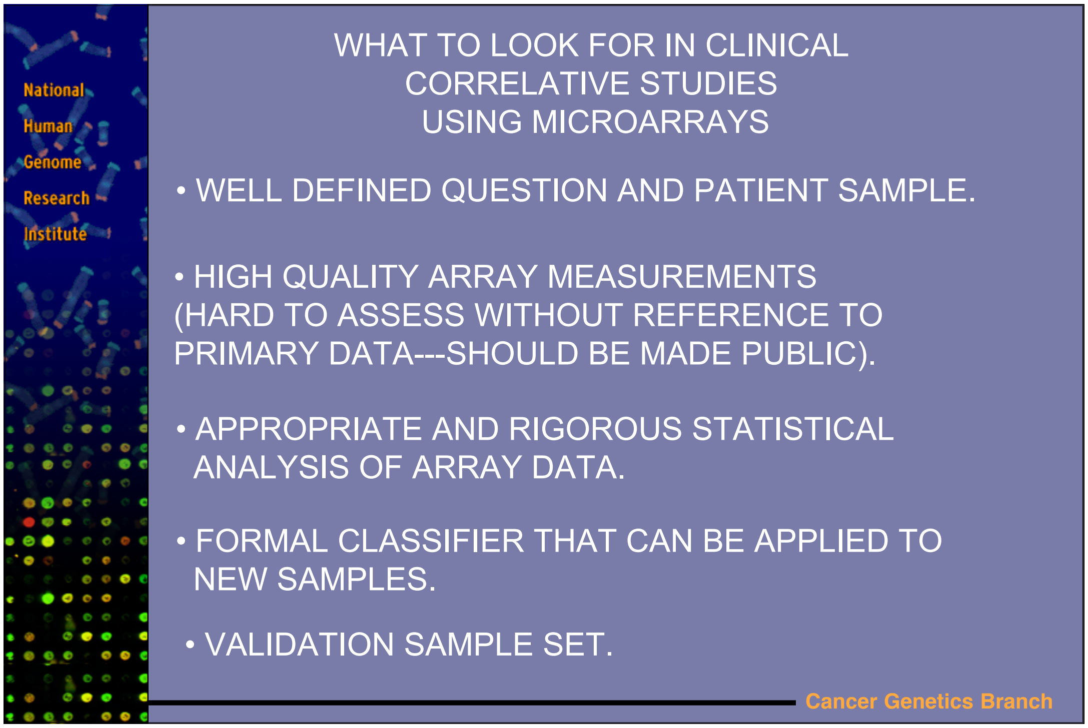
Cancer Genetics Branch



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 .

Cancer Genetics Branch

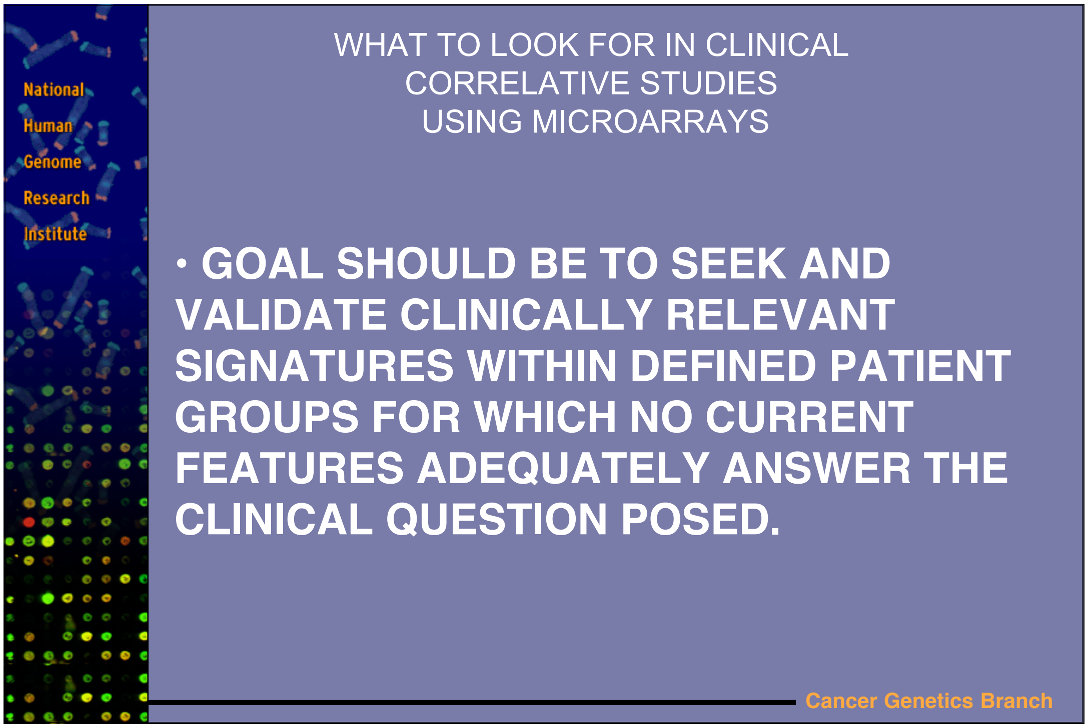


National
Human
Genome
Research
Institute

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

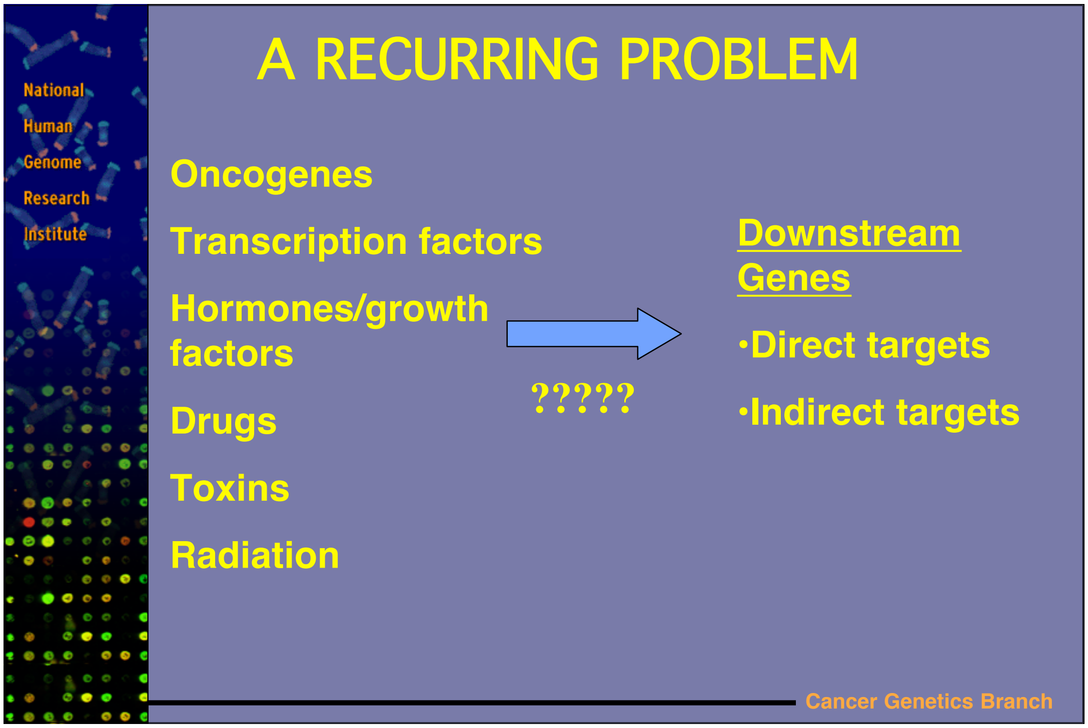


National
Human
Genome
Research
Institute

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.**

Cancer Genetics Branch



National
Human
Genome
Research
Institute

A RECURRING PROBLEM

Oncogenes

Transcription factors

Hormones/growth factors

Drugs

Toxins

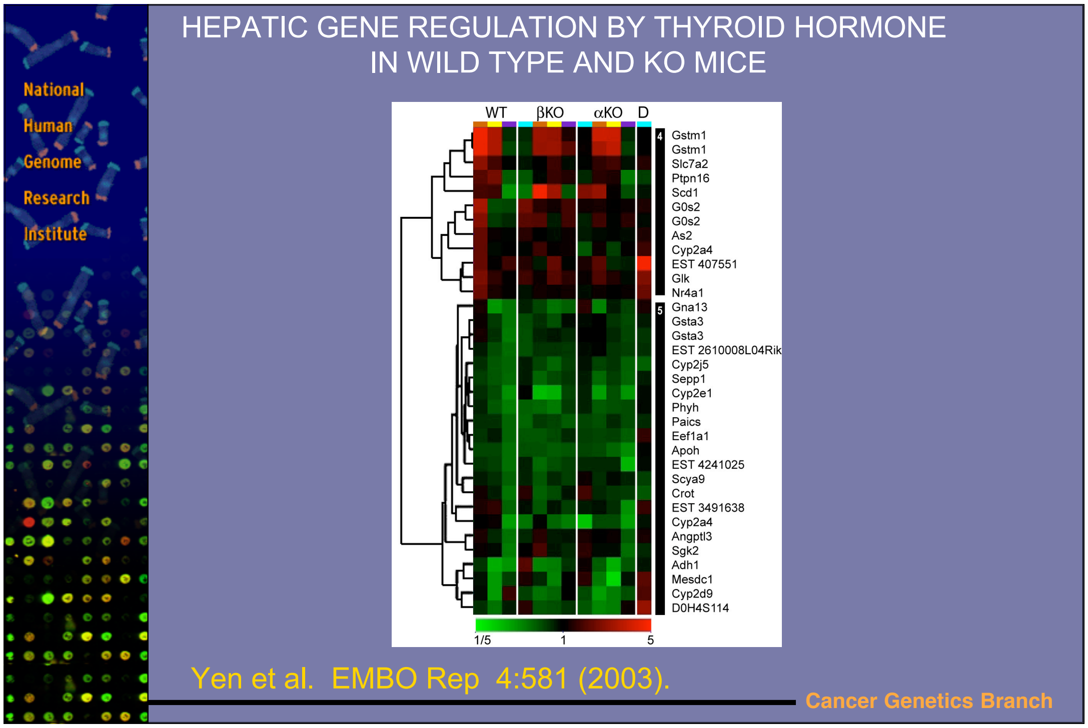
Radiation

?????

Downstream Genes

- **Direct targets**
- **Indirect targets**

Cancer Genetics Branch

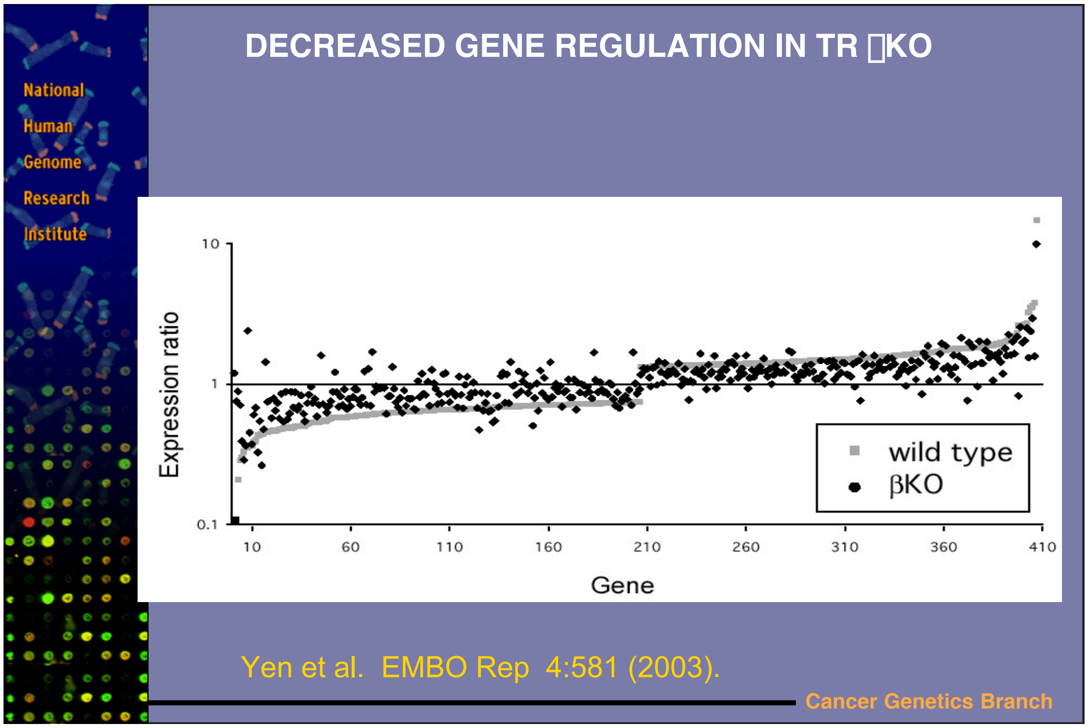


National
Human
Genome
Research
Institute

HEPATIC GENE REGULATION BY THYROID HORMONE IN WILD TYPE AND KO MICE

Yen et al. EMBO Rep 4:581 (2003).

Cancer Genetics Branch

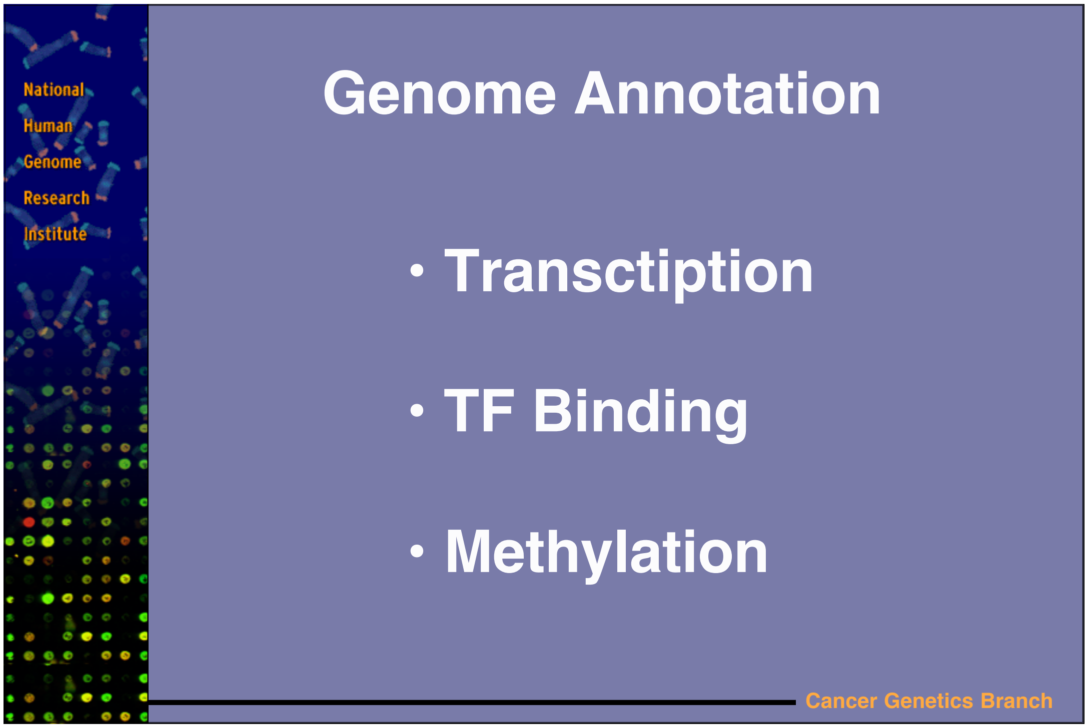


National
Human
Genome
Research
Institute

DECREASED GENE REGULATION IN TR □KO

Yen et al. EMBO Rep 4:581 (2003).

Cancer Genetics Branch



National
Human
Genome
Research
Institute

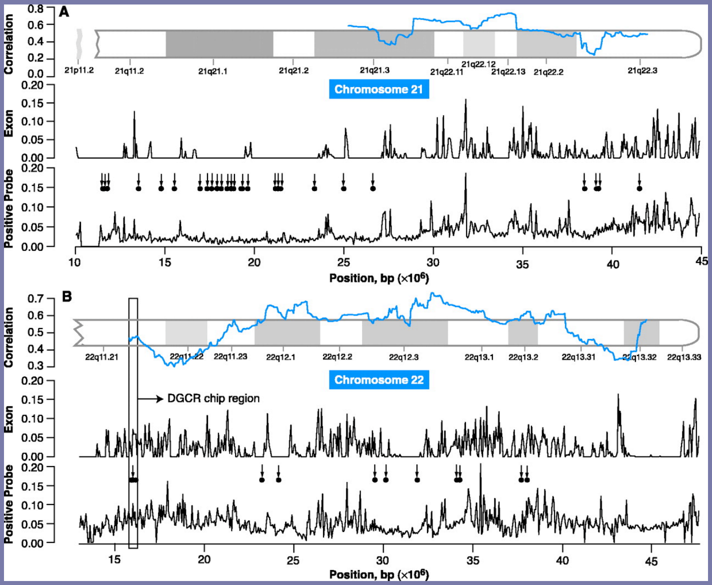
Genome Annotation

- Transctiption
- TF Binding
- Methylation

Cancer Genetics Branch

National
Human
Genome
Research
Institute

Scanning Chromosome with Tiling Path Arrays

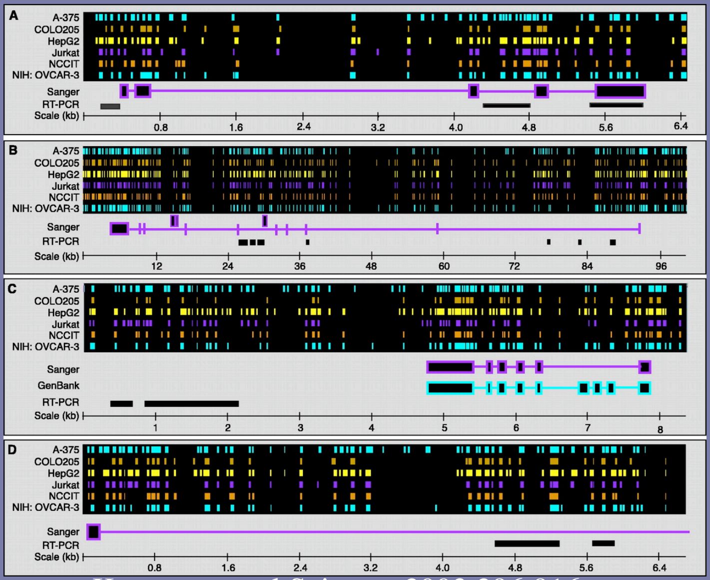


Kapranov et al Science. 2003 296:916.

Cancer Genetics Branch

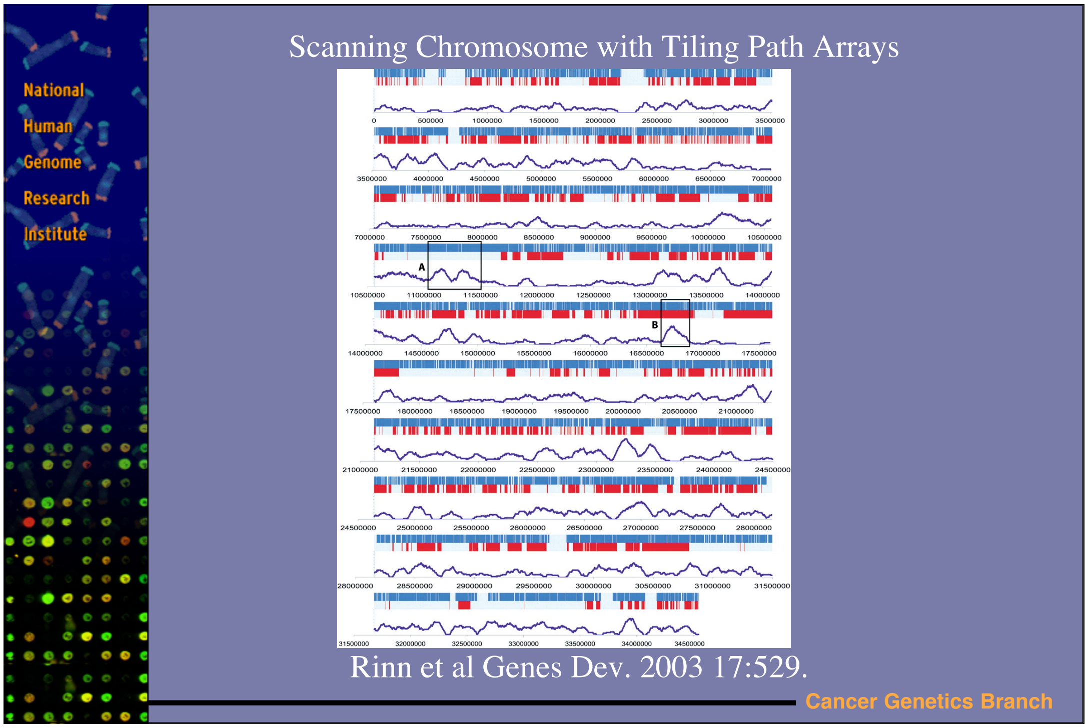
National
Human
Genome
Research
Institute

Scanning Chromosome with Tiling Path Arrays



Kapranov et al Science. 2003 296:916.

Cancer Genetics Branch

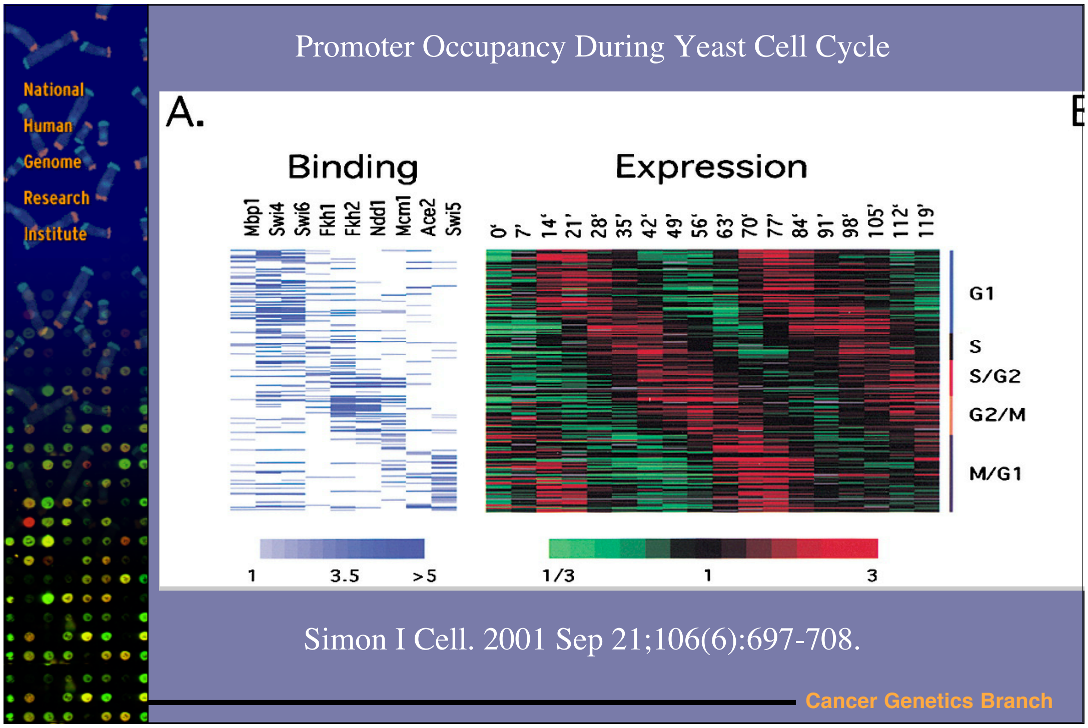


National
Human
Genome
Research
Institute

Scanning Chromosome with Tiling Path Arrays

Rinn et al Genes Dev. 2003 17:529.

Cancer Genetics Branch

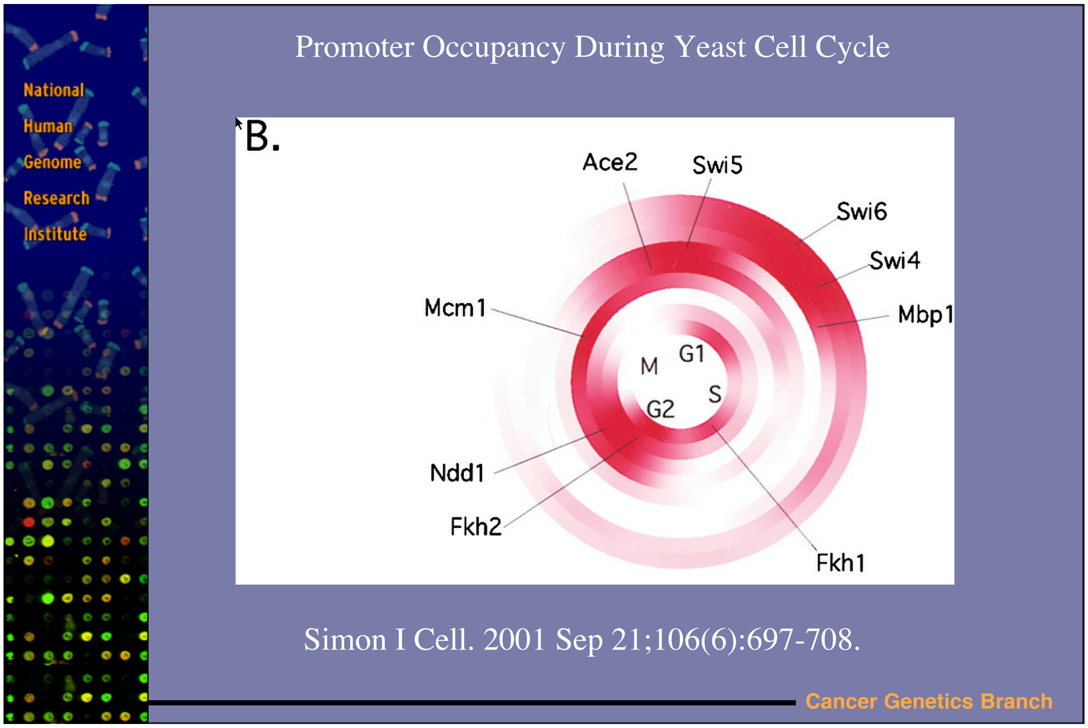


National
Human
Genome
Research
Institute

Promoter Occupancy During Yeast Cell Cycle

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

Cancer Genetics Branch

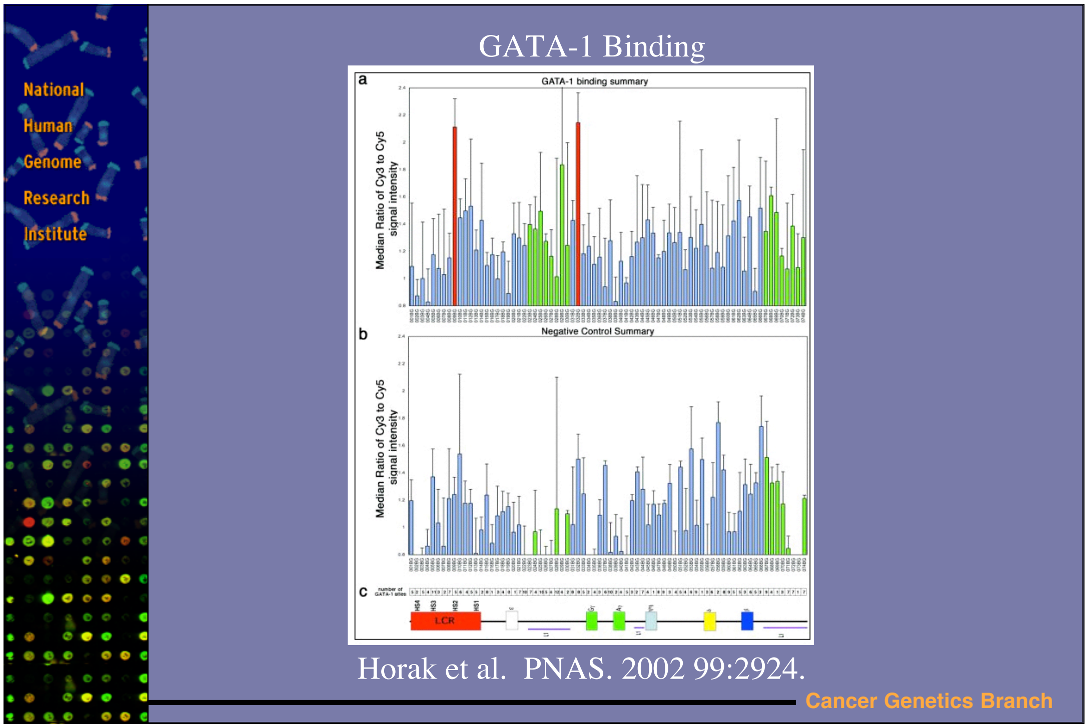


National
Human
Genome
Research
Institute

Promoter Occupancy During Yeast Cell Cycle

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

Cancer Genetics Branch

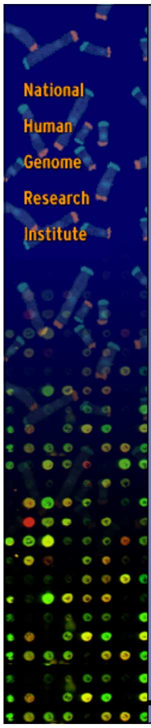


National
Human
Genome
Research
Institute

GATA-1 Binding

Horak et al. PNAS. 2002 99:2924.

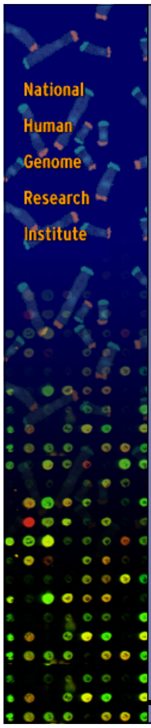
Cancer Genetics Branch

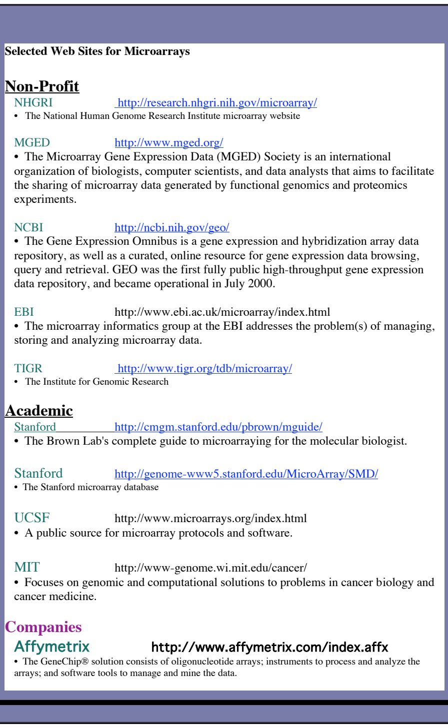


Future Promise of DNA Microarrays

- **Automated processing of gene lists for functional information using standard databases (GeneOntology).**
- **Linking genomic sequence to expression data to define regulatory elements/transcription factors associated with co-regulated genes.**
- **Introduction of technologies which increase the dimensionality of expression data: CGH arrays; methylation arrays; promoter arrays (CHIP on chips).**
- **Development of computational tools which allow predictive modeling of gene networks.**

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.

Companies

Affymetrix	**http://www.affymetrix.com/index.affx**

- The GeneChip® solution consists of oligonucleotide arrays; instruments to process and analyze the arrays; and software tools to manage and mine the data.

Cancer Genetics Branch