



DNA MICROARRAY TECHNOLOGY

**AFTER THE SEQUENCE:
WHOLE GENOME APPROACHES TO
BIOLOGICAL QUESTIONS**

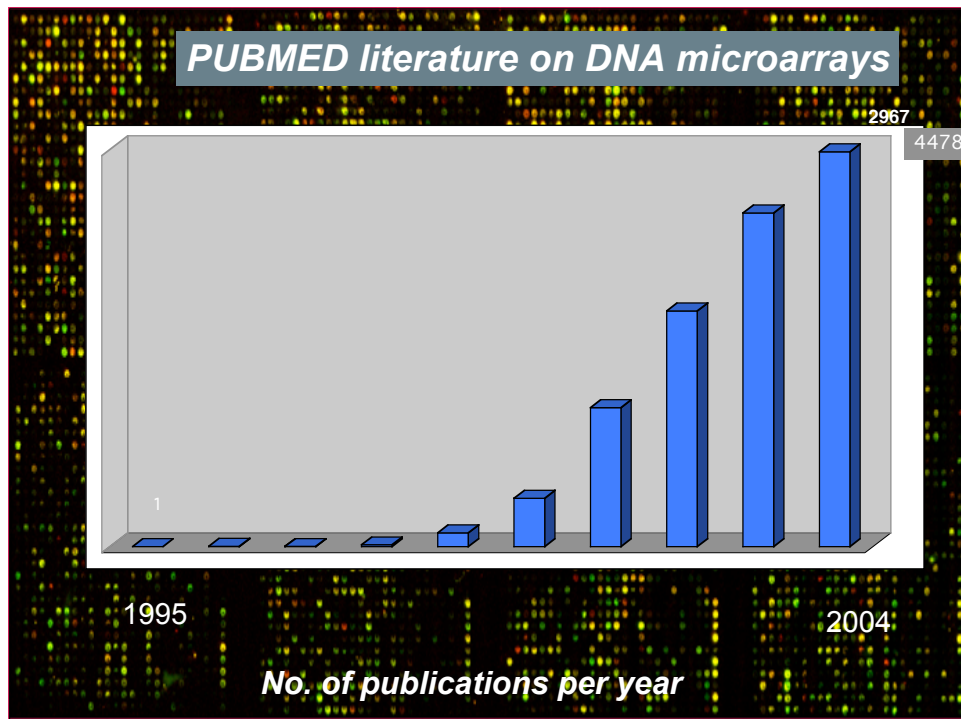
GENE EXPRESSION

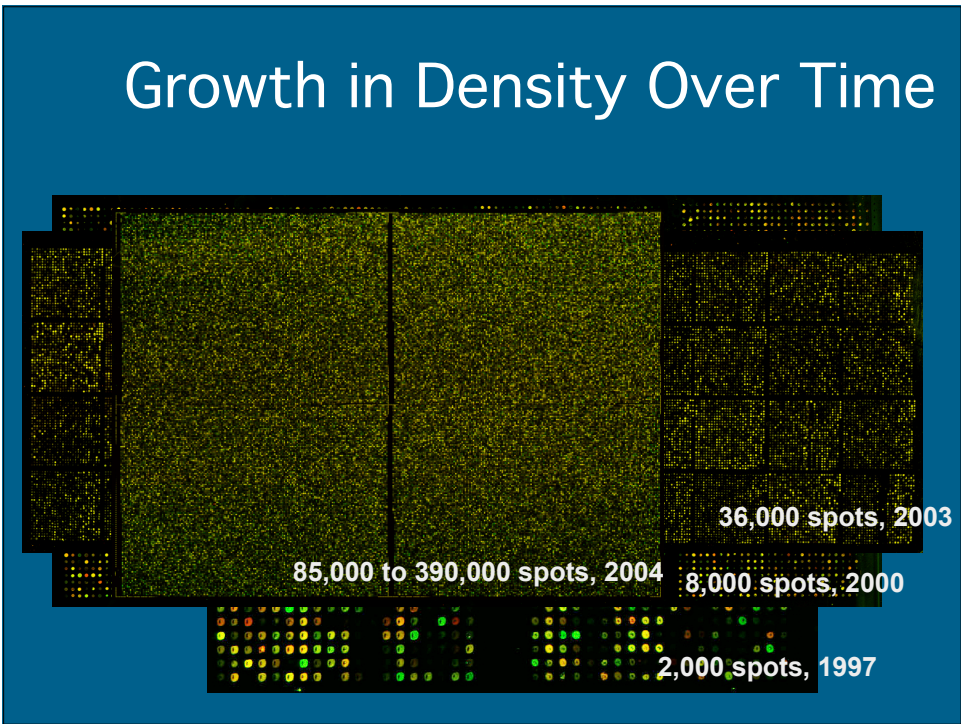
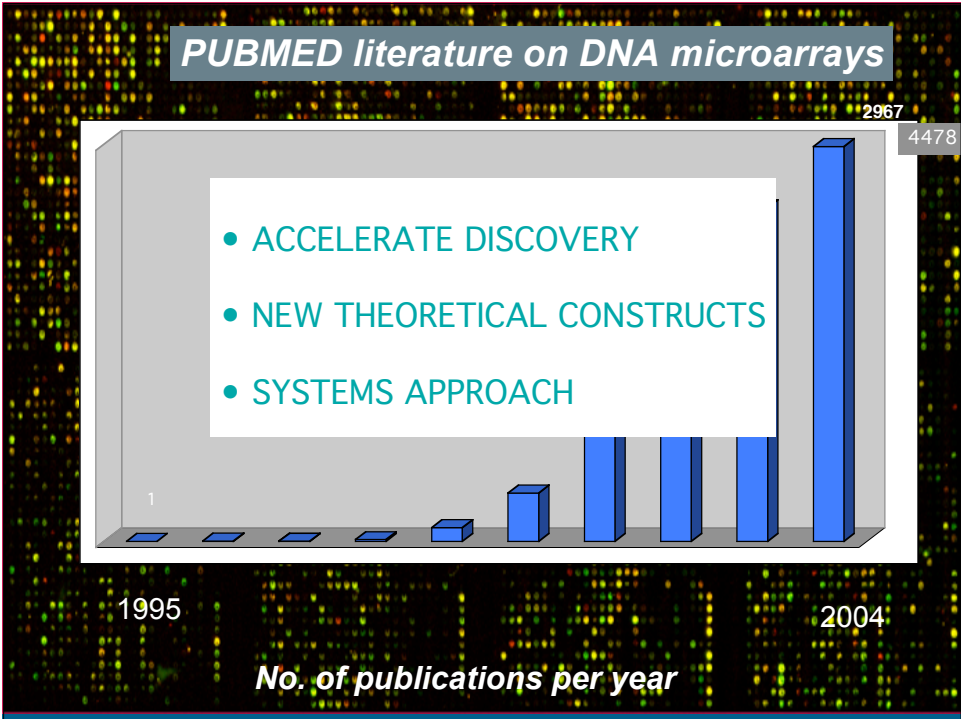
GENE VARIATION

GENE FUNCTION

MICROARRAYS PROVIDE A TOOL FOR WHOLE GENOME ANALYSIS

PRIMARY IMPACT:
ACCELERATED DISCOVERY AND
HYPOTHESIS GENERATION





MICROARRAY TERMINOLOGY

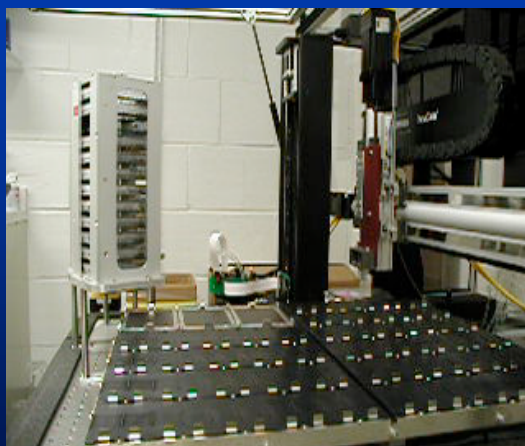
- **Feature--an array element**
- **Probe--a feature corresponding to a defined sequence**
- **Target--a pool of nucleic acids of unknown sequence**

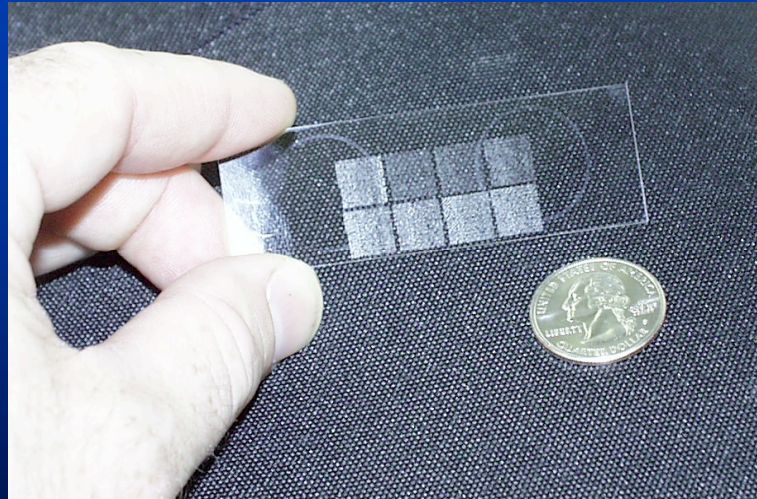
POSSIBLE ARRAY FEATURES

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

Microarray Manufacture

- Printing

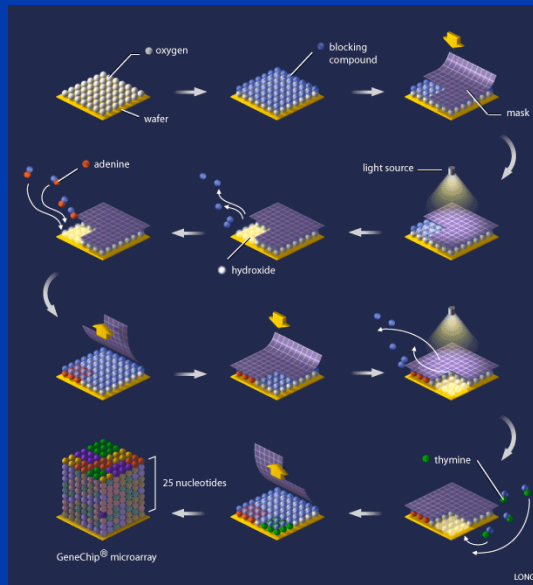




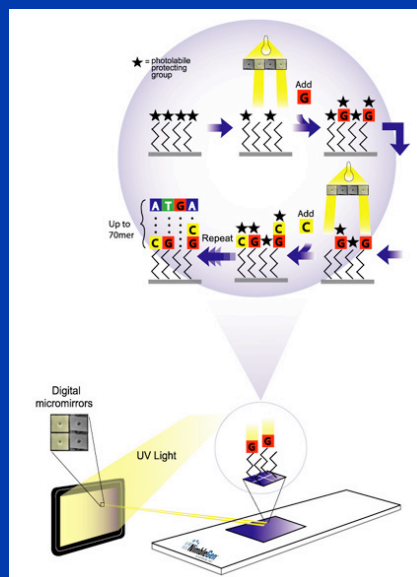
Microarray Manufacture

- **Printing**
- **Synthesis *in situ***
 - light directed
 - mechanically directed

LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS



LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS



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**

Building Microarrays

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

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.

Laboratory Essentials

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

DNA Microarray Applications

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

DNA Microarray Applications

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

DNA Microarray Applications

- Resequencing

Mutations

Polymorphisms

SINGLE NUCLEOTIDE
POLYMORPHISM

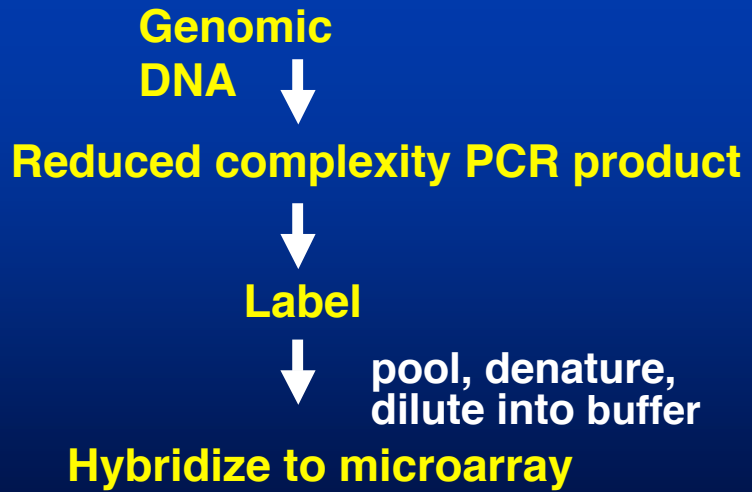
AGGTTACCAGTA

AGGTTGCCAGTA

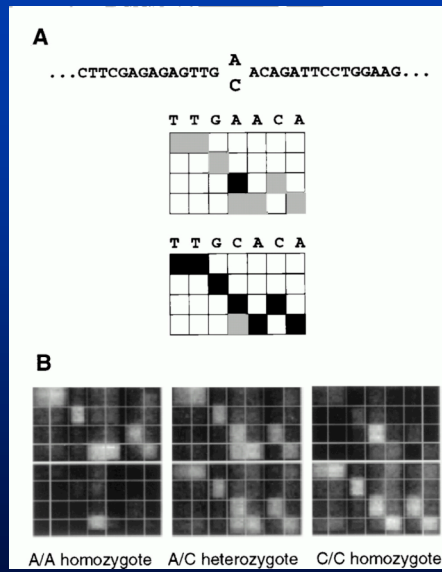
OCCUR ABOUT 1: 1250 BASES

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

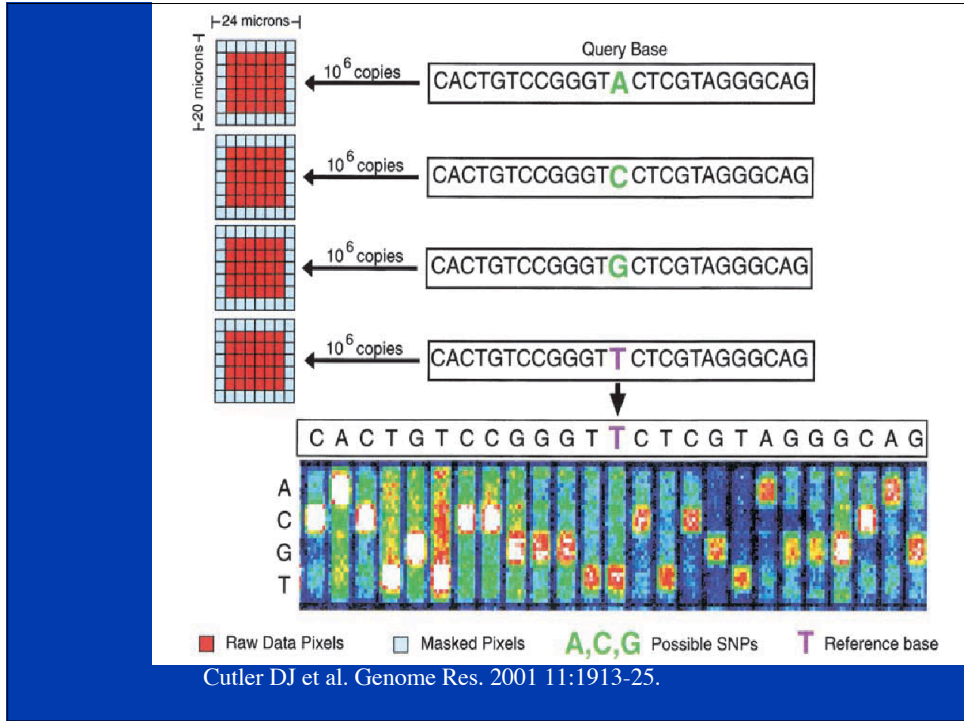
LABELLING SNPs



SNP CHIP*



*Wang et al.
 Science 280:1077
 1998



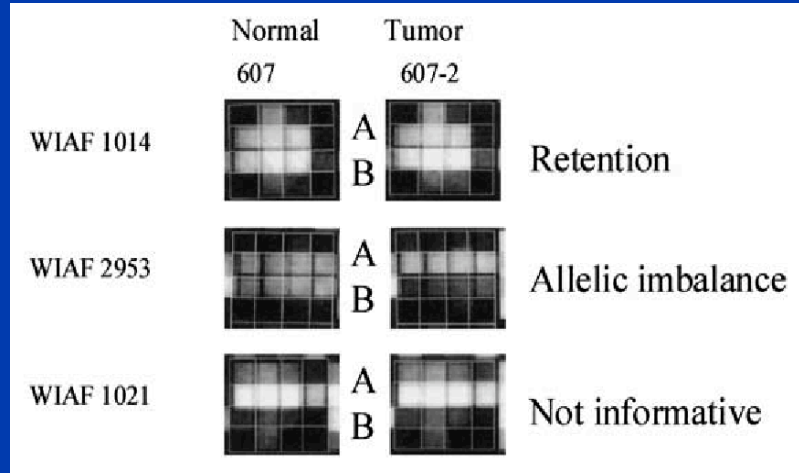
ACCURACY OF SNP CHIP

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.

SNP CHIP FOR ALLELIC IMBALANCE



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

SNP CHIPS

HAVE ACHIEVED HIGH DENSITY

1,586,383 SNPS

HINDS ET AL. SCIENCE 307:1072 (2005)

COMMERCIAL CHIPS AVAILABLE: 500,000 SNPS

WILL INCREASE

CHOICE OF TECHNOLOGY PLATFORMS

VIABLE OPTION FOR:
GENOTYPING.
CANCER ALLELIC IMBALANCE.

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

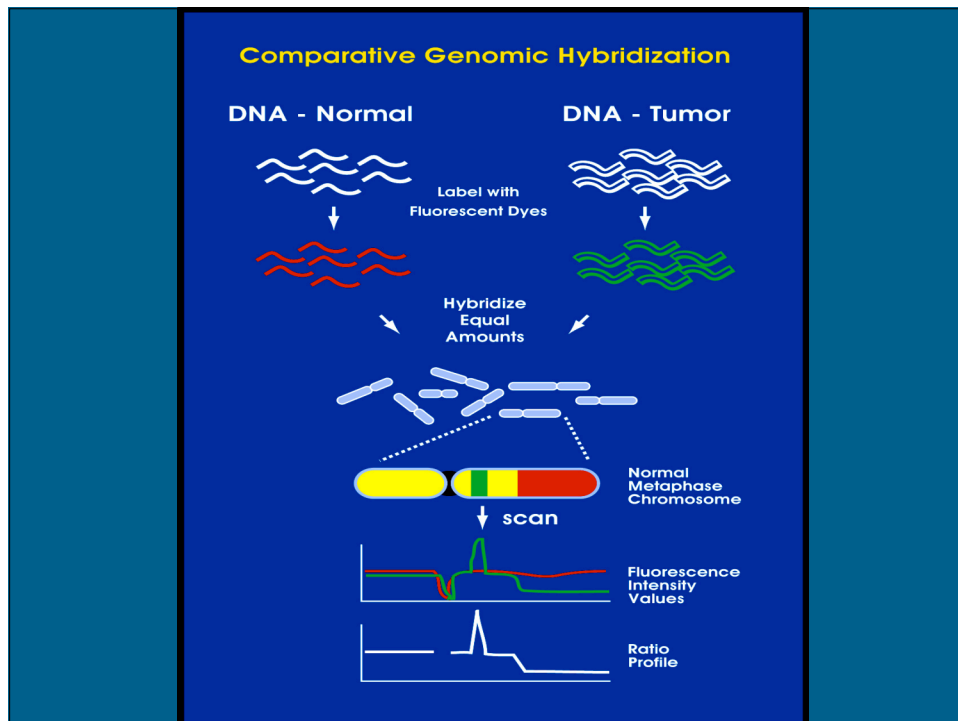
DNA Microarray Applications

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

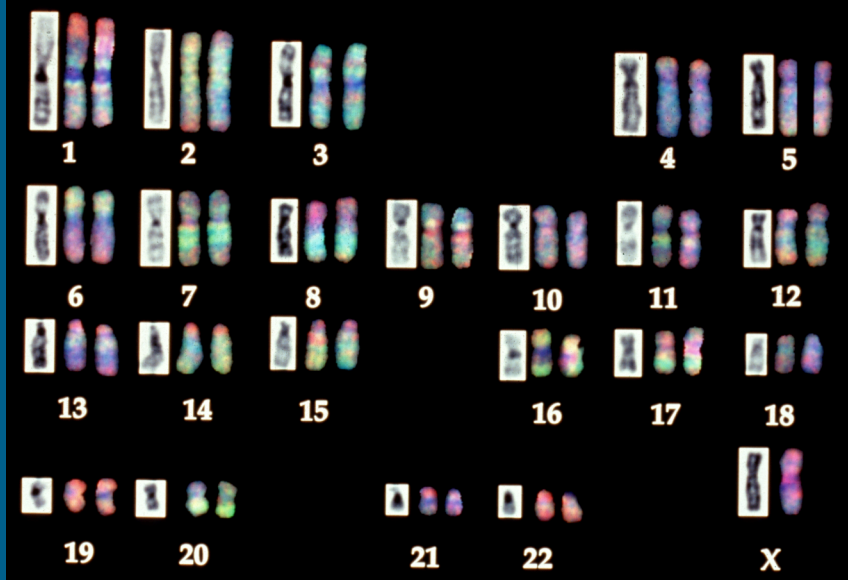
COMPARATIVE GENOMIC HYBRIDIZATION

- 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 and copy number polymorphisms.

Cancer Genetics Branch

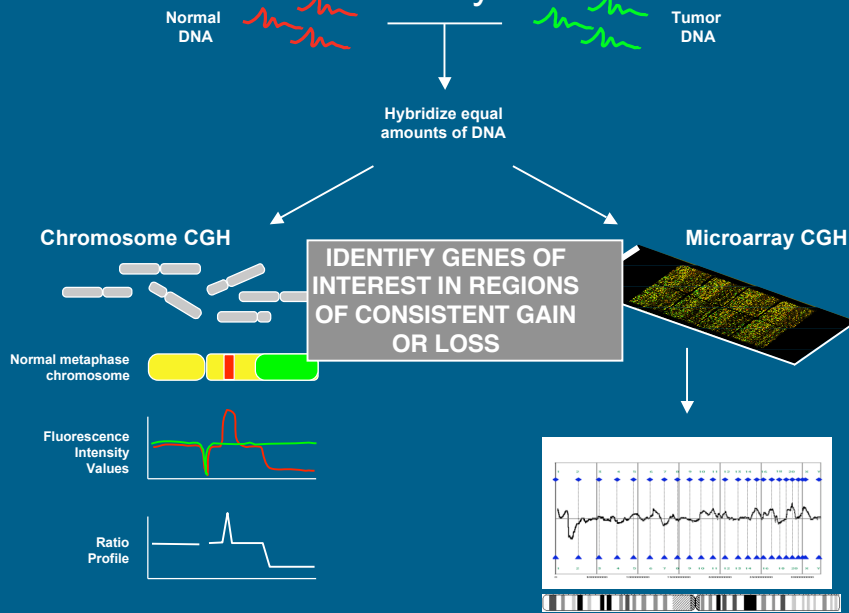


COMPARATIVE GENOMIC HYBRIDIZATION



Cancer Genetics Branch

Comparative Genomic Hybridization (CGH) arrays



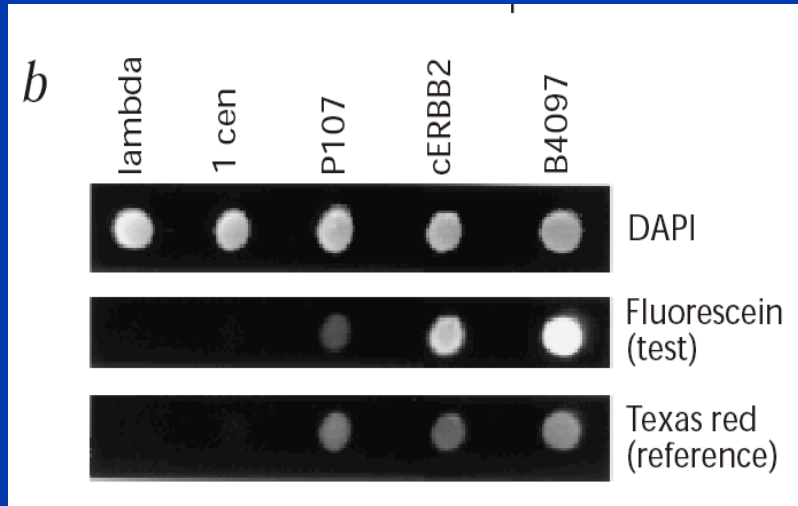
PLATFORMS FOR ARRAY BASED COMPARATIVE GENOMIC HYBRIDIZATION (CGH)

- BACs
- cDNAs
- Oligonucleotides

ARRAY CGH

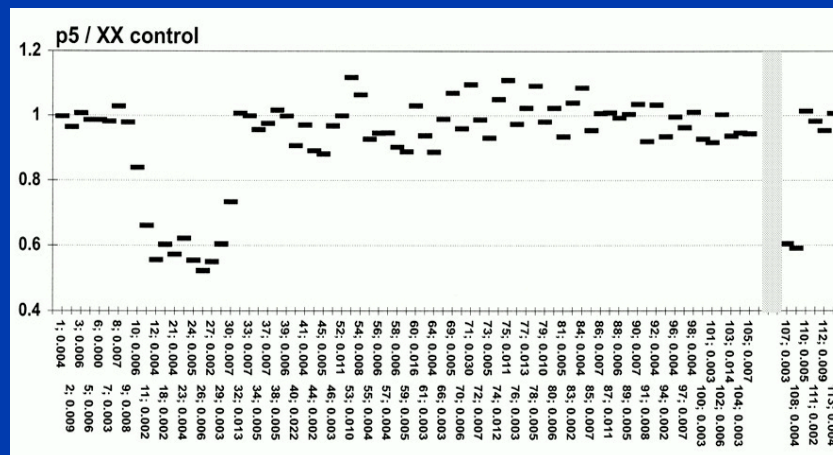
- HIGH RESOLUTION.
- SIMPLIFIED IMAGE ANALYSIS.
- HIGH THROUGHPUT.
- OLIGO STRATEGY ALLOWS GENOME
BASED DESIGN.

CGH BAC ARRAYS

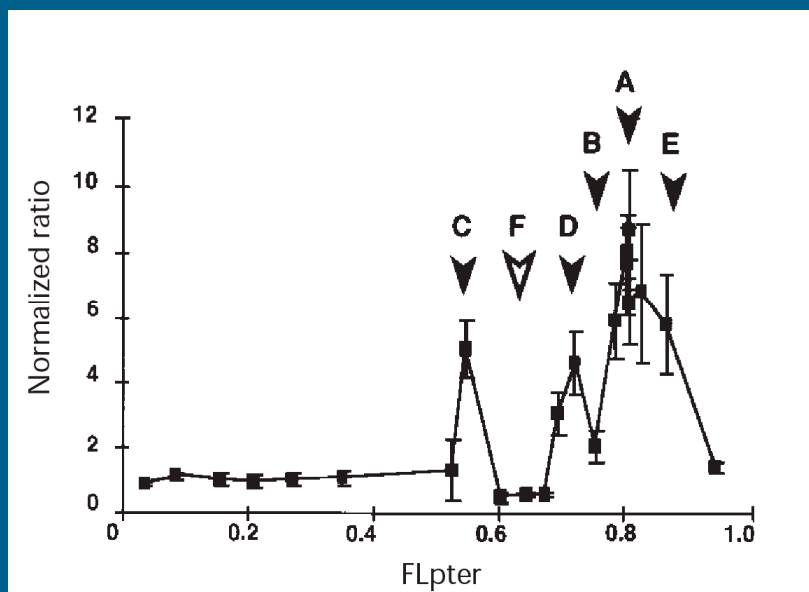


Pinkel D et al., Nature Genetics 20, 207 - 211 ,1998.

CGH BAC ARRAYS

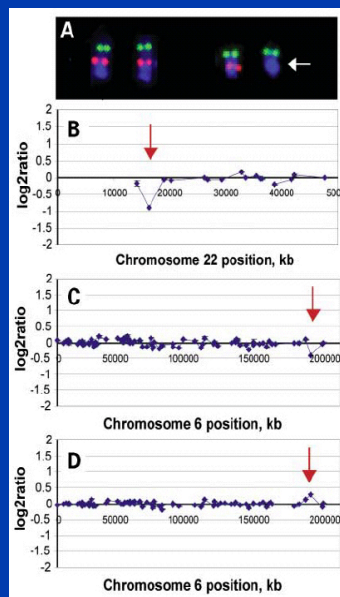


Bruder CE et al., Hum Mol Genet. 2001;10:271-82.



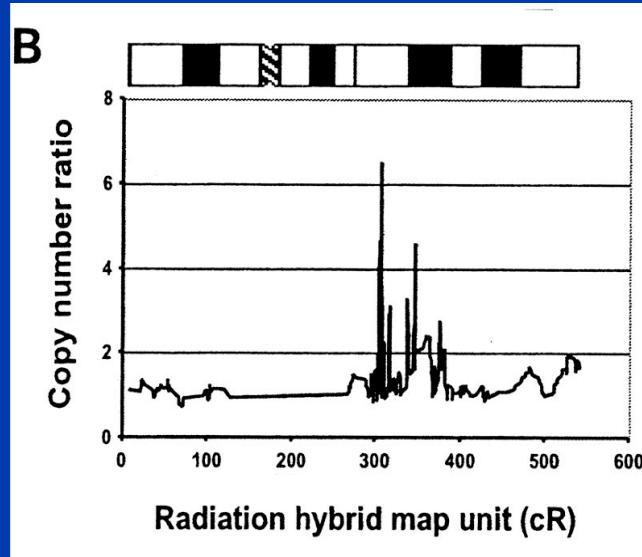
Pinkel et al. Nat Gen 20:207

CGH BAC ARRAYS



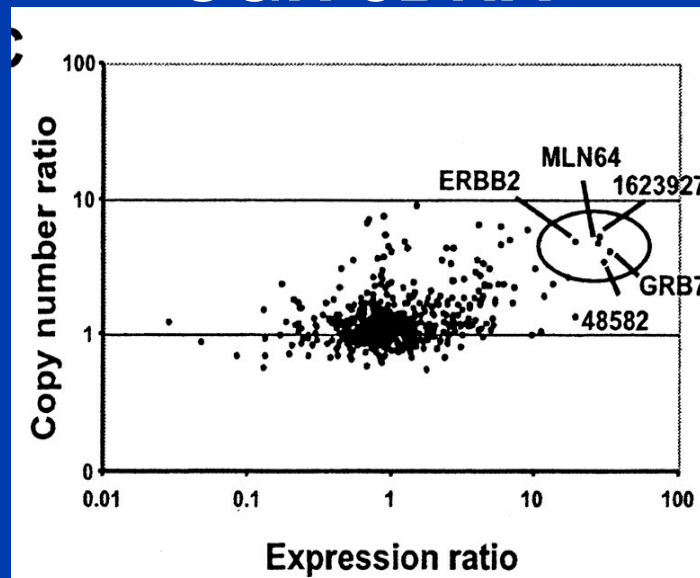
Albertson and Pinkel Hum Mol Genet 2003 12:145

CGH cDNA



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

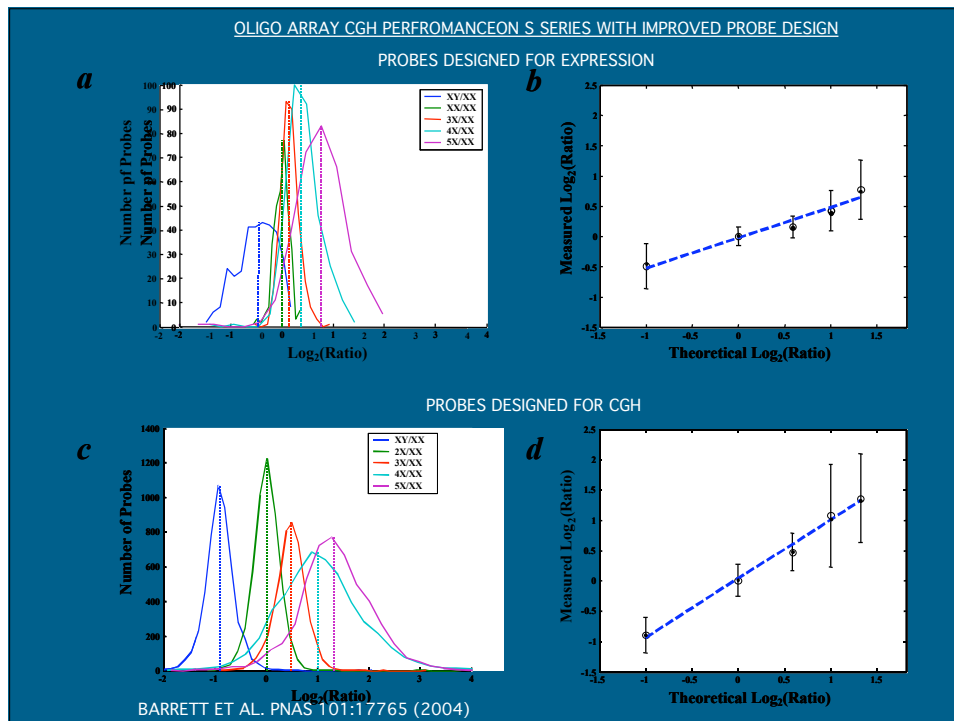
CGH cDNA

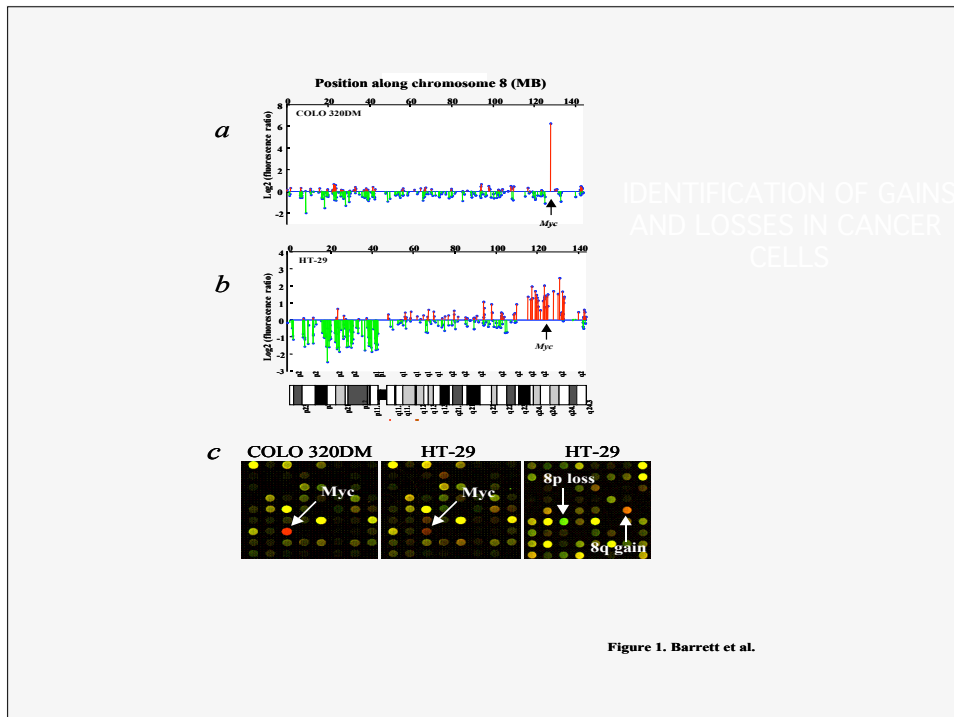
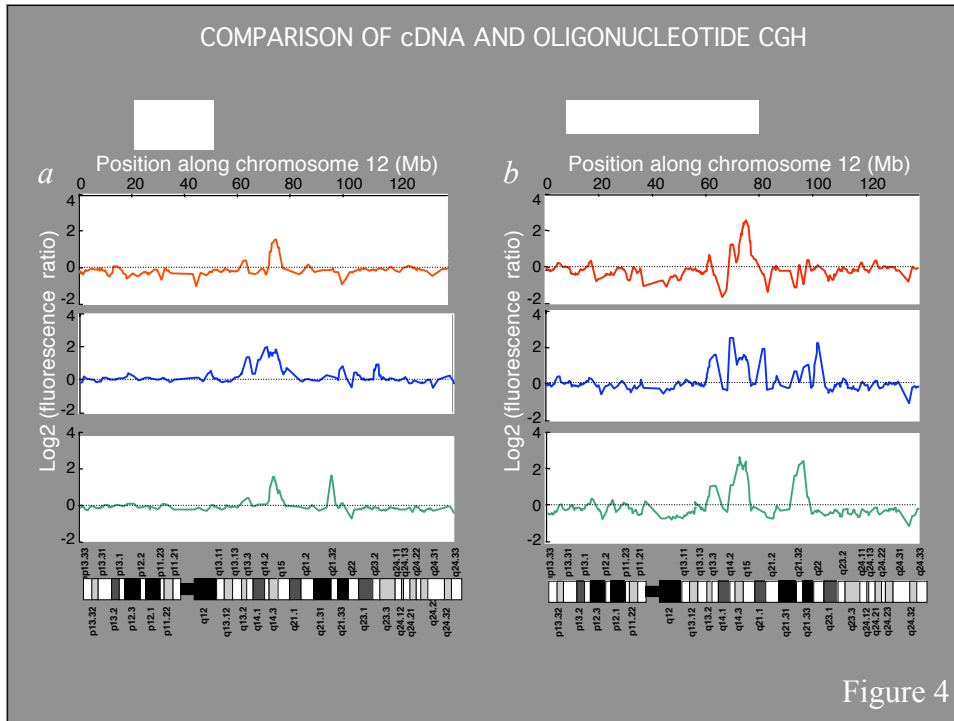


Kauraniemi P et al., Cancer Res. 2001 ;61:8235-40.

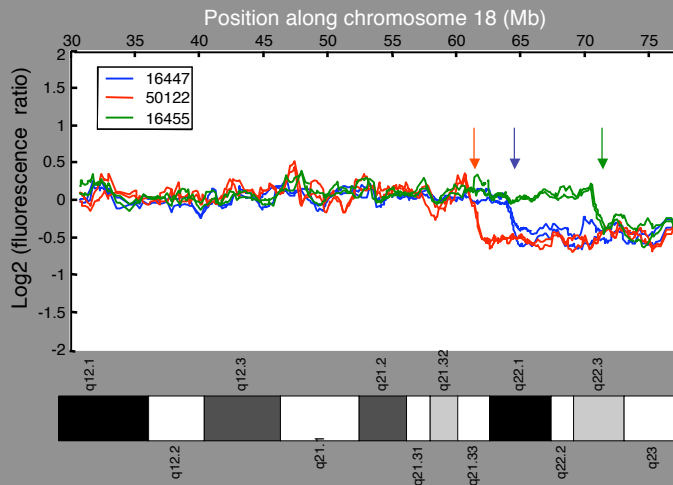
OLIGONUCLEOTIDE BASED CGH

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

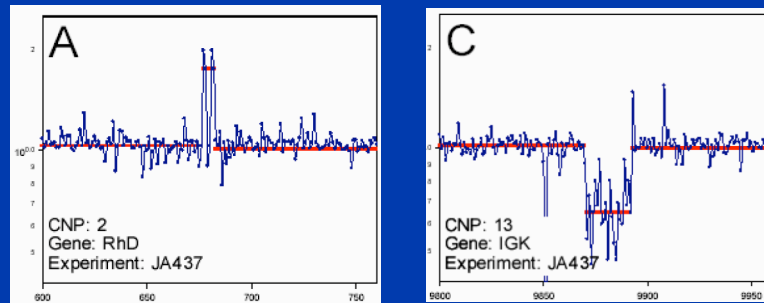




LOCATING CONSTITUTIONAL DELETIONS



HIGH DENSITY OLIGO ARRAYS FOR DETECTING COPY NUMBER POLYMORPHISM



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

DNA Microarray Applications

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

Gene Expression Profiling Technologies

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

Accessing Expression Data

- Individual Lab and Journal Sites; public databases

The screenshot shows the Gene Expression Omnibus (GEO) website. At the top, there is the NCBI logo and the GEO logo with the text "Gene Expression Omnibus". Below the logo, there is a navigation bar with links for "Handout", "NAR 2005 Paper", "NAR 2002 Paper", "FAQ", "MIAME", and "Email GEO".

The main content area is divided into several sections:

- Public data:** A table showing statistics: GPL Platforms: 1192, GSM Samples: 35816, GSE Series: 1816, Total: 38624, dated Apr 08 2005.
- Site contents:** A list of links including Overview, FAQ, Web deposit guide, Batch deposit guide, SOFT examples, Linking & citing, Journal citations, Handout (pdf), DataSet clusters, GEO announce list, Data disclaimer, and GEO staff.
- Query & Browse:** A section with links for DataSet browser, Repository browser, SAGEmap, FTP site, GEO Profiles, GEO DataSets, Deposit & Update, Web deposit, Direct deposit, and New account.
- BROWSE:** A tree view showing "GEO accessions" which branches into "Platforms", "Samples", and "Series".
- QUERY:** A section with input fields for "GEO accession", "Gene profiles", and "DataSets", each with a "GO" button. There is also a "GEO BLAST" button.
- SUBMIT:** A section with buttons for "Direct deposit / update", "Web deposit / update", and "Create new account".
- Retrieve GEO accession:** A search bar with "Scope: Self", "In: HTML", and "view: Quick".
- Depositors only:** A login section with "User:" and "Password:" fields, a "Login" button, and "Unlogged" text.

At the bottom, there is a footer with links for "NLM", "NIH", "GEO Help", "NCBI Help", "Disclaimer", and "Section 508".

GEO

Accessing Expression Data

The screenshot shows the EMBL-EBI ArrayExpress website. At the top, there is the EMBL-EBI logo and the text "European Bioinformatics Institute". Below the logo, there is a navigation bar with links for "EBI Home", "About EBI", "Research", "Services", "Toolbox", "Databases", "Downloads", and "Submissions".

The main content area is divided into several sections:

- ArrayExpress at the EBI:** A section with a description: "ArrayExpress is a public repository for microarray data, which is aimed at storing well annotated data in accordance with MIAME recommendations." Below this is a "Current Content Overview" table:

Experiments:	66	View
Arrays:	89	View
Protocols:	459	View
Hybridizations:	142	
- Latest News:** A section with a "New MIAME Express Release 1.5" dated 08/10/2003. Below this is a link to "MAME Express Package Release 1.5" and a link to "Mapping the MAGE-OM to data within the Stanford Microarray Database" dated 03/05/2003.
- Announcement:** A section with text: "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."
- Navigation:** A sidebar on the left with links for "ArrayExpress Home", "Browse Database >>", "Query Database >>", "Login To Database >>", "Submissions", "Help & Documentation", "Microarray Standards", "Schema", "Implementation", and "EBI Microarray Home >>".
- Footer:** A section with text: "For comments, questions or issues about ArrayExpress, please contact us at arrayexpress@ebi.ac.uk." Below this is a link for "Supplementary information".

Publishing Expression Data

- MIAME standard

Minimum Information about a Microarray Experiment

Format required by many journals

STRATEGIES FOR SIGNAL GENERATION FROM mRNA

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

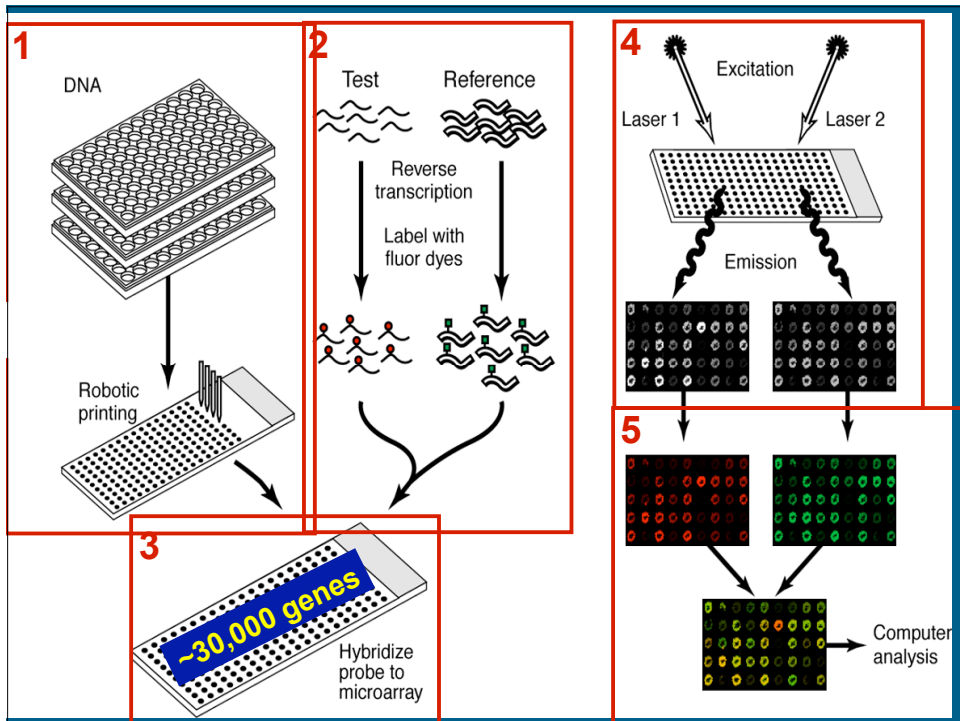
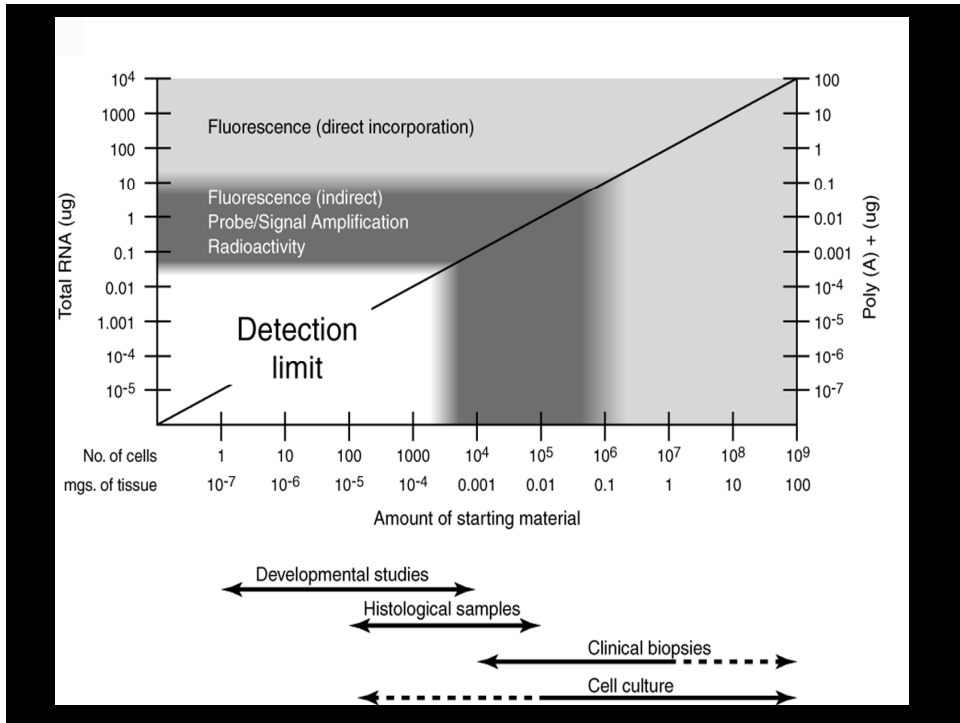
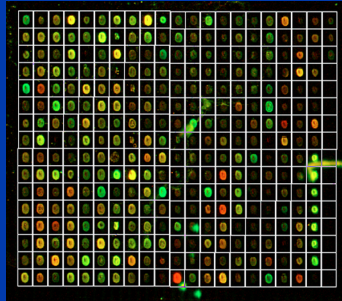


Image Analysis: DeArray

Grid Overlay



Target detection

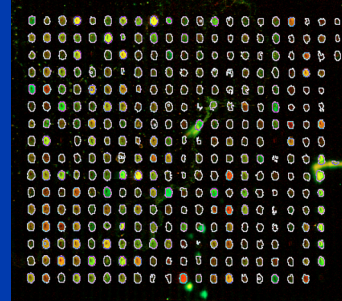
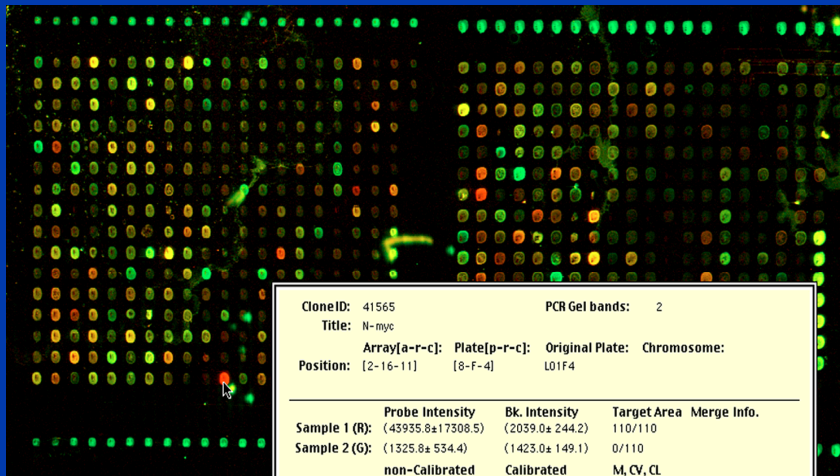
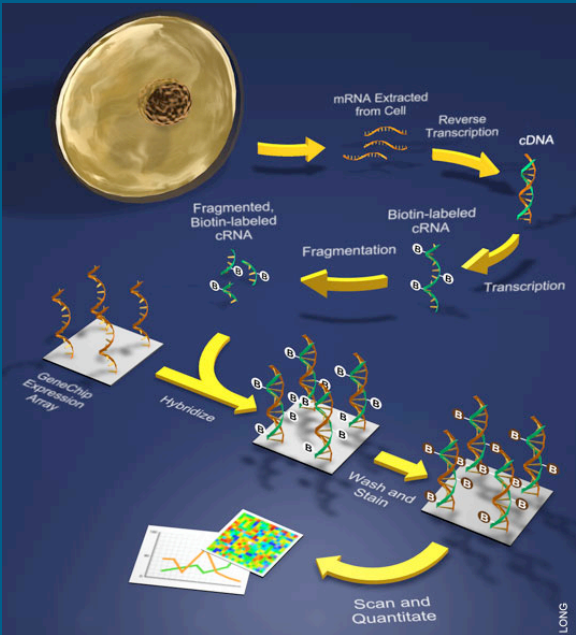
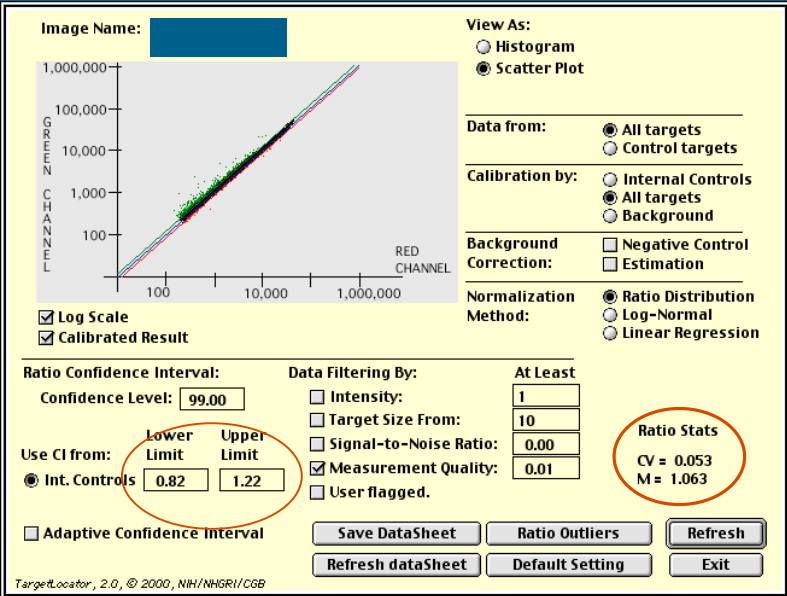


Image Analysis: DeArray



TargetLocator, 1.3, © 1997, NIH/NCRR/LCS

DATA QUALITY IS CRITICAL



ONE COLOR
 HYBRIDIZATION
 ON AN OLIGO
 ARRAY

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.

APPLICATIONS OF EXPRESSION ARRAYS

•Expression profiling

Power arises from increasing sample number

•Direct comparisons (Induction)

Biological system critical

•Genome Annotation

A RECURRING PROBLEM

Disease Genes

Transcription factors

Hormones/growth factors

Drugs

Toxins

Infectious agents

Physical agents



?????

Downstream Genes

•Direct targets

•Indirect targets

EXPRESSION DATA ANALYSIS

- Large amount of data
- Requires visualization and analysis tools

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**

Unsupervised Clustering

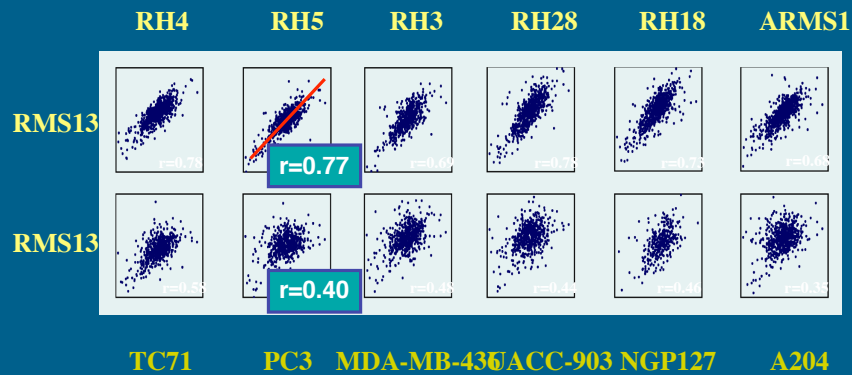
How do genes and samples organize into groups?

Powerful method of data display.

Does not prove the validity of groups.

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

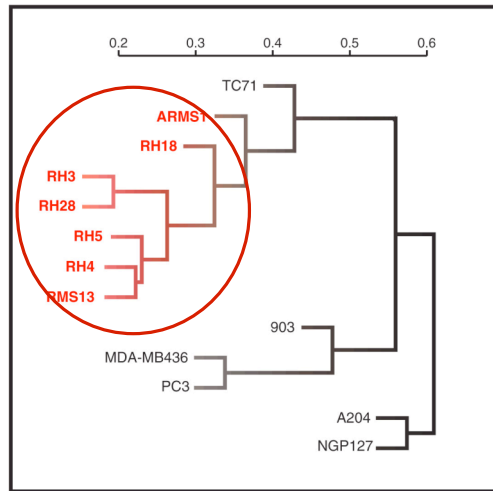
UNSUPERVISED CLUSTERING IS BASED ON A GLOBAL SIMILARITY METRIC



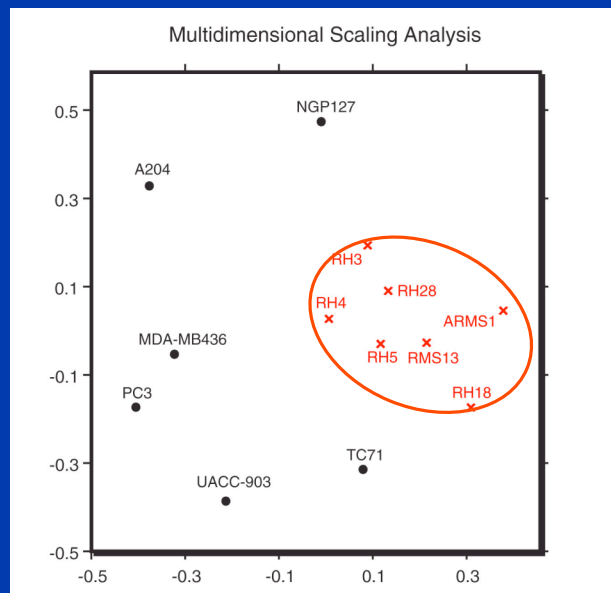
Matrix of Pearson Correlation Coefficients Distance Map

	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	0.615	0.307	0.39	0.498	0.426	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									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													0.662

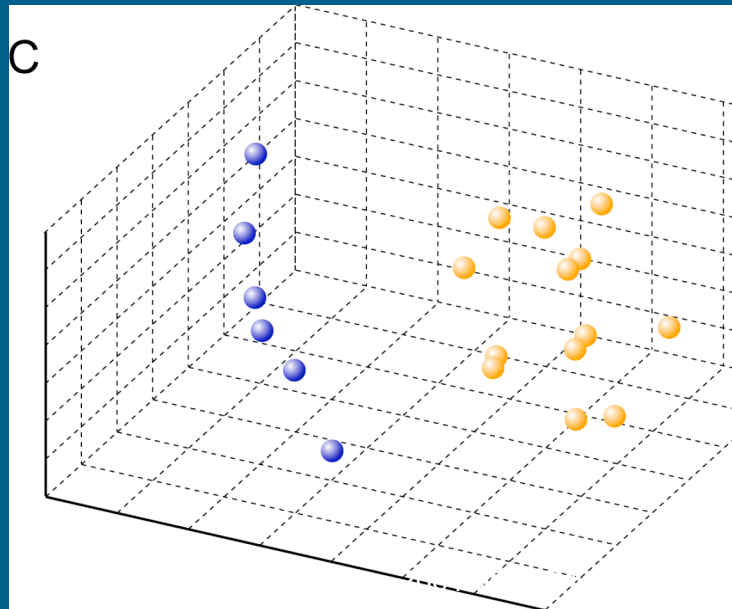
Hierarchical Clustering Dendrogram



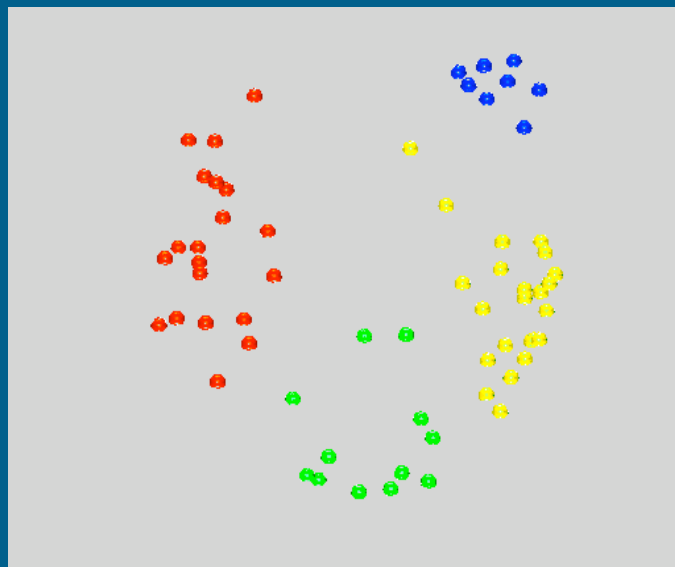
DATA DISPLAY BY MULTIDIMENSIONAL SCALING



MDS PLOT

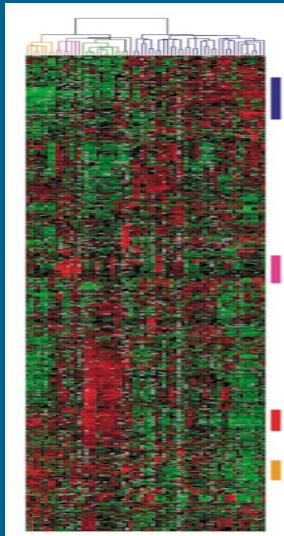


MULTIDIMENSIONAL SCALING



-Lymphoma
RMS
NBL
-EWS

CLUSTERING GENES AND SAMPLES

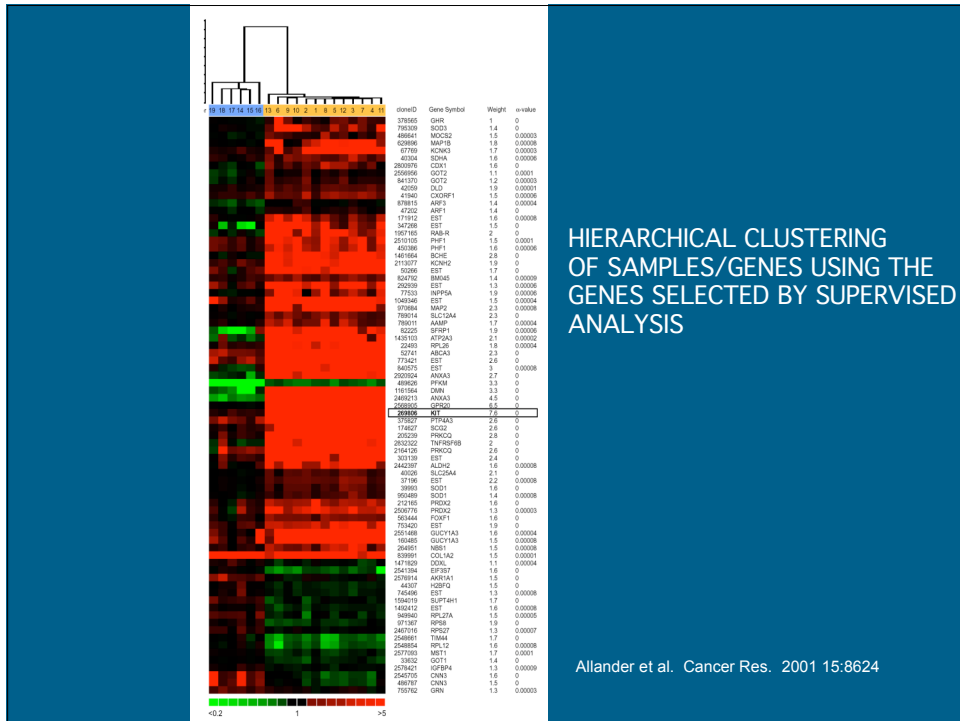


Perou et al. Nature 2000 406:747

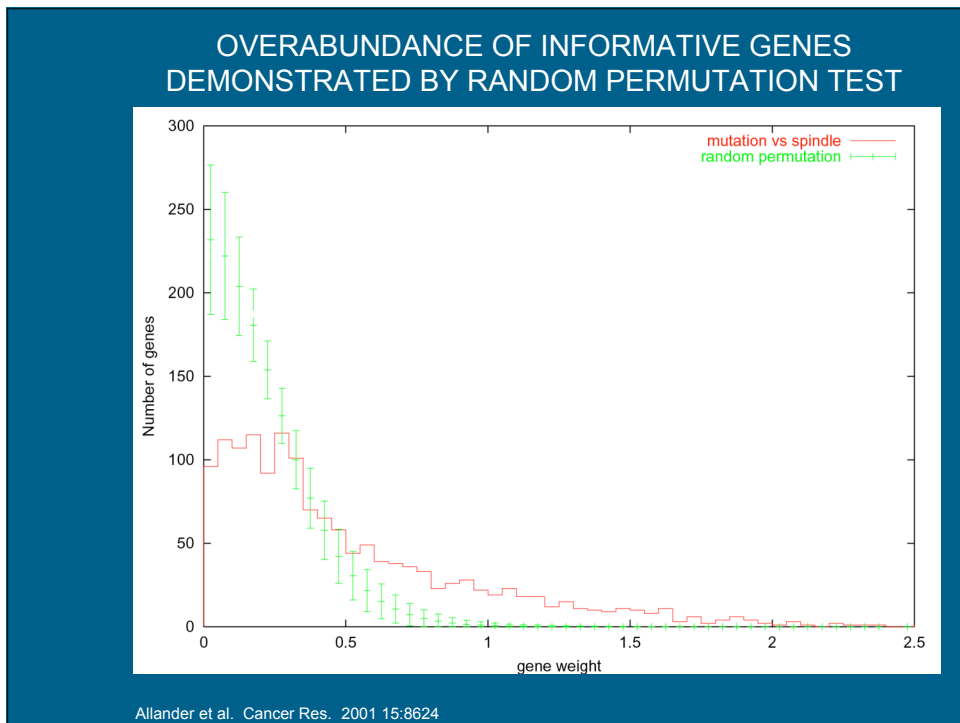
Supervised Clustering

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.



Allander et al. Cancer Res. 2001 15:8624



GENOMICS FROM BENCH TO BEDSIDE

WHOLE GENOME



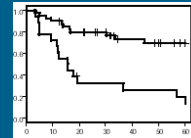
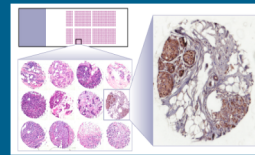
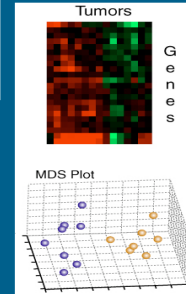
GENE SELECTION



GENE VALIDATION



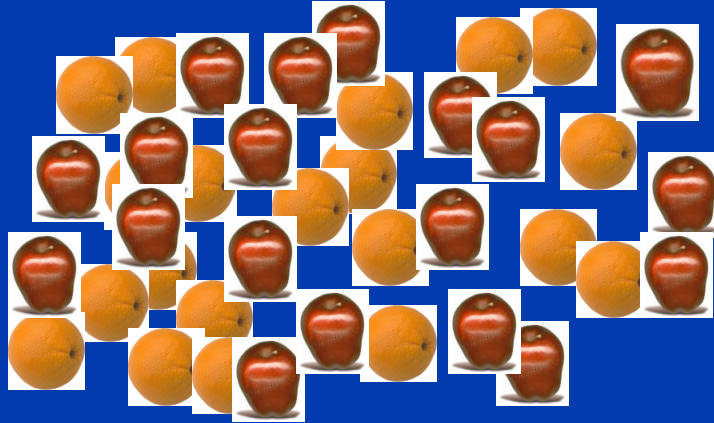
ASSAY DEVELOPMENT



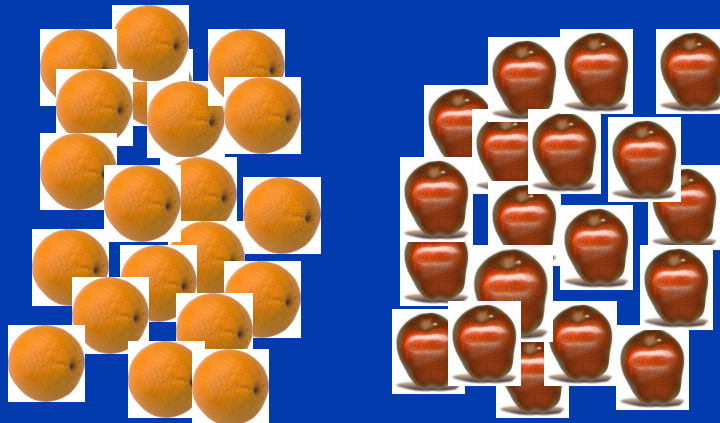
**SIGNAL STRENGTH VARIES IN
TISSUE PROFILING EXPERIMENTS**

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

CONSIDER A SAMPLE SET



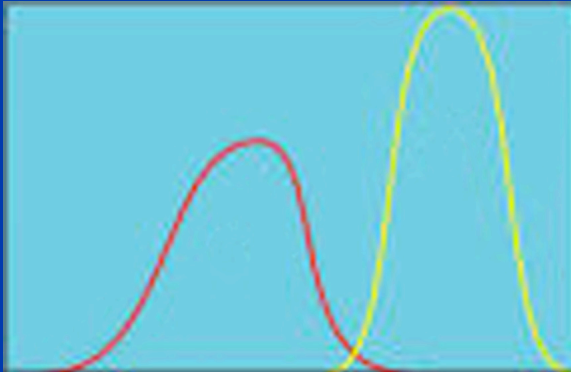
CONSIDER A SAMPLE SET



THESE ARE EASY TO DISTINGUISH BY ONE MEASUREMENT PER INDIVIDUAL.

CONSIDER A SAMPLE SET

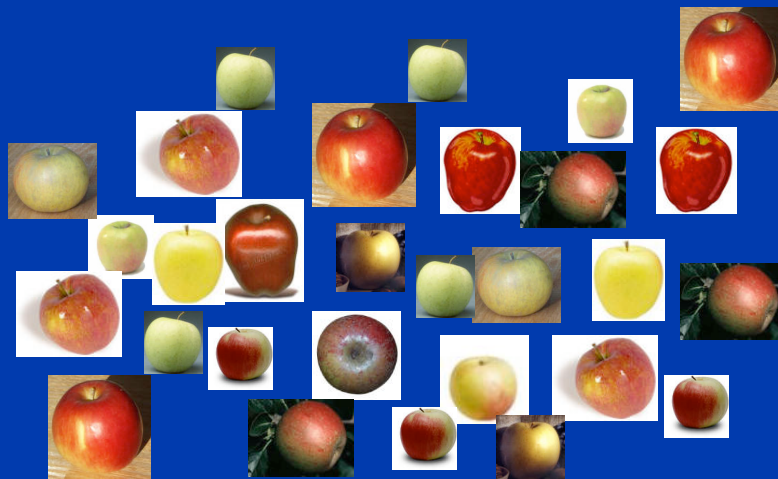
TUMORS



EXPRESSION LEVEL
(HIGHLY INFORMATIVE GENE)

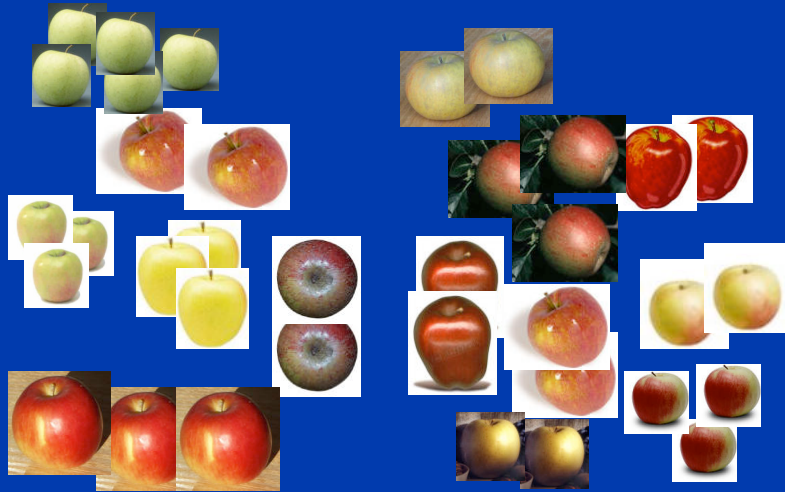
THESE ARE EASY TO DISTINGUISH BY
ONE MEASUREMENT PER INDIVIDUAL.

CONSIDER A SAMPLE SET



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

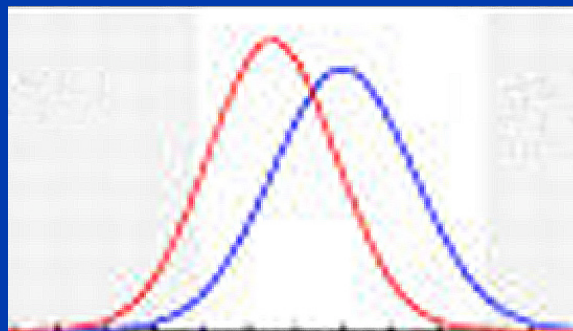
CONSIDER A SAMPLE SET



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

CONSIDER A SAMPLE SET

TUMORS



EXPRESSION LEVEL
(POORLY INFORMATIVE GENE)

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

WE CAN TELL APPLES
FROM ORANGES.

CAN WE DISTINGUISH
DIFFERENT KINDS OF APPLES?

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.

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 .

RISK OF OVERFITTING IN CLINICAL STUDIES WITH SMALL SAMPLE SETS

**NEED INDEPENDENT VALIDATION
SETS.**

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.

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.



WHAT SHOULD YOU LOOK FOR IN A CLINICAL MICROARRAY STUDY?

ARE MICROARRAY TECHNOLOGIES READY TO BE IMPLEMENTED IN CLINICAL PRACTICE?

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.

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.**

EXPRESSION PROFILING IN THE CLINIC?

PROBLEMS:

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

EXPRESSION PROFILING IN THE CLINIC?

OPTIONS:

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

EXPRESSION PROFILING IN THE CLINIC?

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

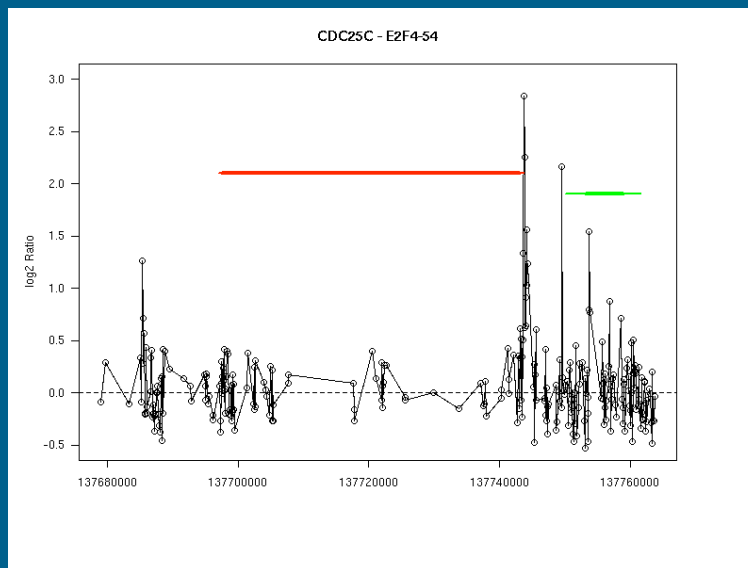
DNA Microarray Applications

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

APPLICATIONS OF TILING PATH ARRAYS

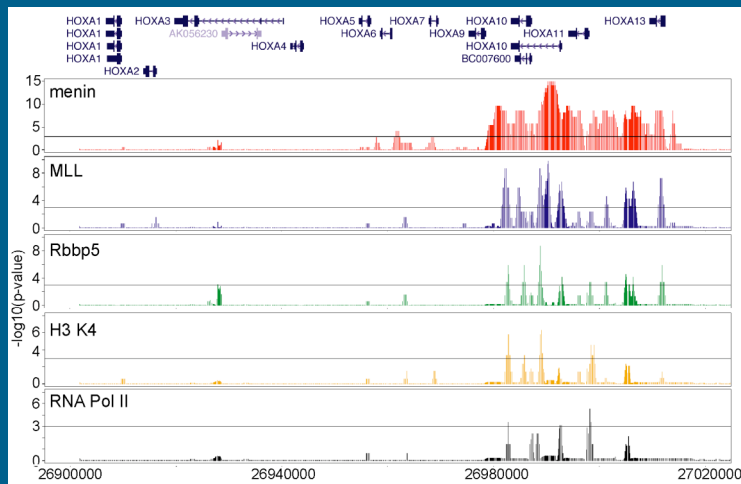
- CGH
- EXPRESSION
- CHIP CHIP
- DNase HYPERSENSITIVE SITES
- ANY ENRICHED PREPARATION OF INTERESTING SEQUENCES

E2F4 IP



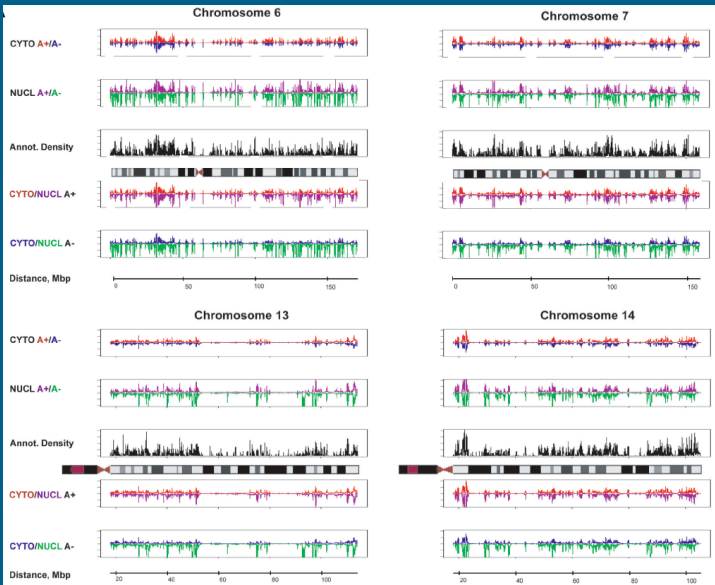
Scacheri et al. PLOS 2006

MULTIPLE IP'S IN HOXA REGION



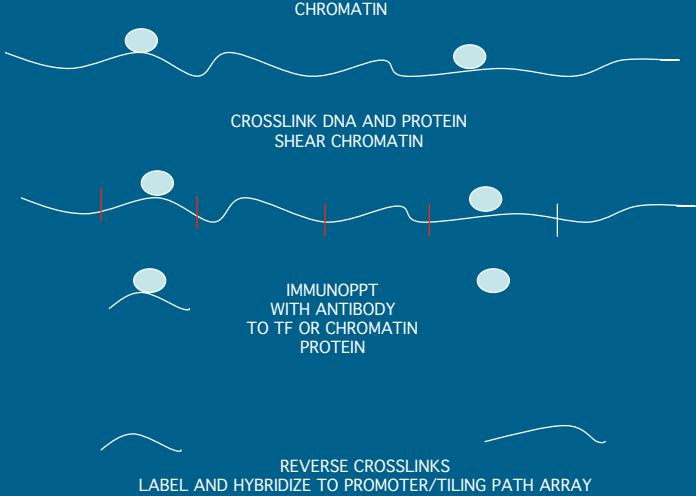
Scacheri et al. PLOS 2006

Scanning Chromosomes with Tiling Path Arrays

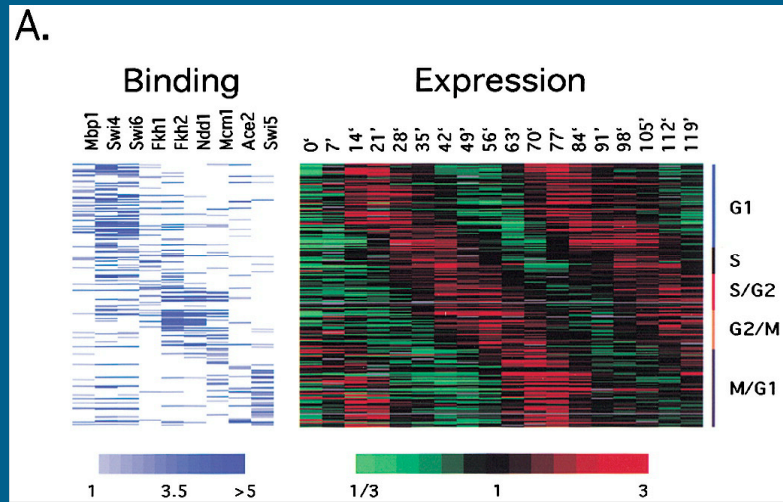


Cheng et al Science March 29, 2005

TRANSCRIPTION FACTOR LOCALIZATION ON ARRAYS

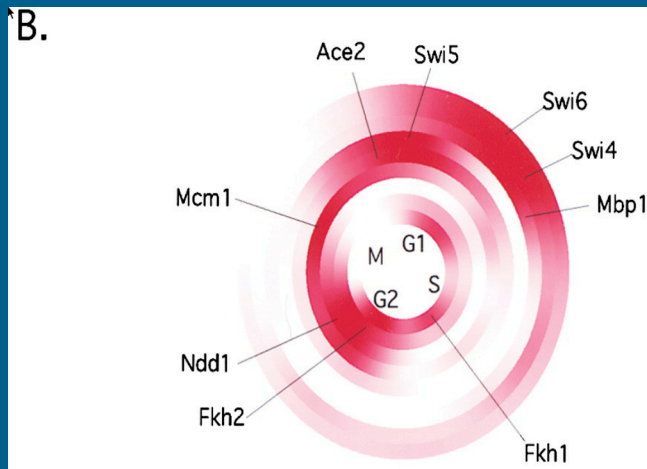


Promoter Occupancy During Yeast Cell Cycle



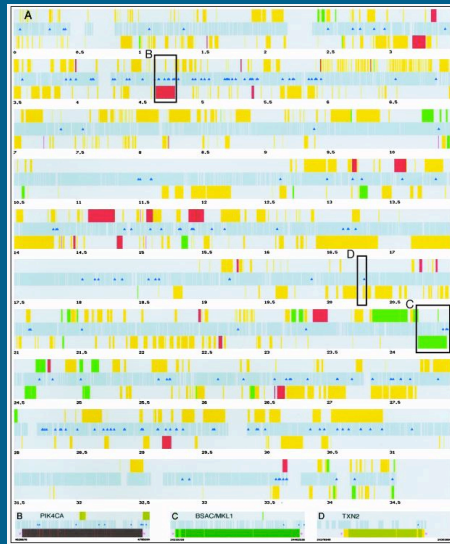
Simon I Cell. 2001 Sep 21;106(6):697-708

Promoter Occupancy During Yeast Cell Cycle



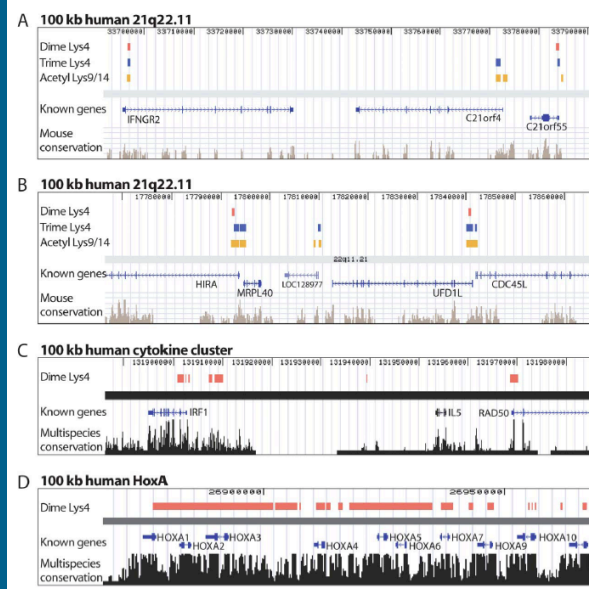
Simon I Cell. 2001 Sep 21;106(6):697-708.

NFKB Binding to Chromosome 22



Martone et al. PNAS. 2004 100:12247.

CHROMATIN MODIFICATION BY CHIP CHIP

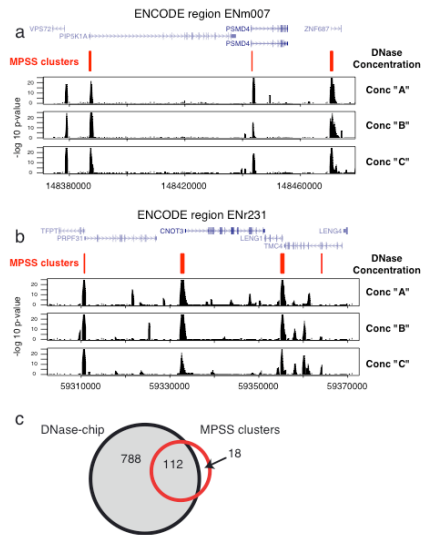


Bernstein et al. Cell 2005 120:169.

DNASE HS SITES

Crawford et al.

Figure 3



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.