

**LARGE SCALE
ANALYSIS OF GENE
EXPRESSION**

Evolution and Revolution

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

GENE EXPRESSION

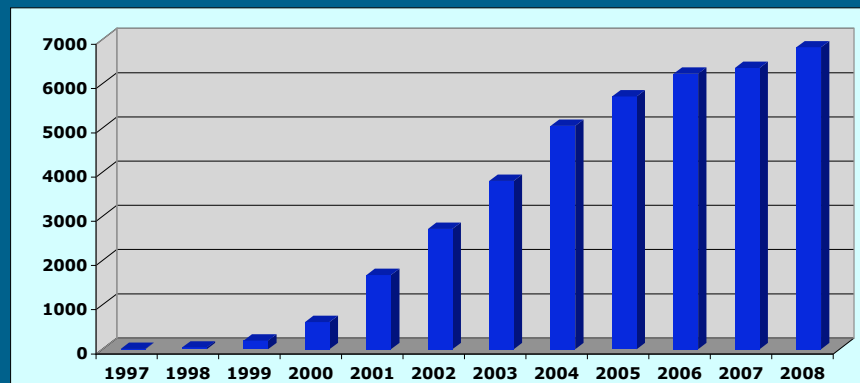
GENE VARIATION

GENE FUNCTION

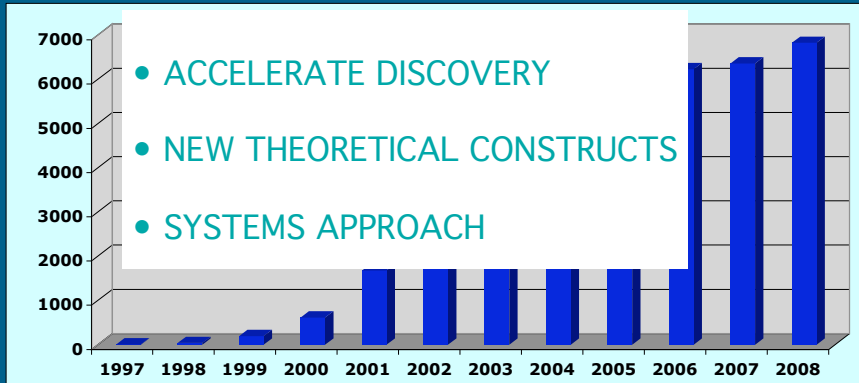
MICROARRAYS PROVIDE A TOOL FOR WHOLE GENOME ANALYSIS

**PRIMARY IMPACT:
ACCELERATED DISCOVERY AND
HYPOTHESIS GENERATION**

PUBMED CITATIONS FOR DNA MICROARRAYS

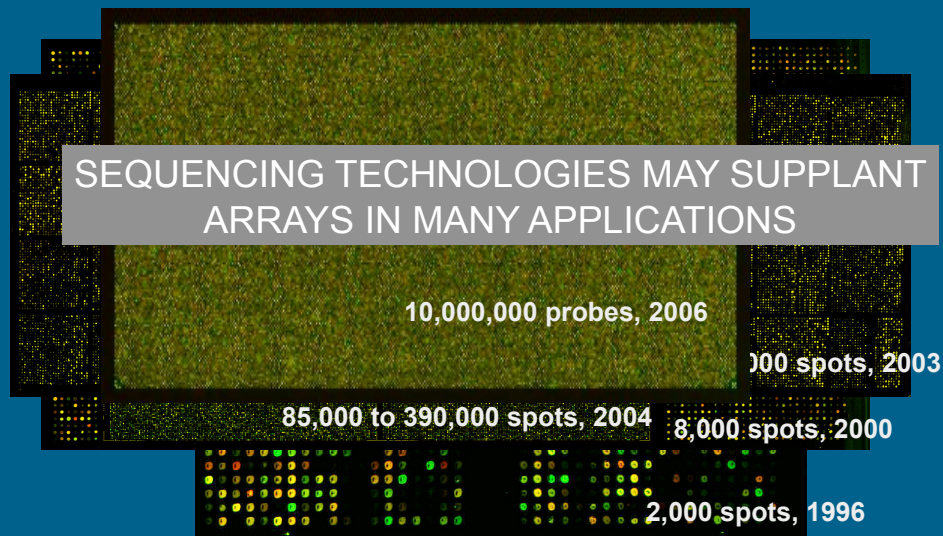


PUBMED CITATIONS FOR DNA MICROARRAYS



INCREASE IN FEATURE DENSITY

SEQUENCING TECHNOLOGIES MAY SUPPLANT ARRAYS IN MANY APPLICATIONS



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

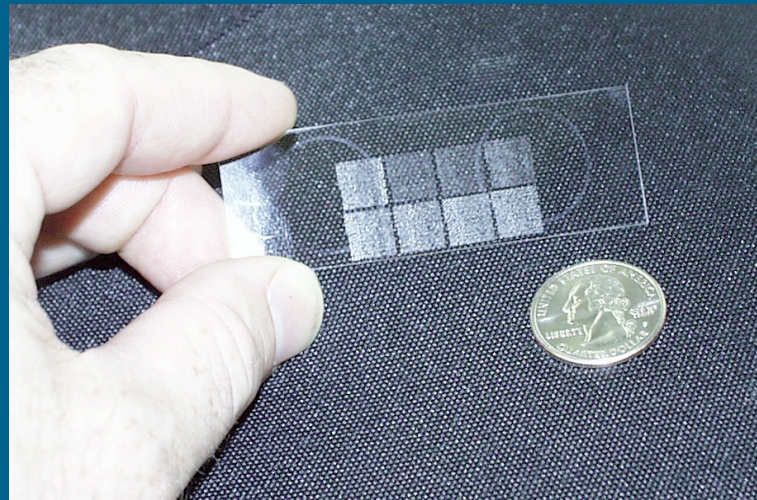
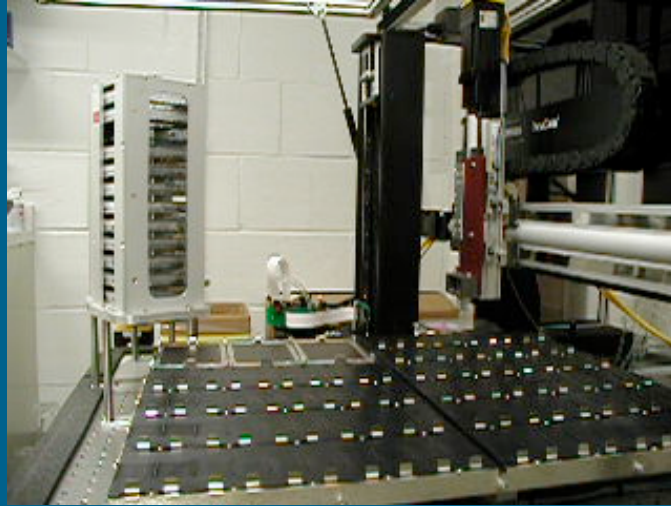
- **Synthetic Oligonucleotides**
- **PCR products from**
 Cloned DNAs
 Genomic DNA
- **Cloned DNA**

OLIGONUCLEOTIDE ARRAY DESIGN

- **Extremely flexible**
 - **3' bias**
 - **full length**
 - **exon specific**
 - **candidate transcripts**
 - **miRNAs**
- **Very high density possible**
- **Requires sequence data**

Microarray Manufacture

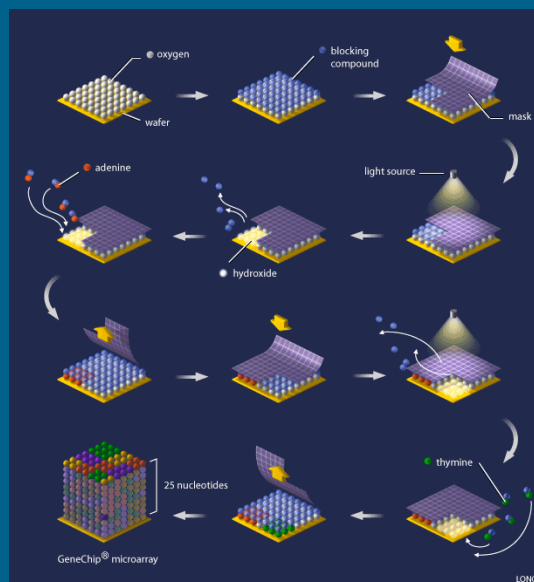
- **Printing**



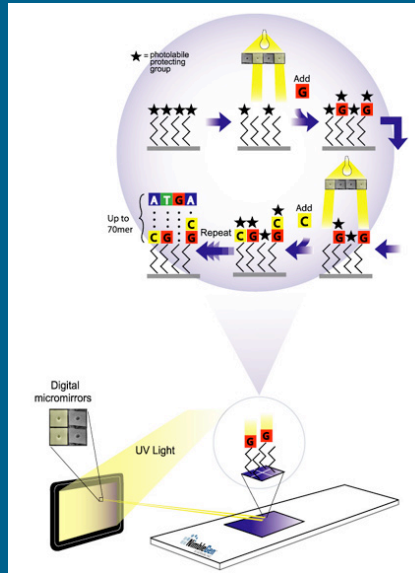
Microarray Manufacture

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

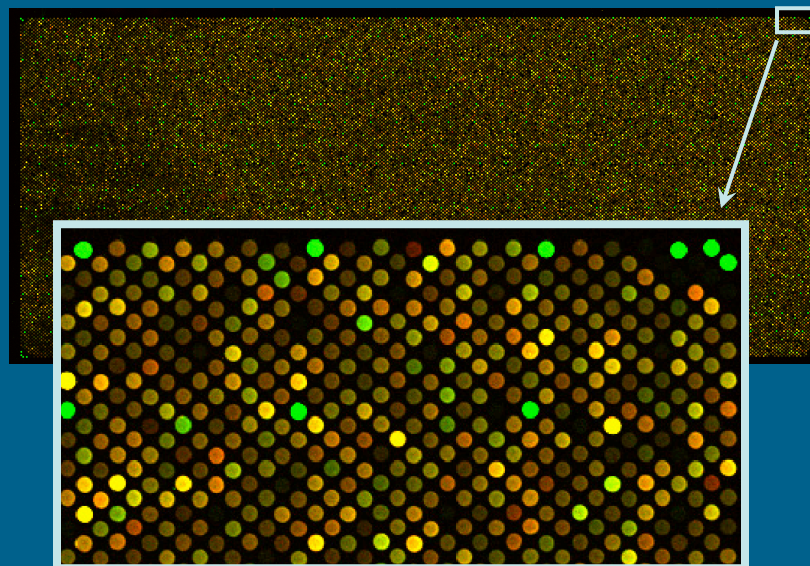
LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS

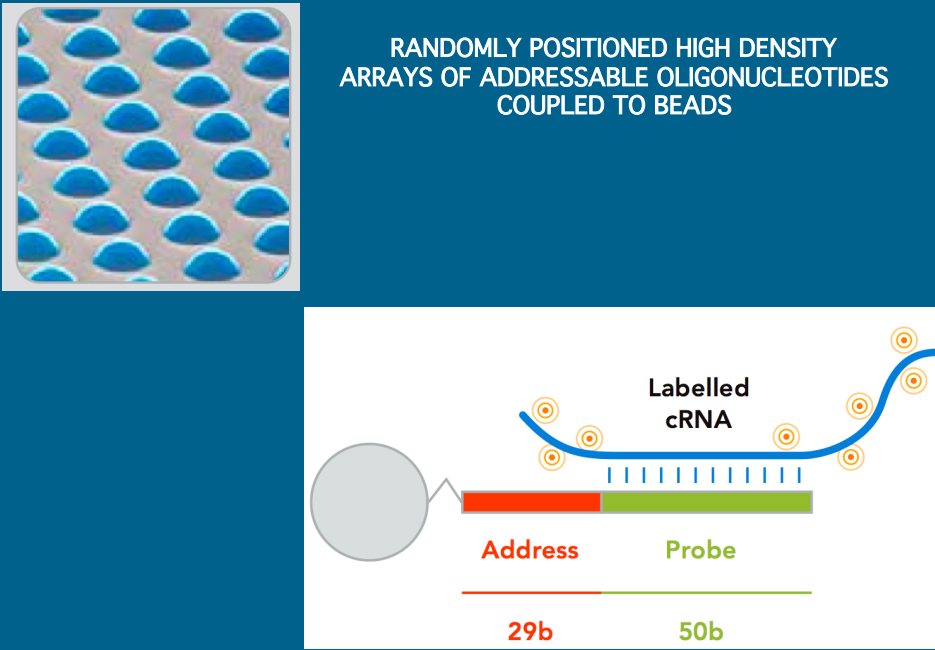


LIGHT DIRECTED OLIGONUCLEOTIDE SYNTHESIS



INK JET DIRECTED SYNTHESIS





RANDOMLY POSITIONED HIGH DENSITY
ARRAYS OF ADDRESSABLE OLIGONUCLEOTIDES
COUPLED TO BEADS

Labelled
cRNA

Address Probe

29b 50b

The image shows a microarray bead array on the left and a schematic diagram on the right. The schematic shows a grey bead attached to a red 'Address' region (29b) and a green 'Probe' region (50b). A blue 'Labelled cRNA' molecule is hybridized to the probe region, with orange circles representing fluorescent labels.

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 useful
- 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**
- **Pin printing based methods limited to 40-50K**
 - **In situ synthesis: millions**
- **Array design is linked to purpose.**

Laboratory Essentials

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

DNA Microarray Applications

- Gene Expression
- Comparative Genomic Hybridization
- Resequencing (SNPs)
- 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**



Reports on Microarray Data Quality

Nature Biotechnology

September 2006

Accessing Expression Data

- Individual Lab and Journal Sites; public databases

GPL Platforms	1192
GSM Samples	35816
GSE Series	1824
Total	38824
Apr 08 2005	

GEO

Currently contains
 expression data on
 342,783 samples

<http://www.ncbi.nlm.nih.gov/geo/>

Accessing Expression Data

Experiments:	66	View
Arrays:	89	View
Protocols:	459	View
Hybridizations:	142	View

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.

288524 assays
 including 34,264
 curated, reannotated
 Assays

<http://www.ebi.ac.uk/microarray-as/ae/>

Publishing Expression Data

- MIAME standard

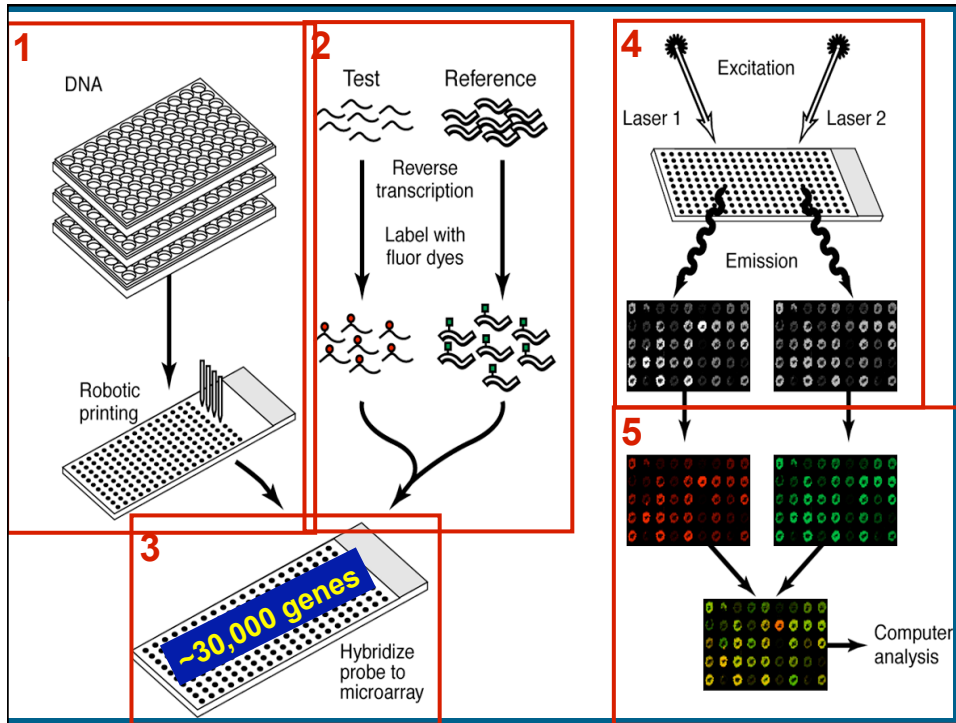
Minimum Information about a Microarray Experiment

- Format required by many journals
- Essential for database submissions

<http://www.mged.org/Workgroups/MIAME/miame.html>

STRATEGIES FOR SIGNAL GENERATION FROM mRNA

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



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

siRNA's



?????

Downstream Genes

•**Direct targets**

•**Indirect targets**

EXPRESSION DATA ANALYSIS

•**Large amount of data**

Examples: 200 samples x 25000 probes= 5,000,000 data points

•**Requires analysis and visualization tools**

Recent overview of microarray bioinformatics:
Simon R, Curr Opin Biotechnol. 2008 Feb;19(1):26-9.

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.

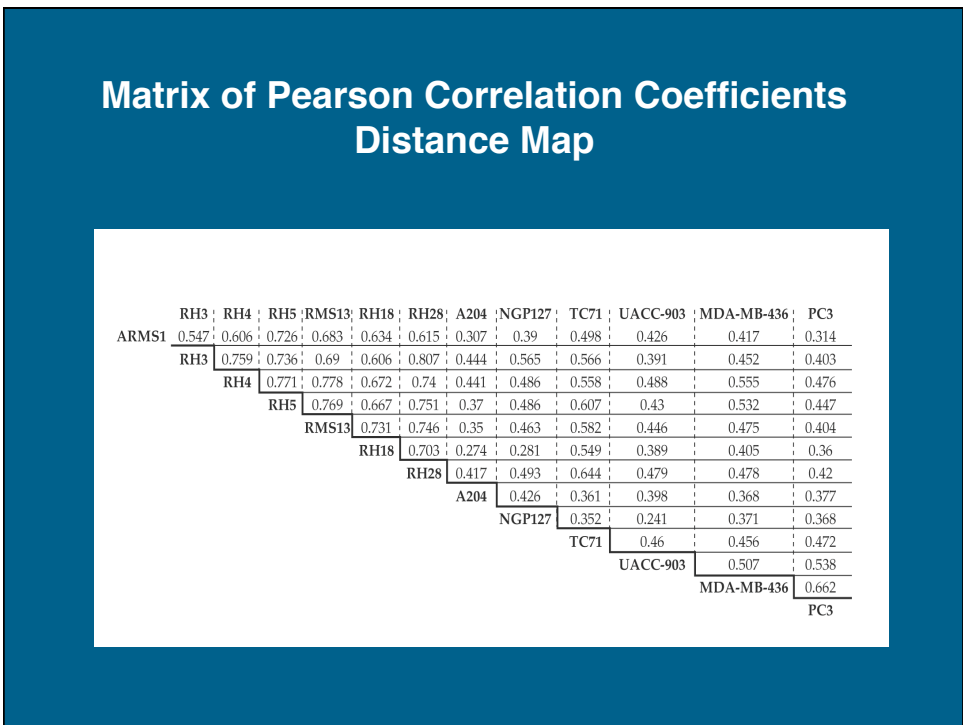
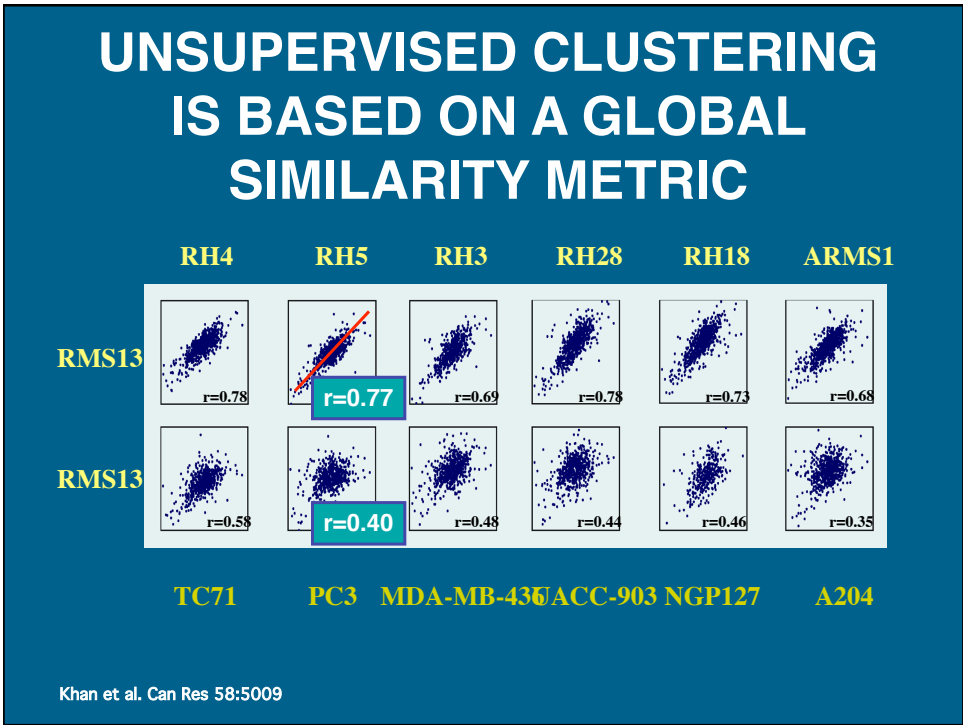
Does not prove the validity of groups.

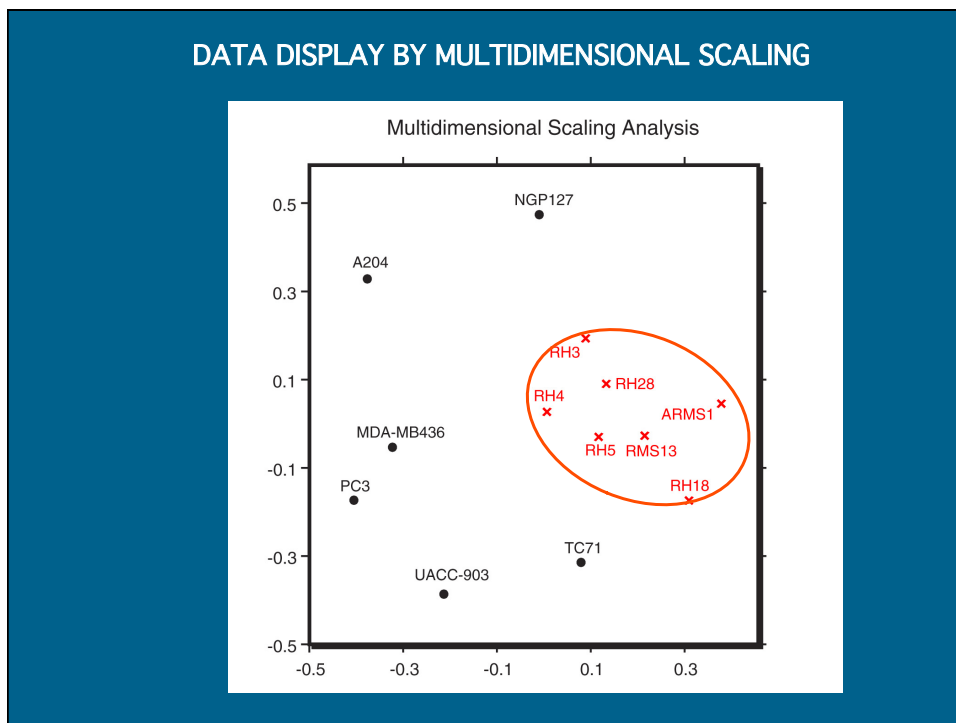
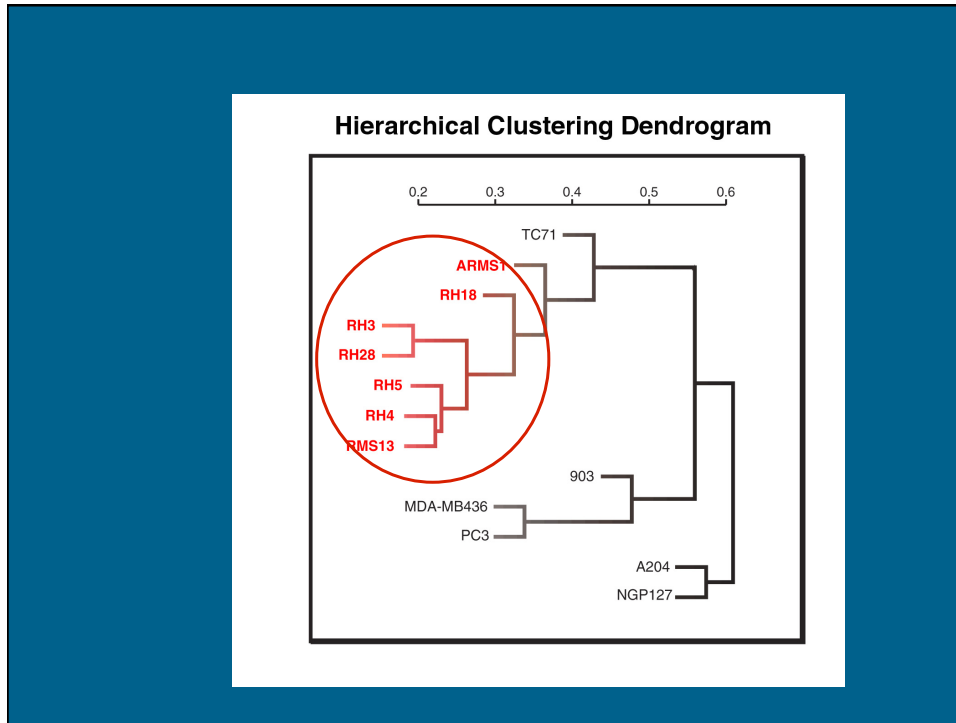
- Clustered Samples Are Biologically Similar

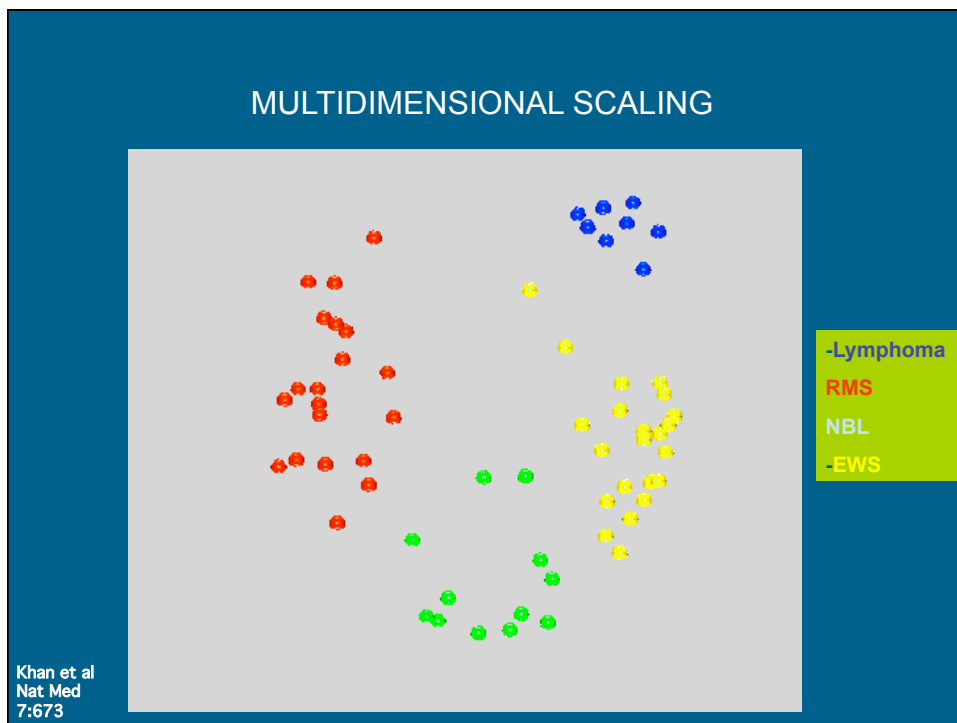
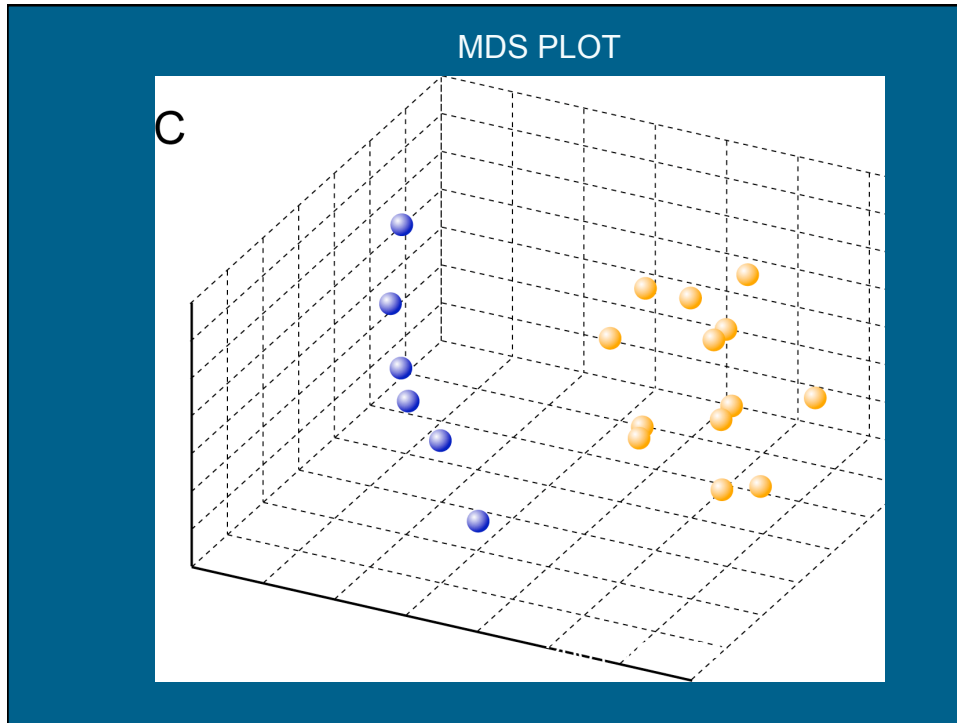
- Clusters of Co-expressed genes

- May be functionally related

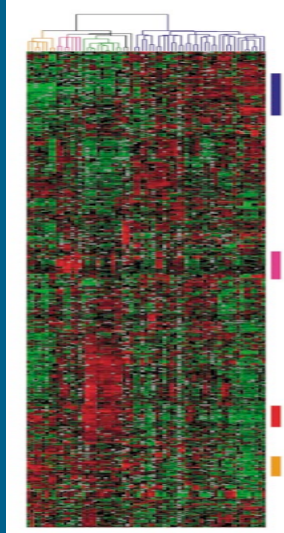
- May be enriched for pathways







CLUSTERING GENES AND SAMPLES

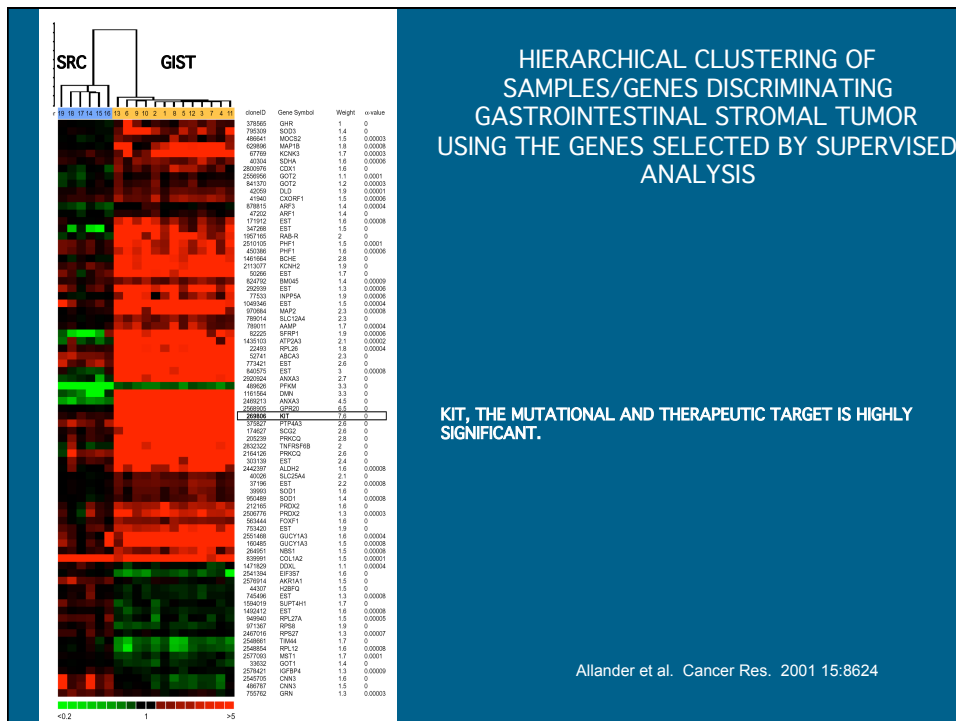
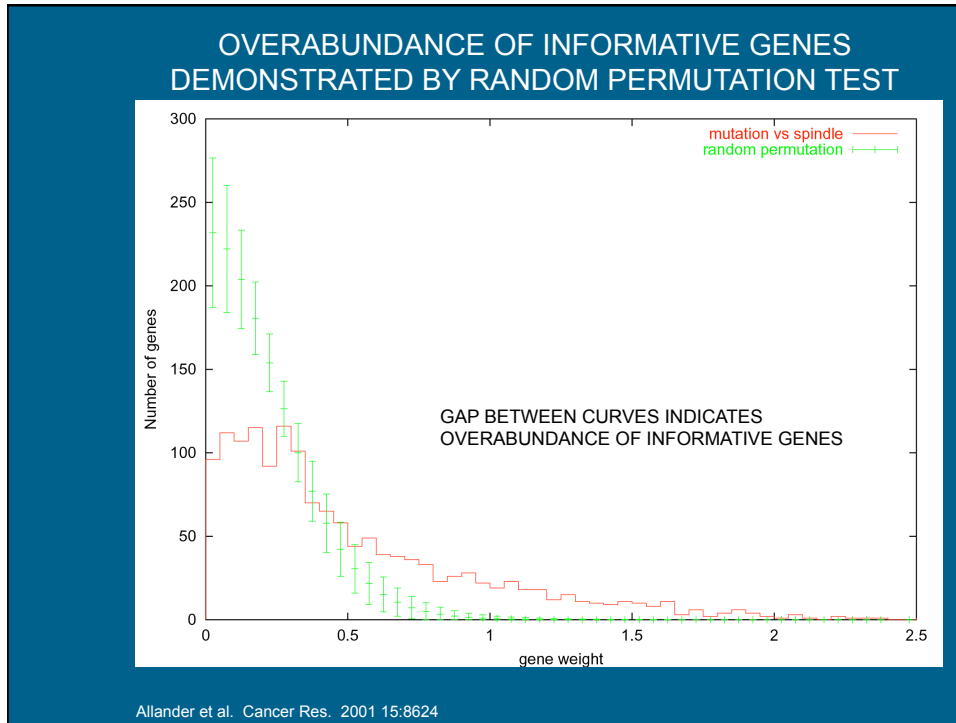


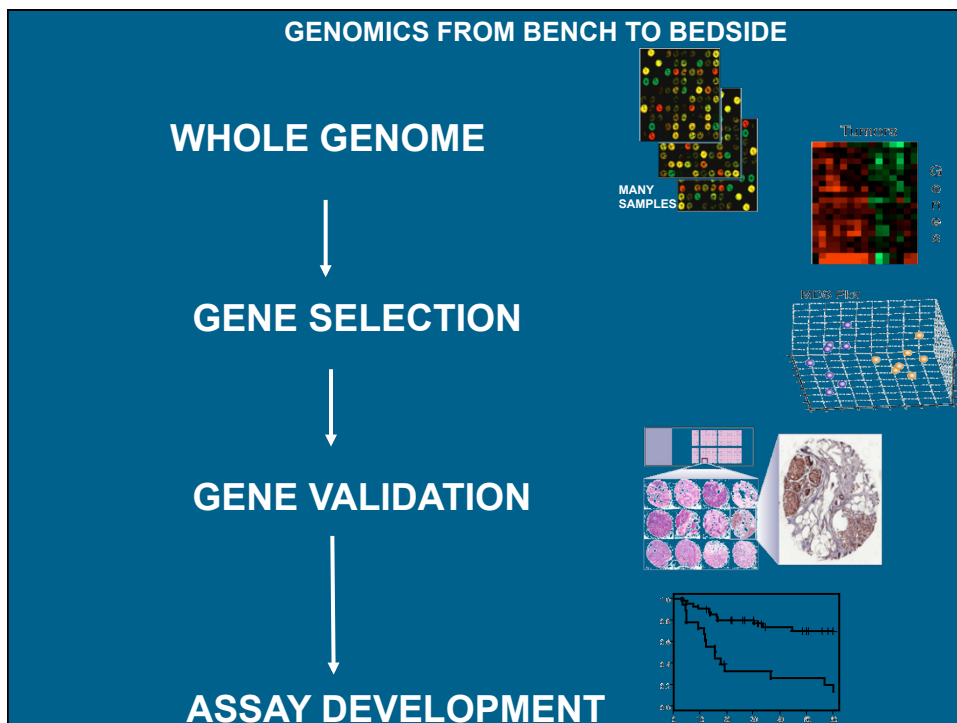
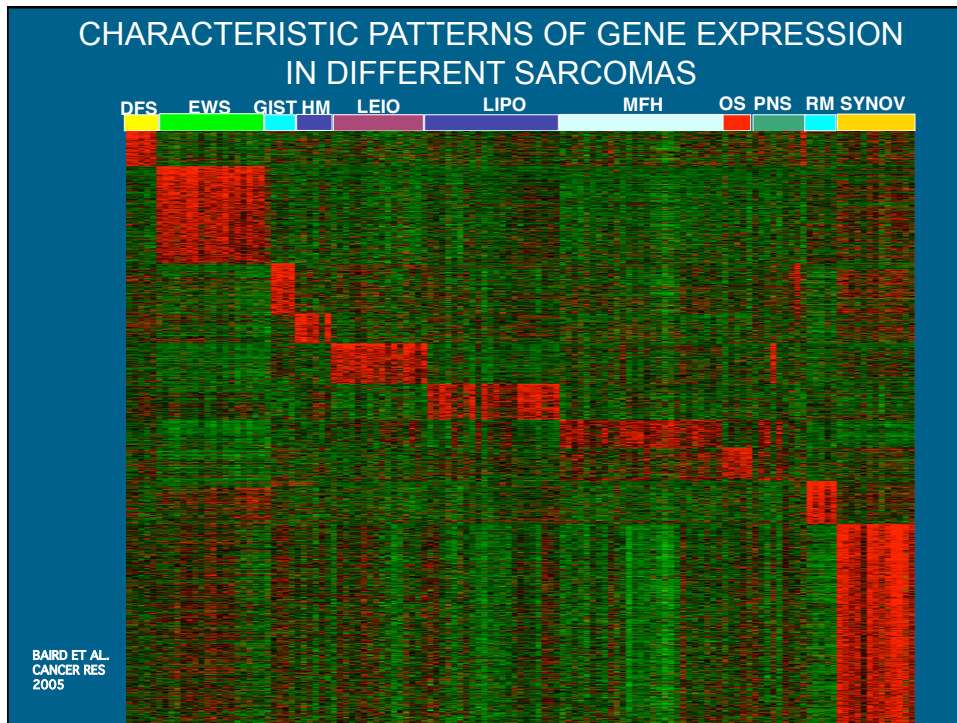
Perou et al. Nature 2000 406:747

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.

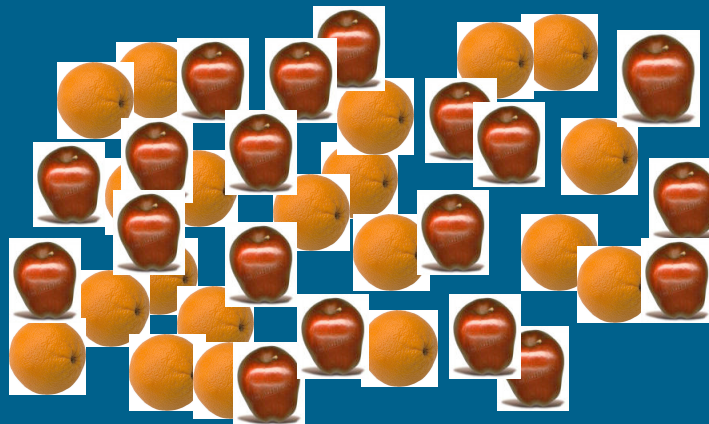




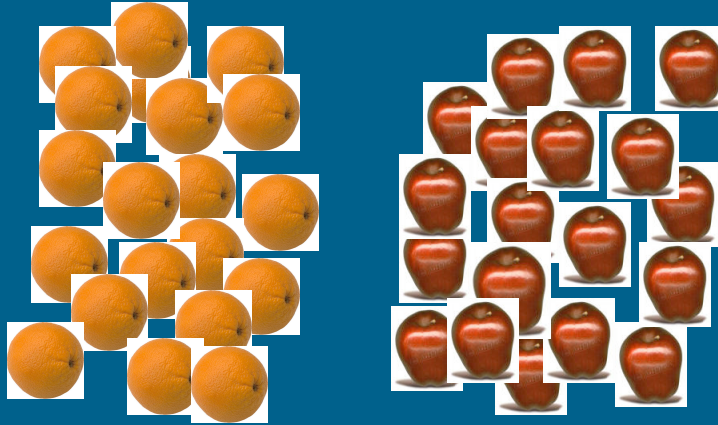
**SIGNAL STRENGTH VARIES IN
TISSUE PROFILING EXPERIMENTS**

**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)

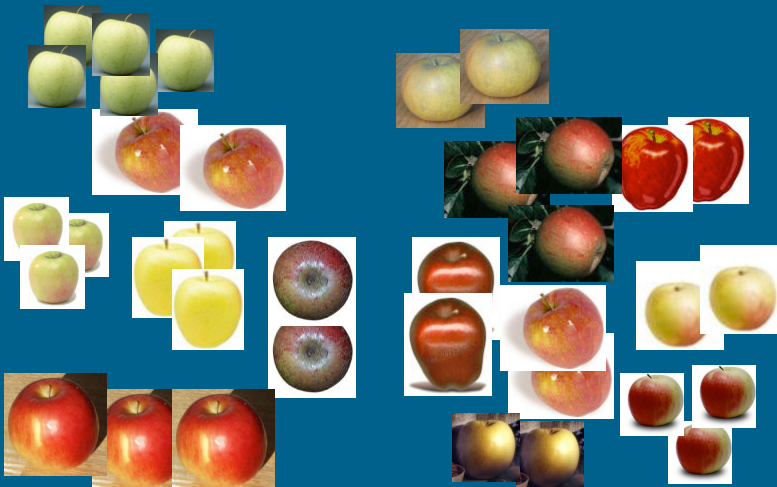
THESE ARE EASY TO DISTINGUISH BY
ONE MEASUREMENT PER INDIVIDUAL.

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**

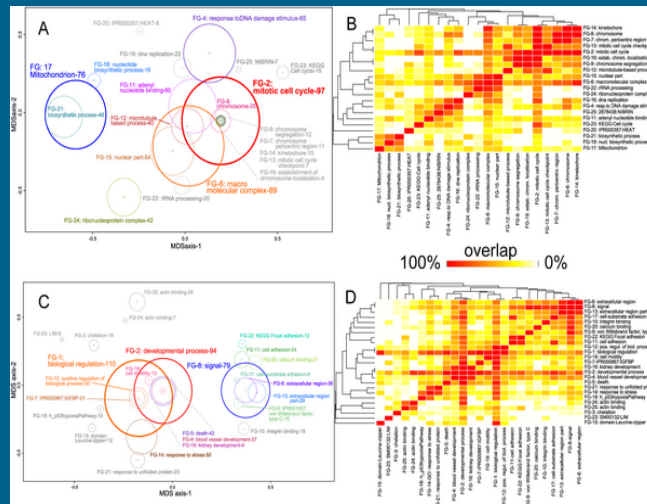
- Often cryptic and hard to interpret.
- Hypothesis generating, but this is often rather subjective.
- Seldom provide strong evidence for a specific mechanism.
- Expression data is intrinsically limited.

GETTING BEYOND GENE LISTS

- Optimal use of gene annotations.
- Gene Ontology
(<http://david.abcc.ncifcrf.gov/>)
- Optimizing use of public data.
 - GEO, ARRAY EXPRESS, ACADEMIC DATA
 - GENE SIGNATURE BASED METHODS (Gene Set Enrichment Analysis).

GENE ONTOLOGY AND PROMOTER DATABASES CAN HELP FIND BIOLOGY

GENE ONTOLOGY CATEGORIES AFFECTED BY ONCOGENE KNOCKDOWN IN EWING'S SARCOMA



KAUER ET AL. PLOS ONE 4:e5415 2009

GETTING BEYOND GENE LISTS

- Incorporating data from model systems.
- Linking expression data to sequence (e.g. Regulatory elements).
- Integrating other types of genome scale data.



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.

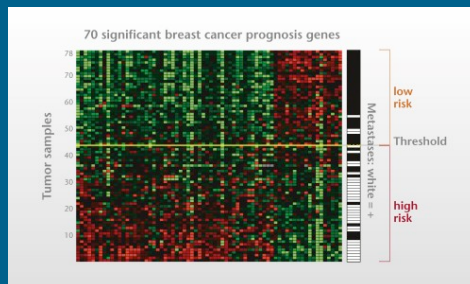
EXPRESSION PROFILING IN THE CLINIC?

- COMMERCIAL TESTS BEGINNING TO APPEAR.
- FDA IS ADDRESSING MULTIPLEX GENE EXPRESSION TESTS.
- LIMITED CLINICAL VALIDATION SO FAR

FDA APPROVED TESTS FOR BREAST CANCER BASED ON EXPRESSION STUDIES

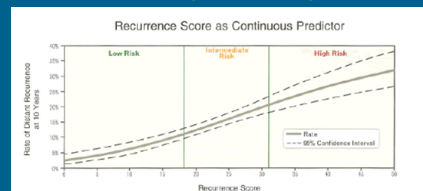
70 GENE MICROARRAY SIGNATURE

Van de Vijver et al
NEJM 347:1999 .



Multigene RT-PCR Signature

Paik et al NEJM 351:2817



ARRAYS VS. NEXT GENERATION SEQUENCING

- ARRAY TECHNOLOGIES MEASURE THE RELATIVE ABUNDANCE OF NUCLEIC ACIDS OF DEFINED SEQUENCE IN A COMPLEX MIXTURE.
- SEQUENCING CAN ACCOMPLISH THE SAME THING.

ARRAYS VS. NEXT GENERATION SEQUENCING

MICROARRAYS

- READILY AVAILABLE MATURE TECHNOLOGY
- RELATIVELY INEXPENSIVE
- EFFECTIVE WITH VERY COMPLEX SAMPLES
- HUNDREDS OF SAMPLES PRACTICAL
- CAN TARGET SUBSET OF GENOME

PROS

CONS

- REQUIRE PLATFORM AND APPLICATION SPECIFIC DATA PROCESSING
- PRONE TO PLATFORM SPECIFIC ARTIFACTS
- MANY SOURCES OF NOISE
- WHOLE GENOME STUDIES GENERALLY REQUIRE MANY ARRAYS, INCREASING SAMPLE REQUIREMENTS AND COMPLICATING ANALYSIS

MICROARRAYS

SEQUENCING

- WHOLE GENOME DATA
- RELATIVELY UNIFORM ANALYTICAL PIPELINE
- FREE OF HYBRIDIZATION ARTIFACTS
- POSSIBILITY OF ONE PLATFORM FOR ALL APPLICATIONS

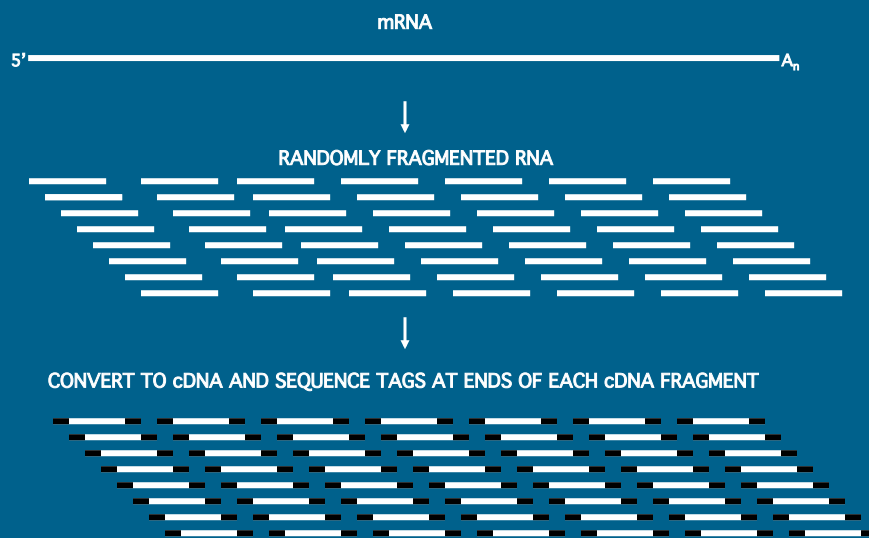
- IMMATURE TECHNOLOGY
- HIGH COSTS
- COMPUTATIONALLY INTENSIVE
- LIMITED SAMPLE THROUGHPUT

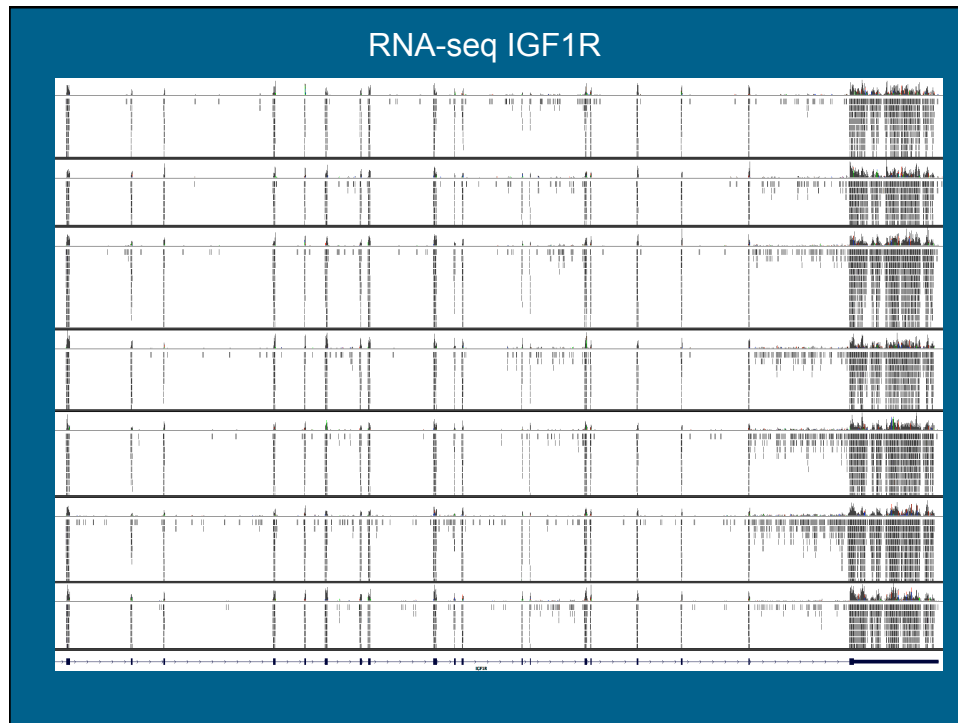
SEQUENCING

MEASURING GENE EXPRESSION BY RNA SEQUENCING

- TAG SEQUENCING (SAGE-LIKE)
- FULL LENGTH mRNA----RNA-Seq
- 3' fragment mRNA sequencing
- miRNA sequencing

MEASURING GENE EXPRESSION BY RNA SEQUENCING



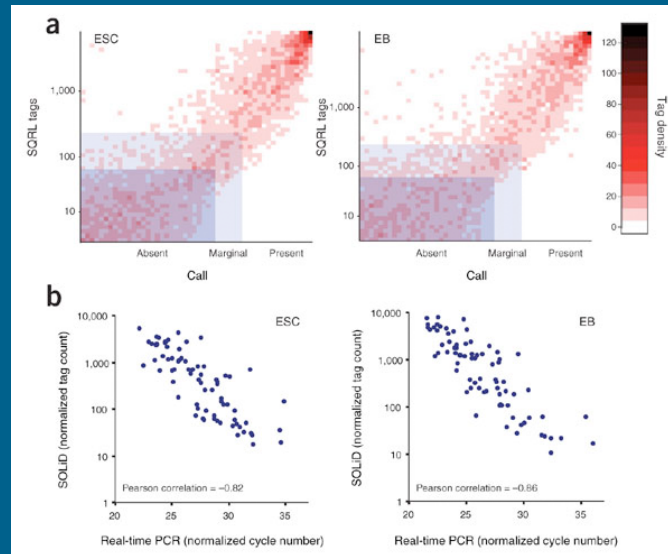


MEASURING GENE EXPRESSION BY RNA SEQUENCING: PROS AND CONS

ADVANTAGES

- RNA SEQUENCE VARIATIONS DETECTED AT SINGLE NUCLEOTIDE RESOLUTION
 - ALLELE SPECIFIC EXPRESSION
 - MUTATIONS
- RNA STRUCTURE: SPLICING, START SITES, TERMINATION SITES; REARRANGEMENTS
- DETECTED SIGNALS ARE RELATIVELY UNAMBIGUOUS; POTENTIAL TO OUTPERFORM MICROARRAY
- DE NOVO ASSEMBLY IS POSSIBLE

MEASURING GENE EXPRESSION BY RNA SEQUENCING



Cloonan et al.

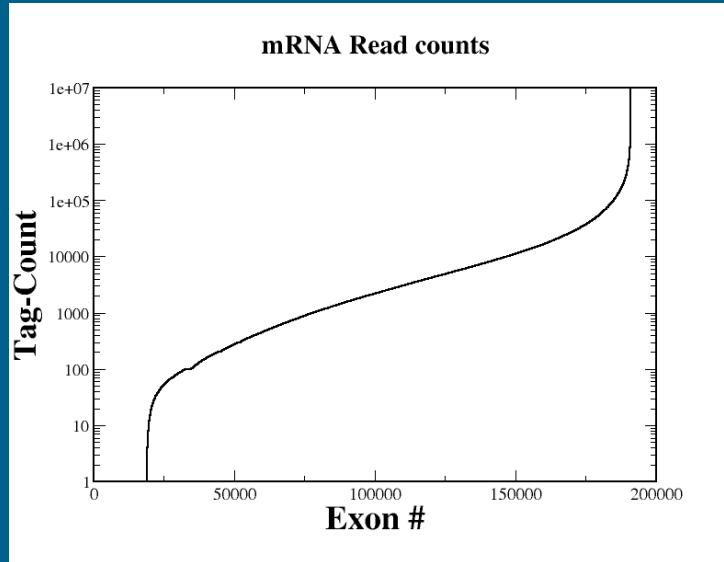
Nature Methods - 5, 613 - 619 (2008)

MEASURING GENE EXPRESSION BY RNA SEQUENCING: PROS AND CONS

LIMITATIONS

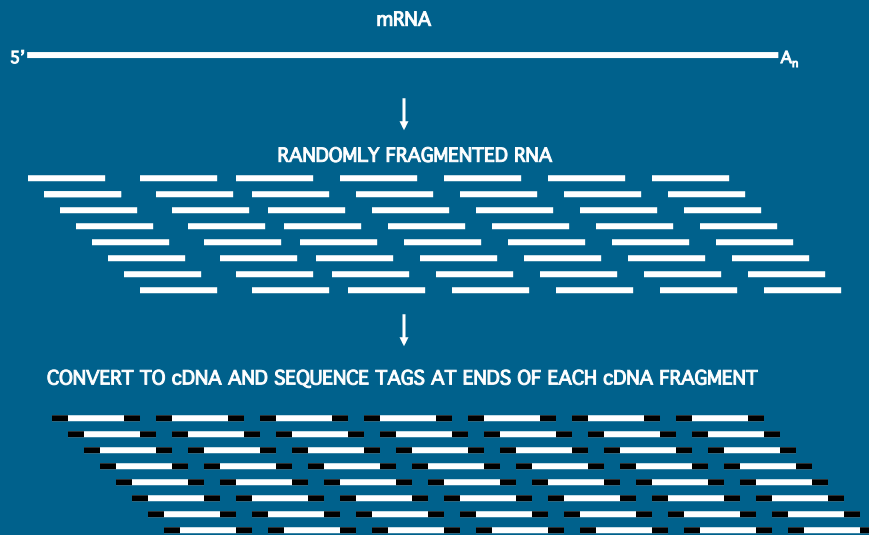
- LOWER LIMIT OF DETECTION IS CONSTRAINED BY THE mRNA ABUNDANCE DISTRIBUTION AND THE NUMBER OF ALIGNED READS PER SAMPLE.
- LARGE SAMPLE NUMBERS DIFFICULT TO ACHIEVE, EXCEPT IN TAG MODE.
- SOFTWARE IS STILL DEVELOPMENTAL: REQUIRES SOPHISTICATED BIOINFORMATICS COLLABORATION. [For review see Pepke et al. Nat Methods 6:S22 (2009)]

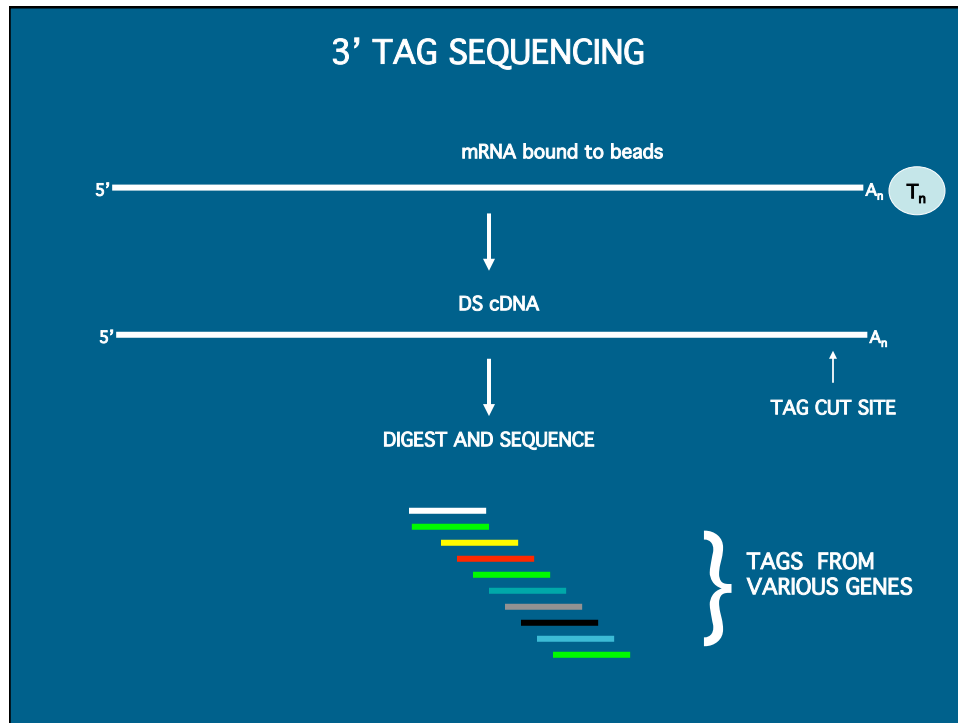
mRNA ABUNDANCE VARIES OVER A LARGE DYNAMIC RANGE



Sven Bilke

MEASURING GENE EXPRESSION BY
RNA SEQUENCING





- ### 3' TAG SEQUENCING
- SEQUENCES ALIGNED AND COUNTED
 - LIBRARIES OF TAGS FROM MANY SAMPLES CAN BE IDENTIFIED BY ADDING A “BARCODE” AND POOLED BEFORE SEQUENCING
 - POTENTIAL TO ANALYZE LARGE NUMBERS OF SAMPLES IN PARALLEL

THE FUTURE?

AS SEQUENCE THROUGHPUT INCREASES AND COSTS PER READ DECLINE, SEQUENCING IS LIKELY TO BECOME AN ATTRACTIVE ALTERNATIVE TO MICROARRAYS IN MORE AND MORE APPLICATIONS.

USEFUL WEB SITES

MGEGD The Microarray Gene Expression Data Society:

<http://www.mged.org/>

NCBI Gene Expression Omnibus:

<http://ncbi.nih.gov/geo/>

NCBI Sequence Read Archive (SRA):

<http://www.ncbi.nlm.nih.gov/sra>

EBI Microarray informatics:

<http://www.ebi.ac.uk/microarray/index.html>

Stanford Microarray Database:

<http://smd.stanford.edu/>

UCSF DeRisi lab:

<http://derisilab.ucsf.edu/data/microarray/index.html>

Broad Institute:

Gene Set Enrichment Analysis (GSEA)

<http://www.broadinstitute.org/gsea/>

Connectivity Map:

<http://www.broadinstitute.org/cmap/>