THE OHIO STATE UNIVERSITY
UNIVERSITY MEDICAL CENTER

Biomedical Informatics

3184 Graves Hall 333 W. 10th Ave. Columbus, OH 43210 614-292-4778 (phone) 614-292-7659 (fax)

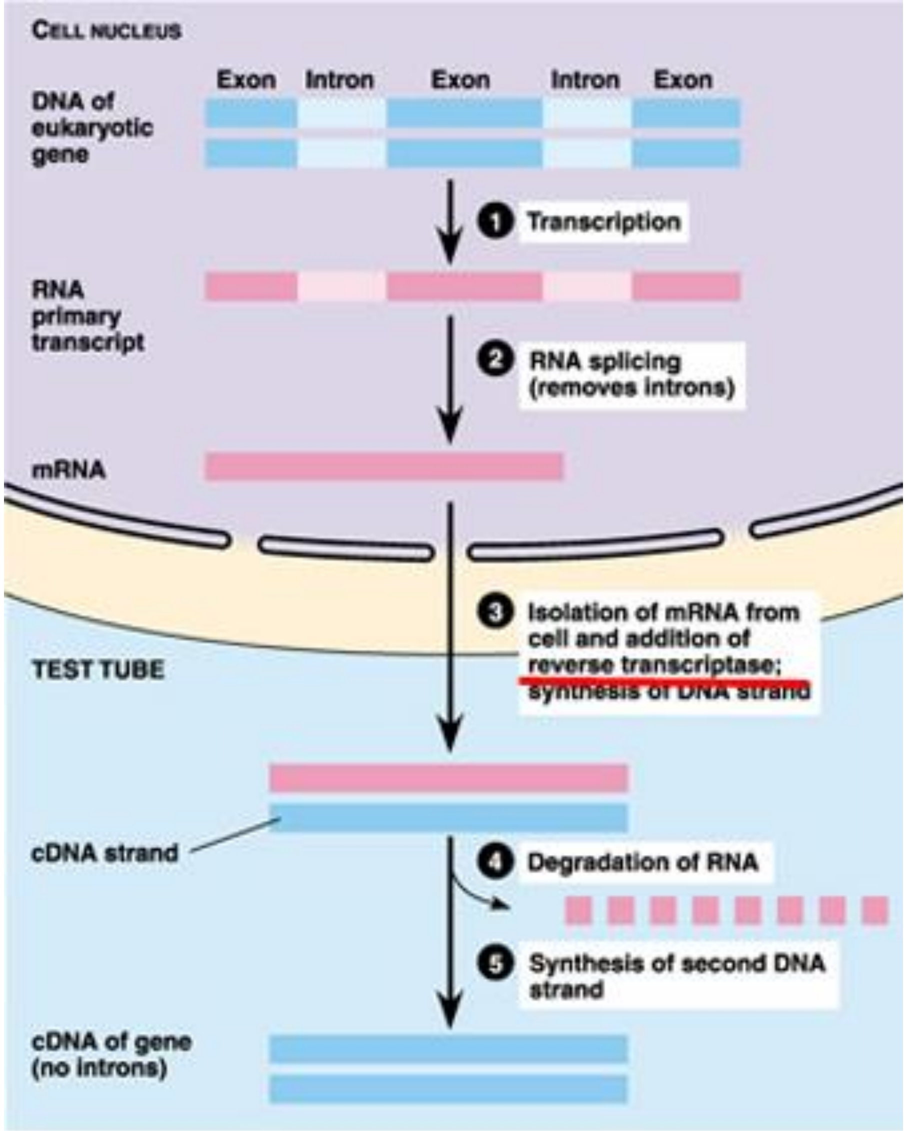
Introduction to Microarray Data Analysis

- Introduction to gene expression microarray
  - A middle-man’s approach
  - Applications of microarray
- Microarray data processing/analysis workflow
  - Data format and visualization
  - Data normalization
    - Two-color array
    - Affymetrix array
- Software and databases

Review of Biology

mRNA, cDNA,

exon, intron

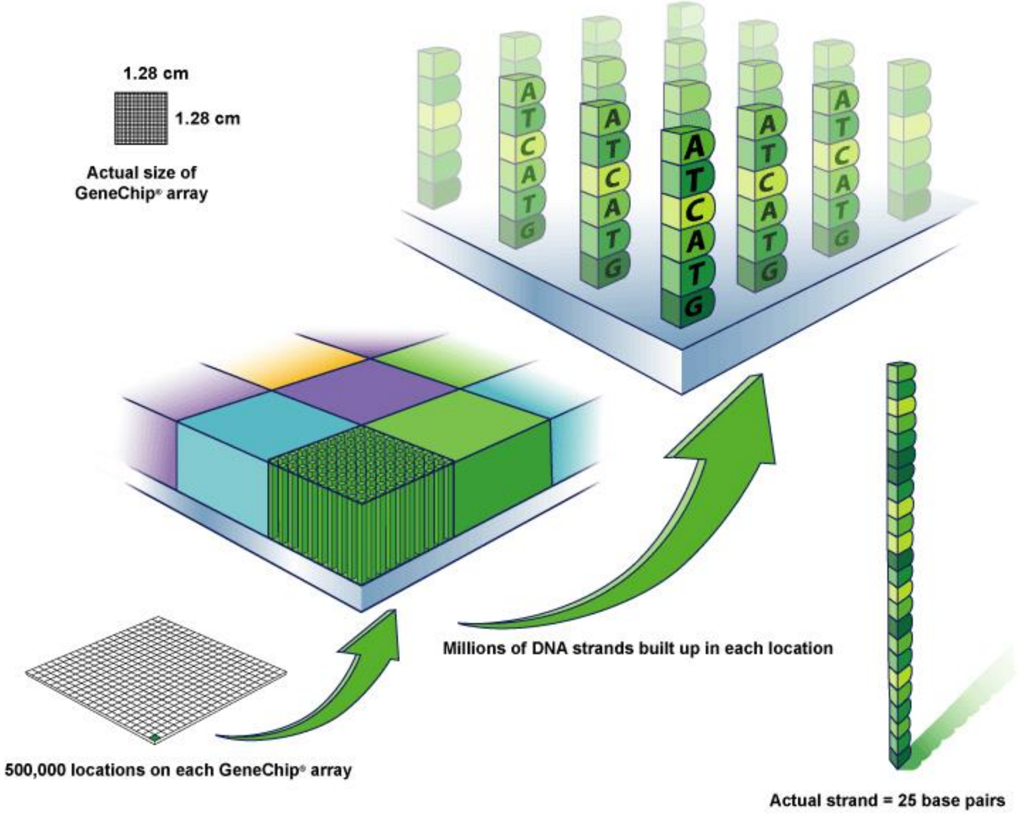

CELL NUCLEUS
Exon
Intron
Exon
Intron
Exon
DNA of eukaryotic gene
1 Transcription
RNA primary transcript
2 RNA splicing (removes introns)
mRNA
3 Isolation of mRNA from cell and addition of reverse transcriptase; synthesis of DNA strand
TEST TUBE
cDNA strand
4 Degradation of RNA
5 Synthesis of second DNA strand
cDNA of gene (no introns)

What is microarray?

- If we can assay every single molecule of DNA/RNA of interest directly, do we still need microarray?
- Currently direct single-molecule sequencing is still not mature, probes are used instead. Probe is a “middle-man”.

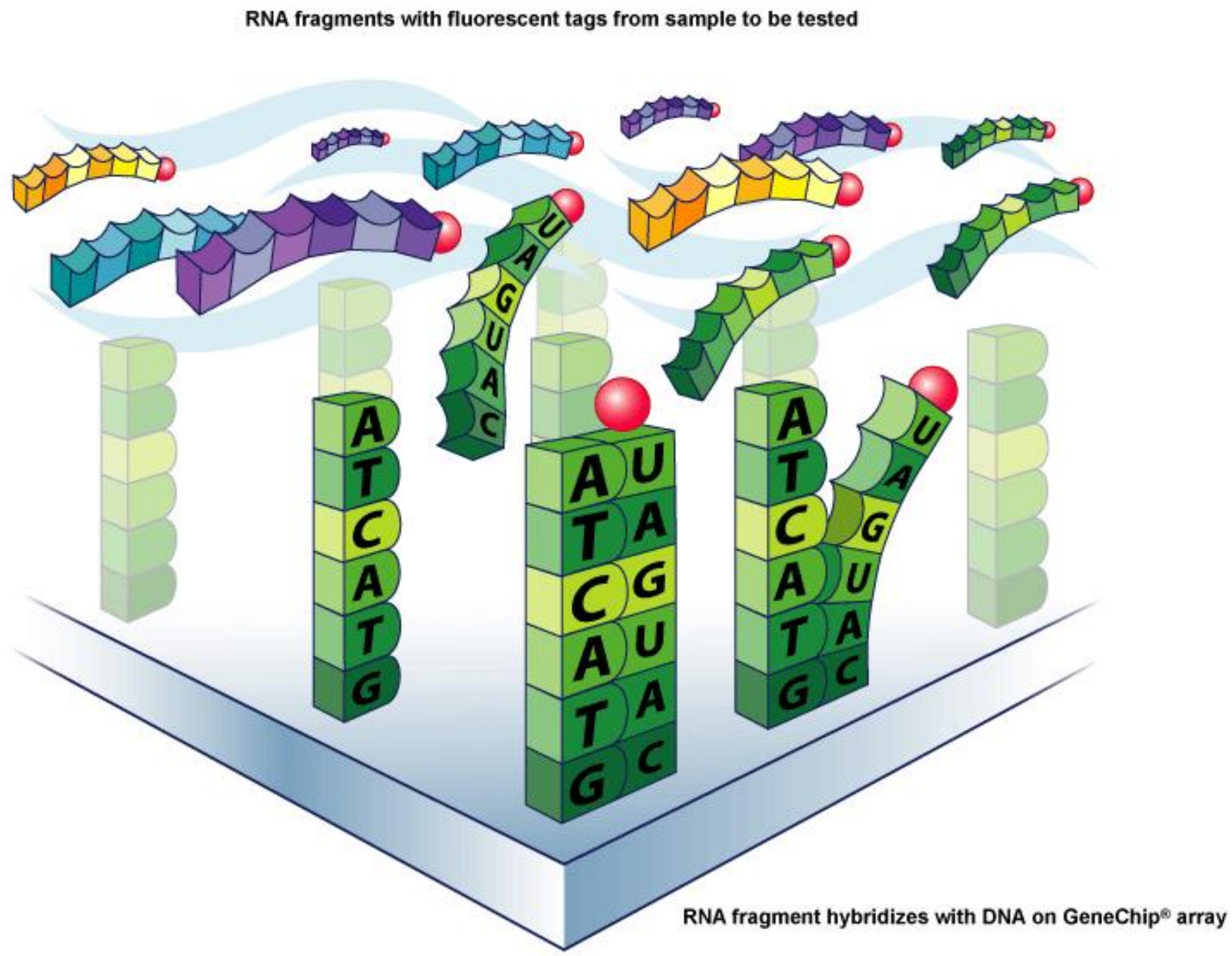
How is microarray manufactured?

- Affymetrix GeneChip
  - silicon chip
  - oligonucleiotide probes lithographically synthesized on the array
  - cRNA is used instead of cDNA



1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
500,000 locations on each GeneChip® array
Millions of DNA strands built up in each location
Actual strand = 25 base pairs

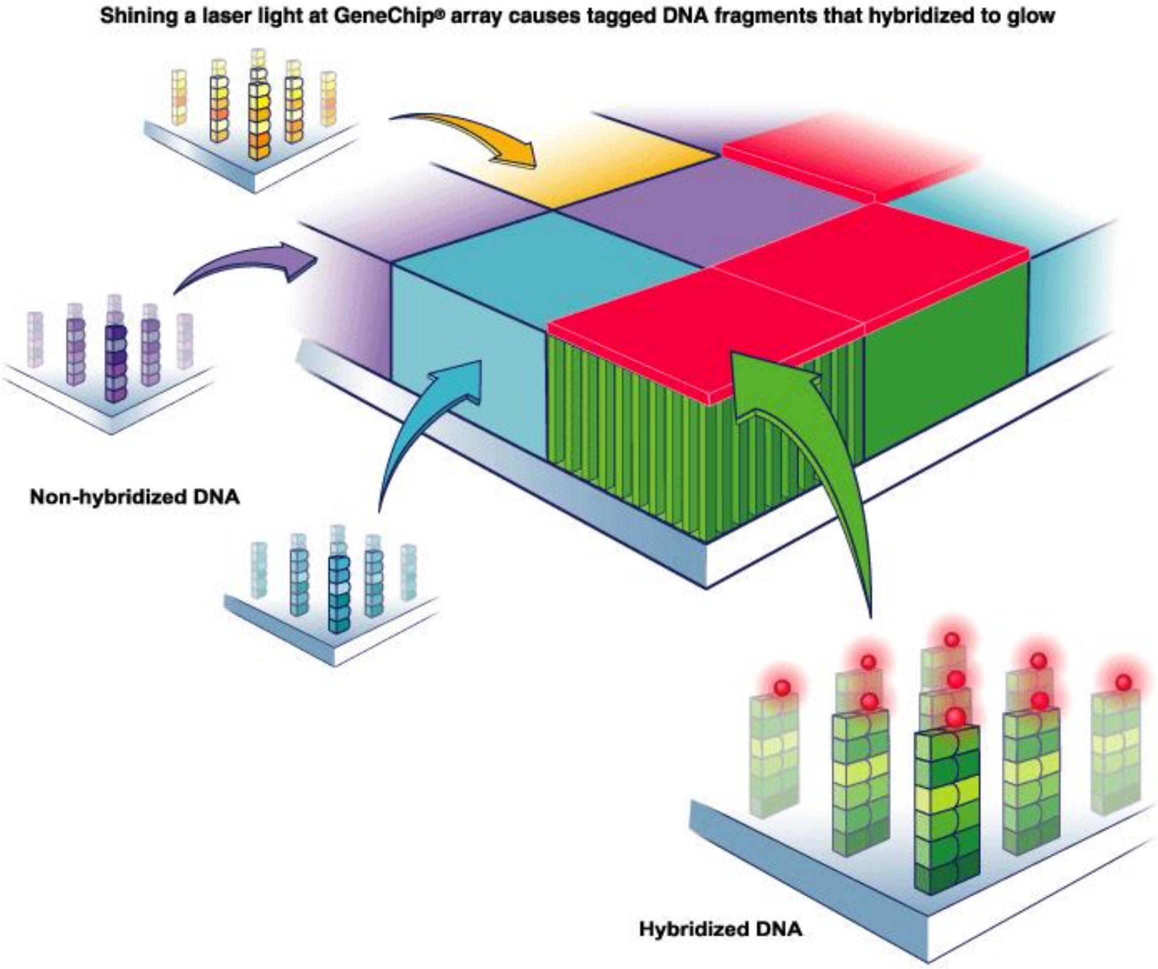
How does microarray work?



RNA fragments with fluorescent tags from sample to be tested

RNA fragment hybridizes with DNA on GeneChip® array

How does microarray work?

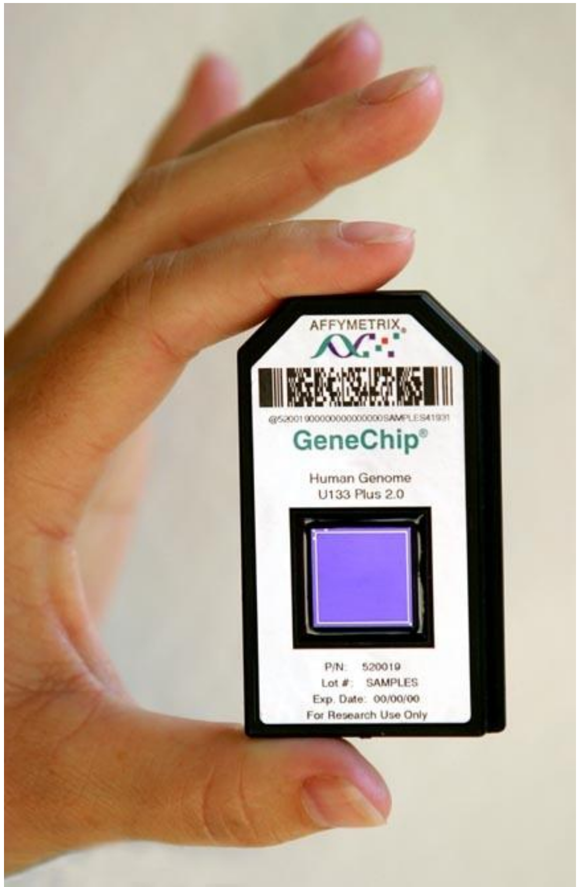

Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

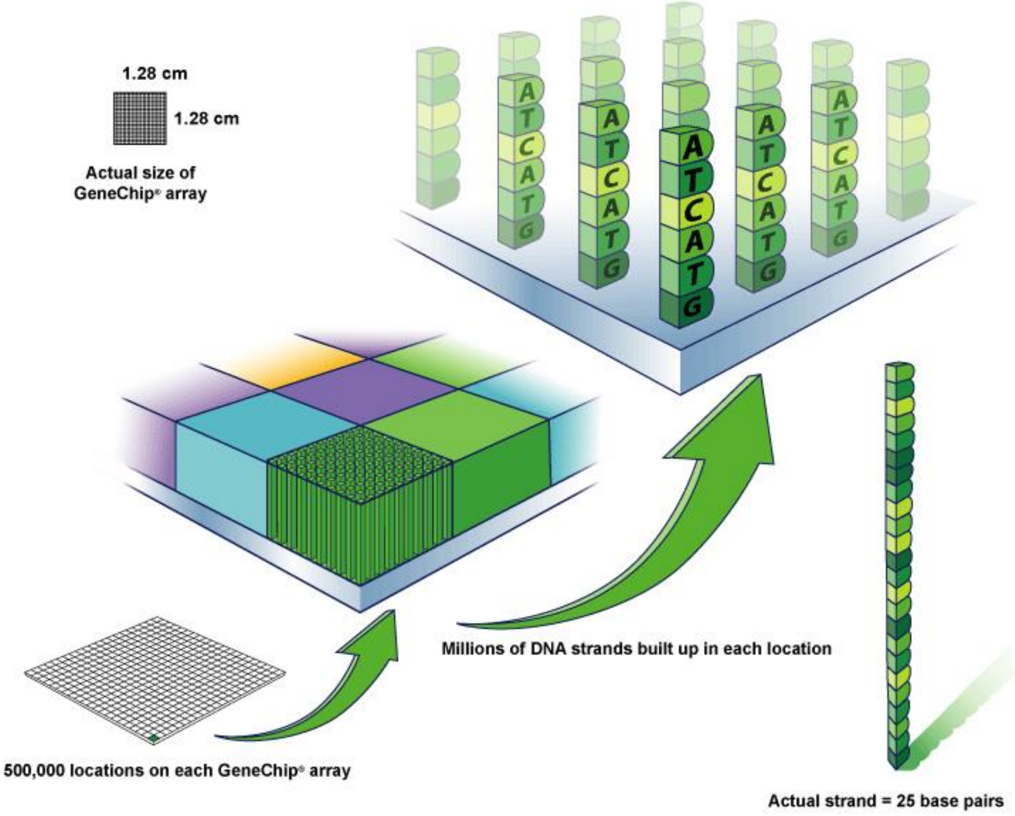
Two-major types of microarray

- Affymetrix-like arrays – single channel (background-green, foreground-red)
- cDNA arrays – two channel (red, green, yellow)

Affymetrix GeneChip

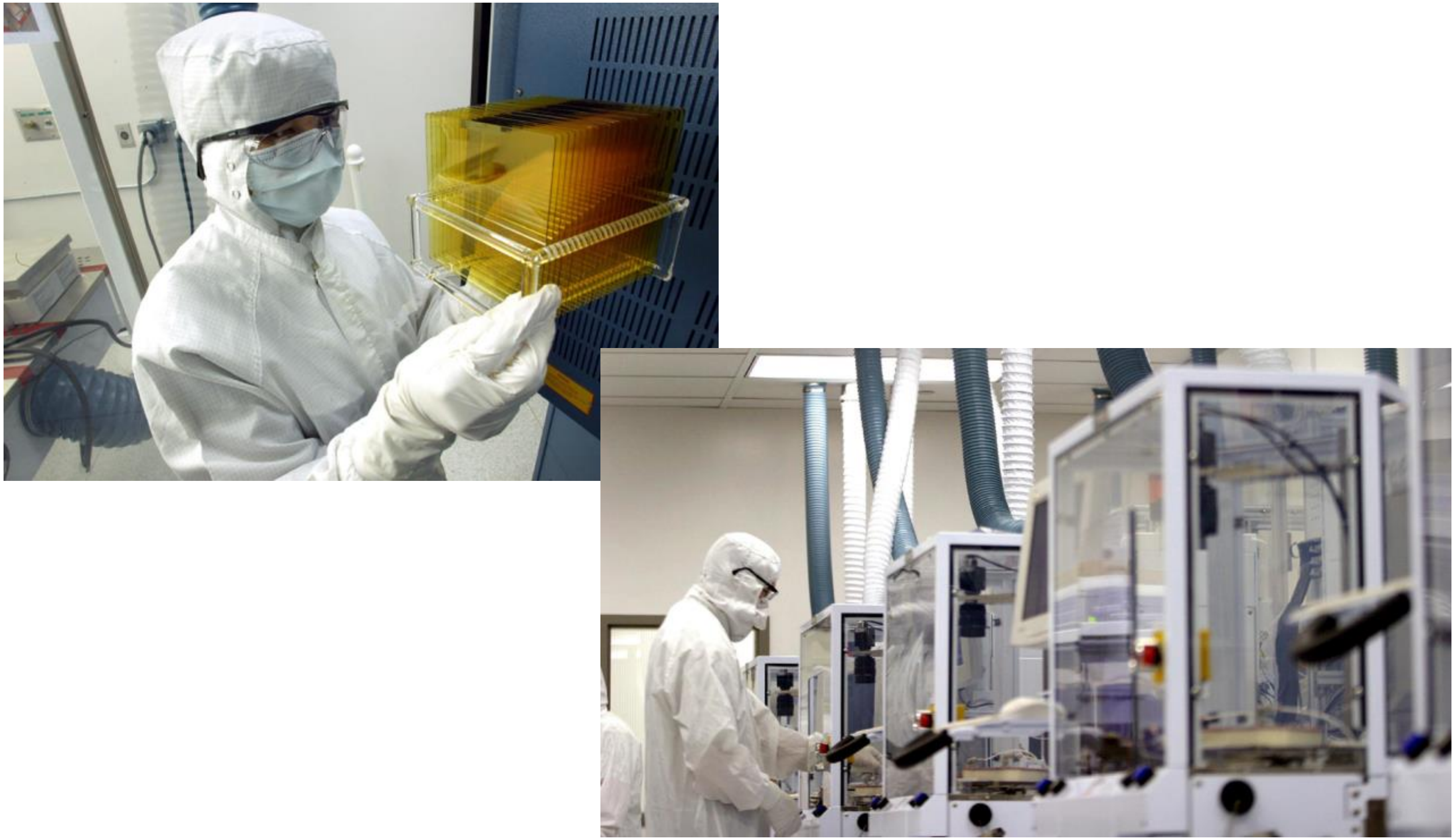
- silicon chip
- oligonucleiotide probes lithographically synthesized on the array
- cRNA is used instead of cDNA




1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

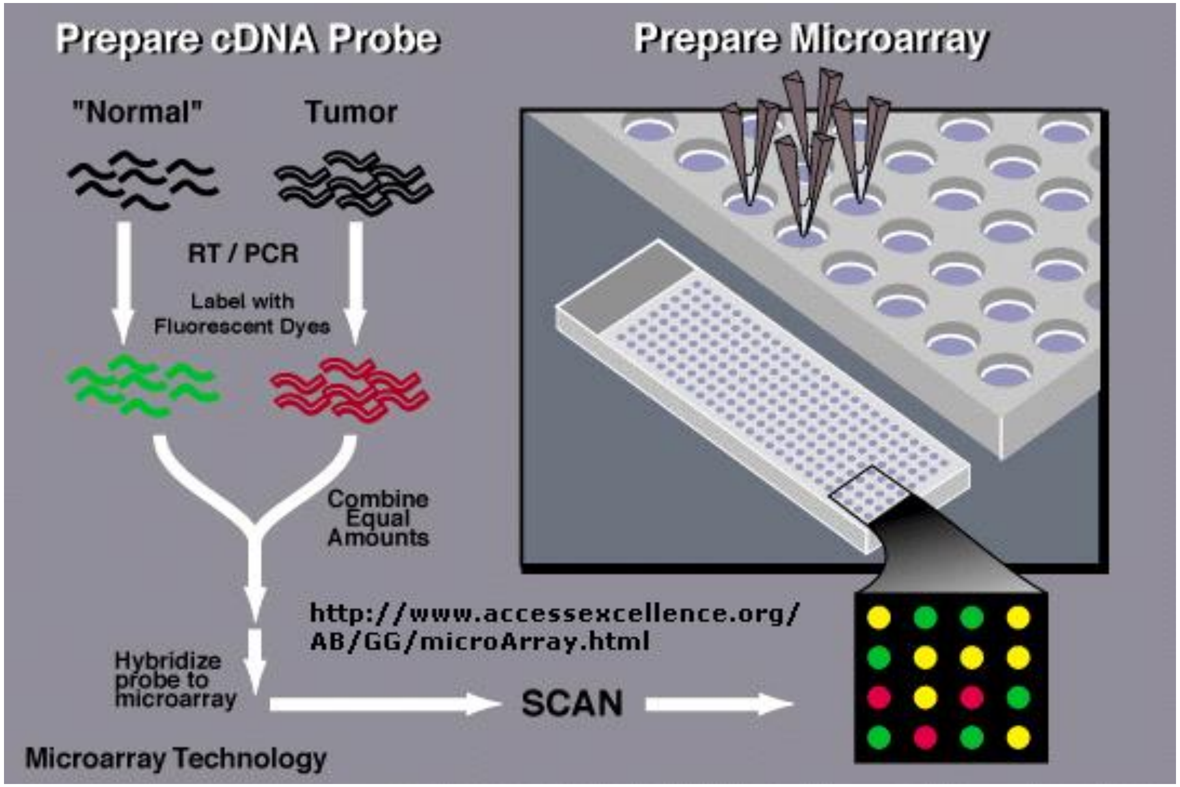
Affymetrix GeneChip

- silicon chip
- oligonucleiotide probes lithographically synthesized on the array

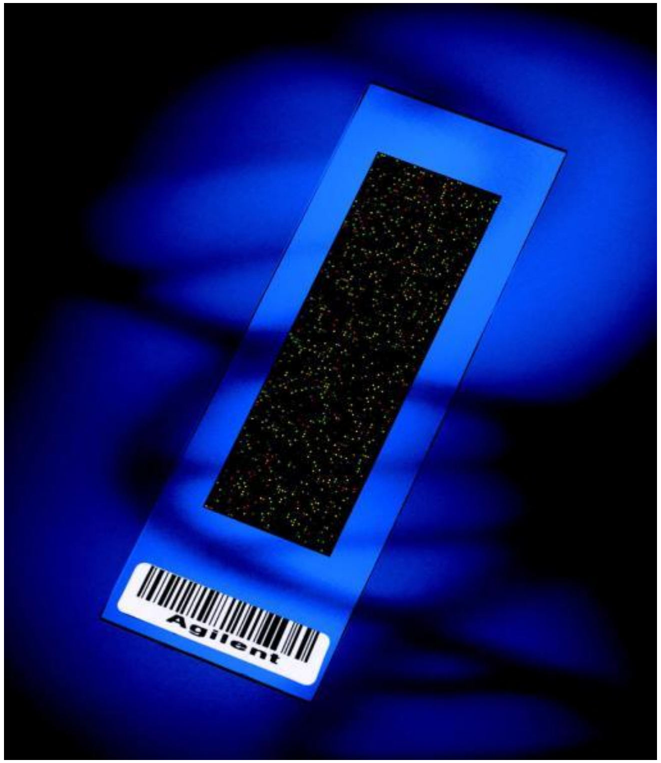


Two-channel microarray

- Printed microarrays
- Long probe oligonucleotides (80-100) long are “printed” on the glass chip
- Comparative hybridization experiment

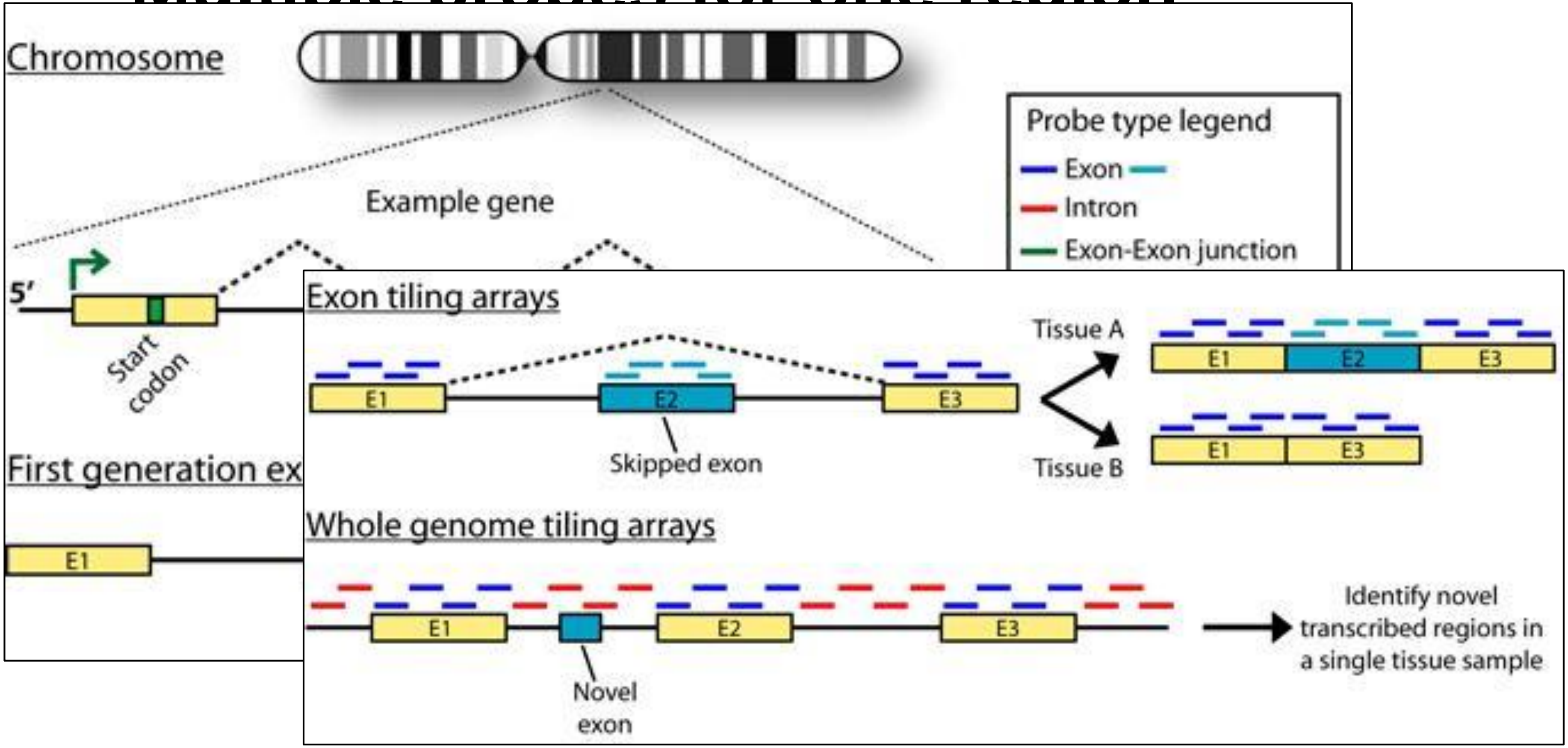


http://www.accessexcellence.org/AB/GG/microArray.html



Probe selection

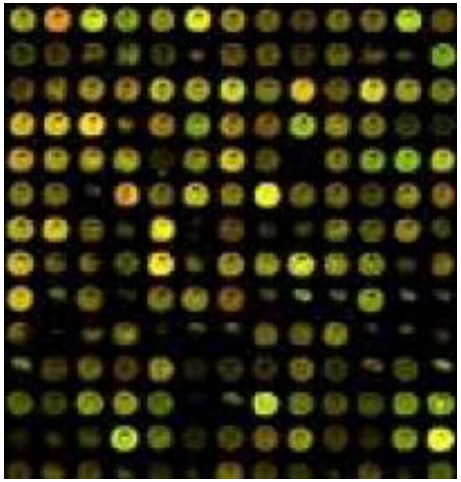
- Protocol for extracting mRNA
- 3’ bias – why? Think degradation.
- Multiple probes for one region



http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm

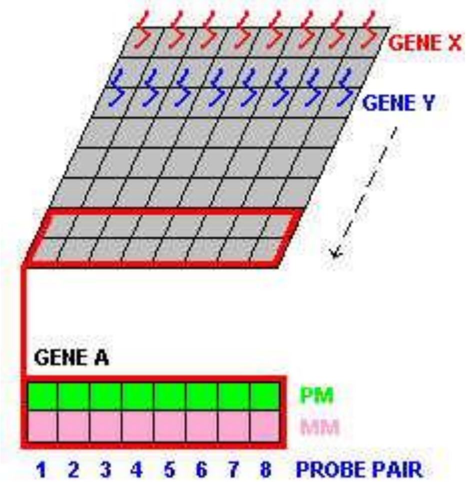
How do we process microarray data (measurement)?

- **cDNA array – ratio, log ratio**



$$T_i = \frac{R_i}{G_i} \text{ OR } \log \text{ratio} = \log_2 \frac{R_i}{G_i}$$

- **Affymetrix array**



$$\text{Difference}_{\text{probe pair}} = PM - MM$$

$$\text{Average Difference}_{\text{probe set}} = \sum_{i=1}^{n} \frac{(PM_i - MM_i)}{n}$$

Applications of microarrays

- Gene expression
- Exon expression
- SNP detection
- Copy number variance (arrayCGH)
- Tiling array (e.g., ChIP-chip)

Major vendors

- Affymetrix
- Agilent
- Illumina
- Nimblegen

- Introduction to gene expression microarray
  - A middle-man’s approach
  - Applications of microarray
- **Microarray data processing/analysis workflow**
  - **Data format and visualization**
  - **Data normalization**
    - **Two-color array**
    - **Affymetrix array**
- Software and databases

Typical workflow

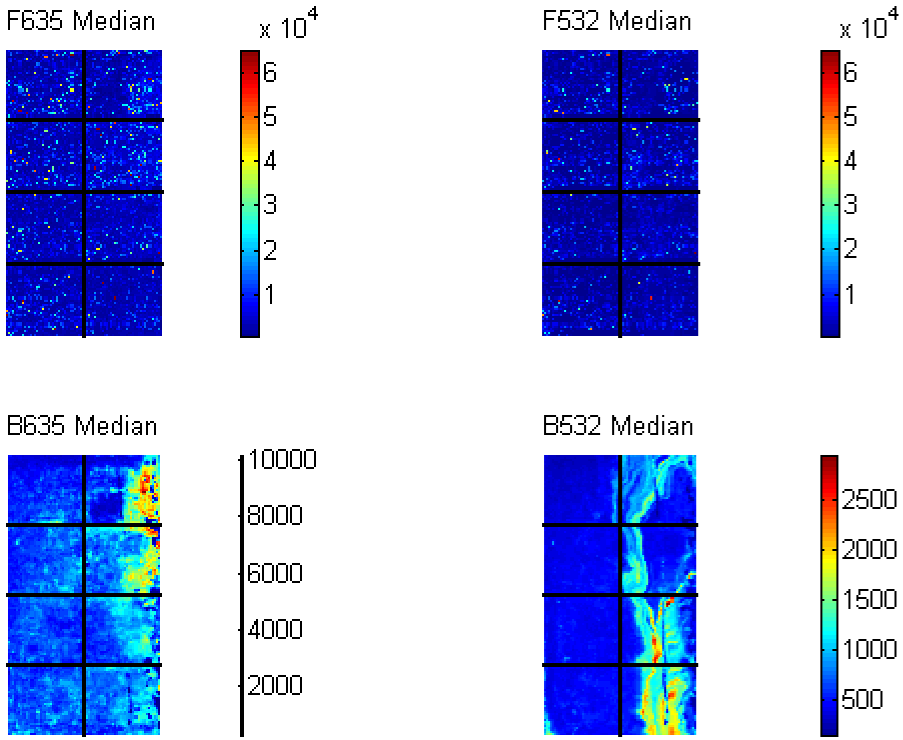
- QC
- Normalization
- Visualization (boxplot, PCA, RI plot, etc).

- Comparative study (volcano plot)
- Clustering

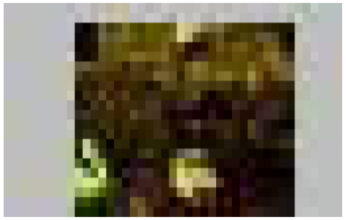
- Network/pathway inference
- Motif finding

Spatial Images of the Microarrays

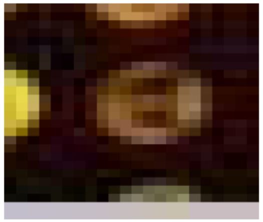
- Data for the same brain voxel but for the untreated control mouse
- Background levels are much higher than those for the Parkinson’s disearse model mouse
- There appears to be something non random affecting the background of the green channel of this slide



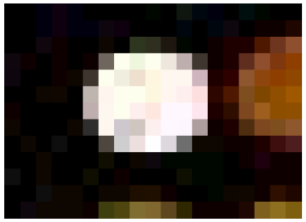
Take a look …



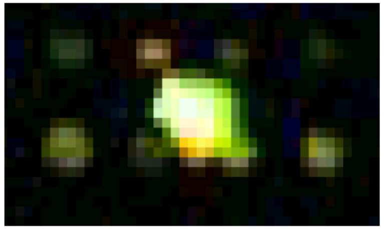
Poorly defined borders



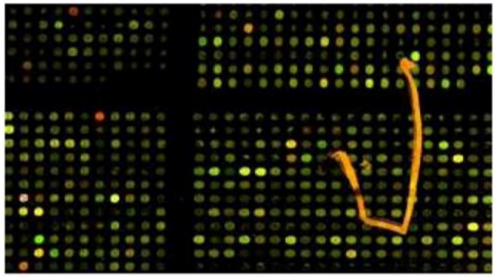
Large holes



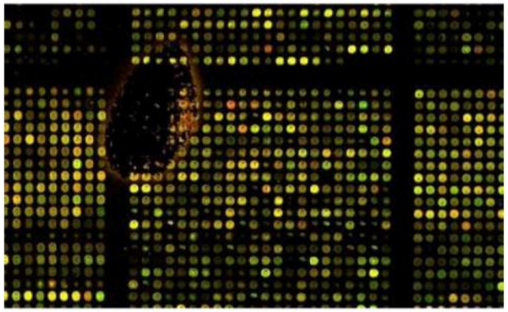
Saturated spot



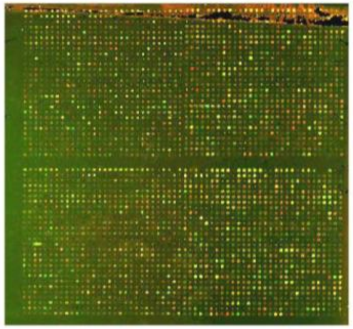
Dust specs



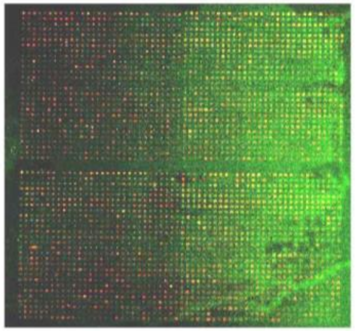
Fiber or scratch?



Bubble



Edge effect

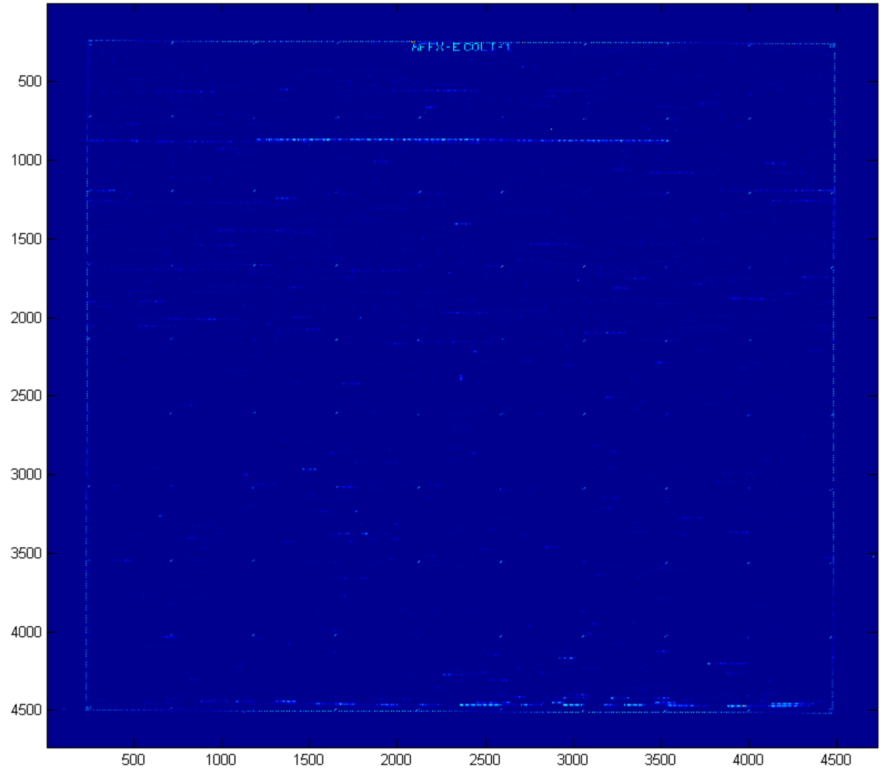


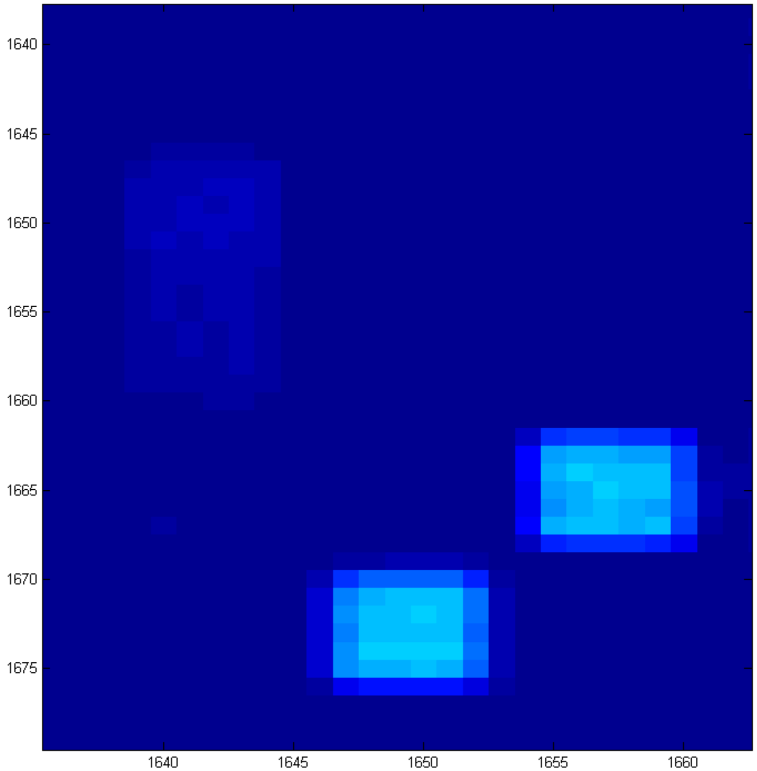
Background haze

(McShane, NCI)

Example – Affymetrix Data Files

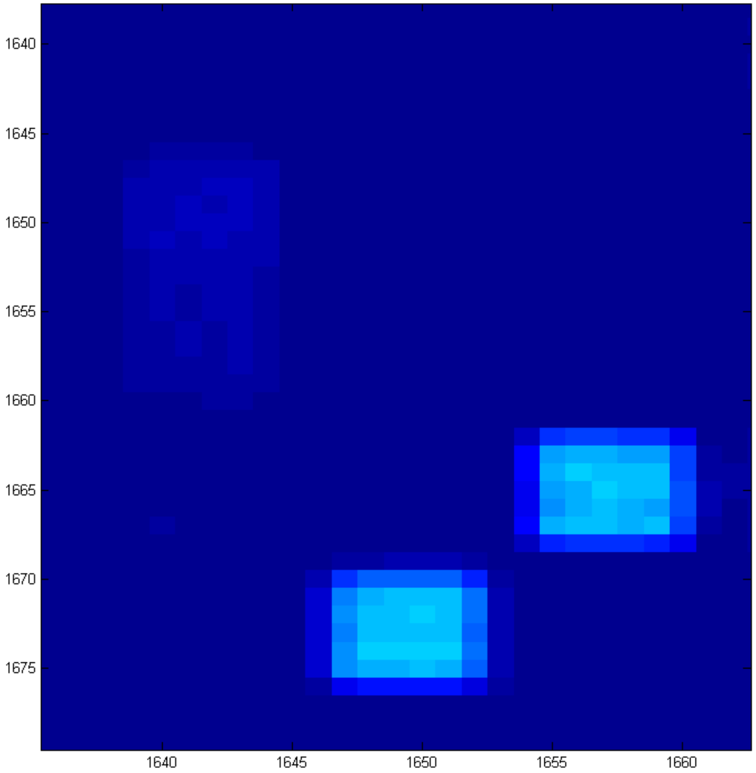
- Image file (.dat file)
- Probe results file (.cel file)
- Library file (.cdf, .gin files)
- Results file (.chp file)

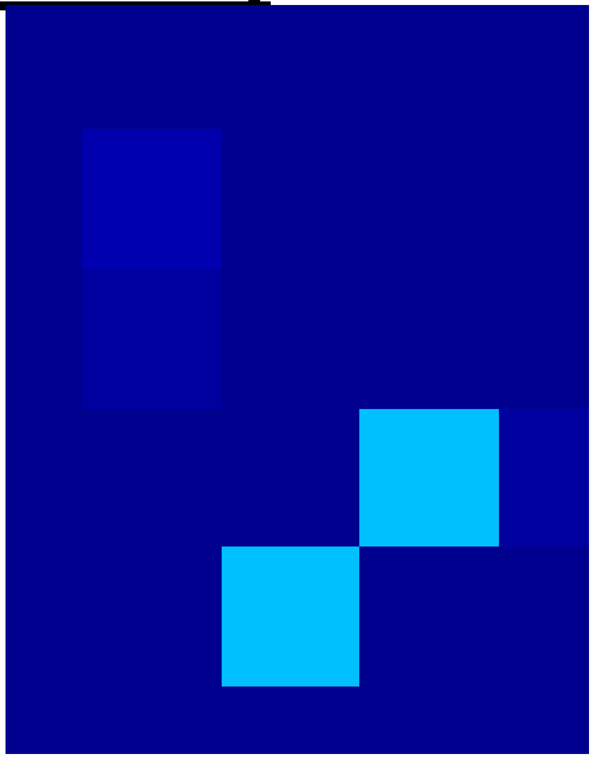




Example – Affymetrix Data Files

- Image file (.dat file)
- Probe results file (.cel file)

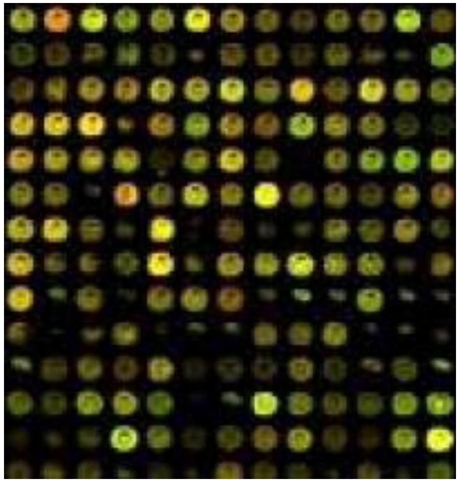


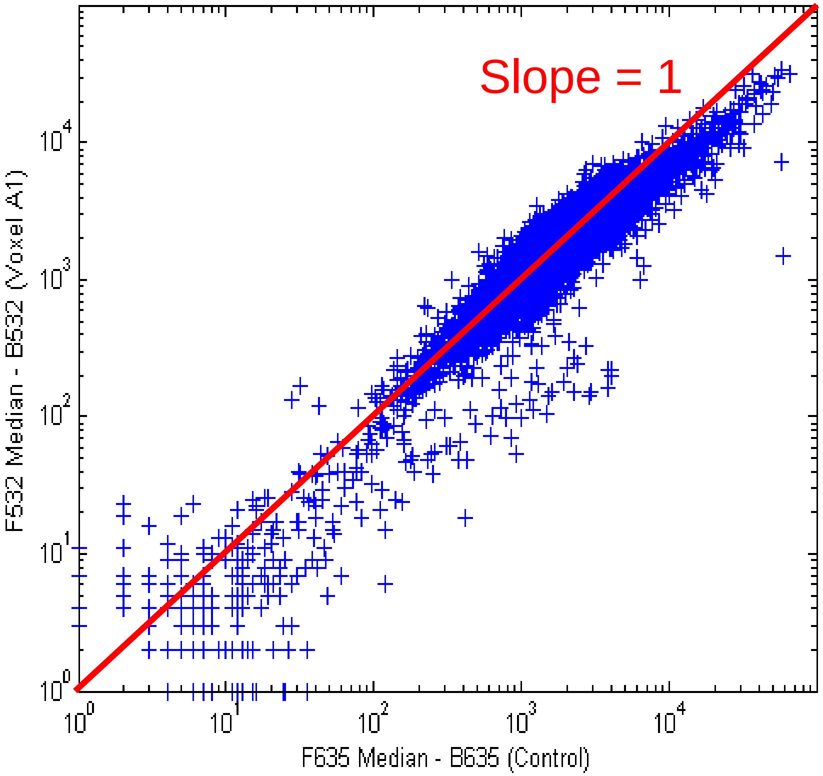


Scatter plots of the Microarrays

- **A measure of the actual expression levels**, i.e., differences between the median foreground and the median background for the red channel and green channel:

  "F635 Median - B635"
  "F532 Median - B532”

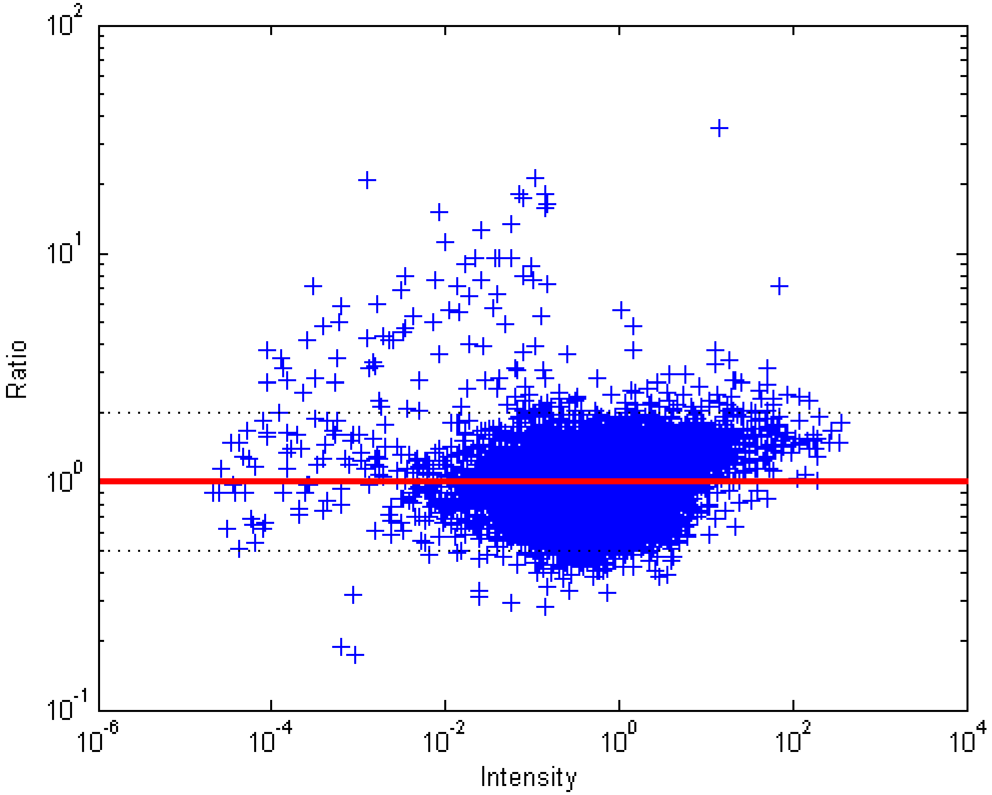




RI plots of the Microarrays

- RI (ratio-intensity) plot or MA plot

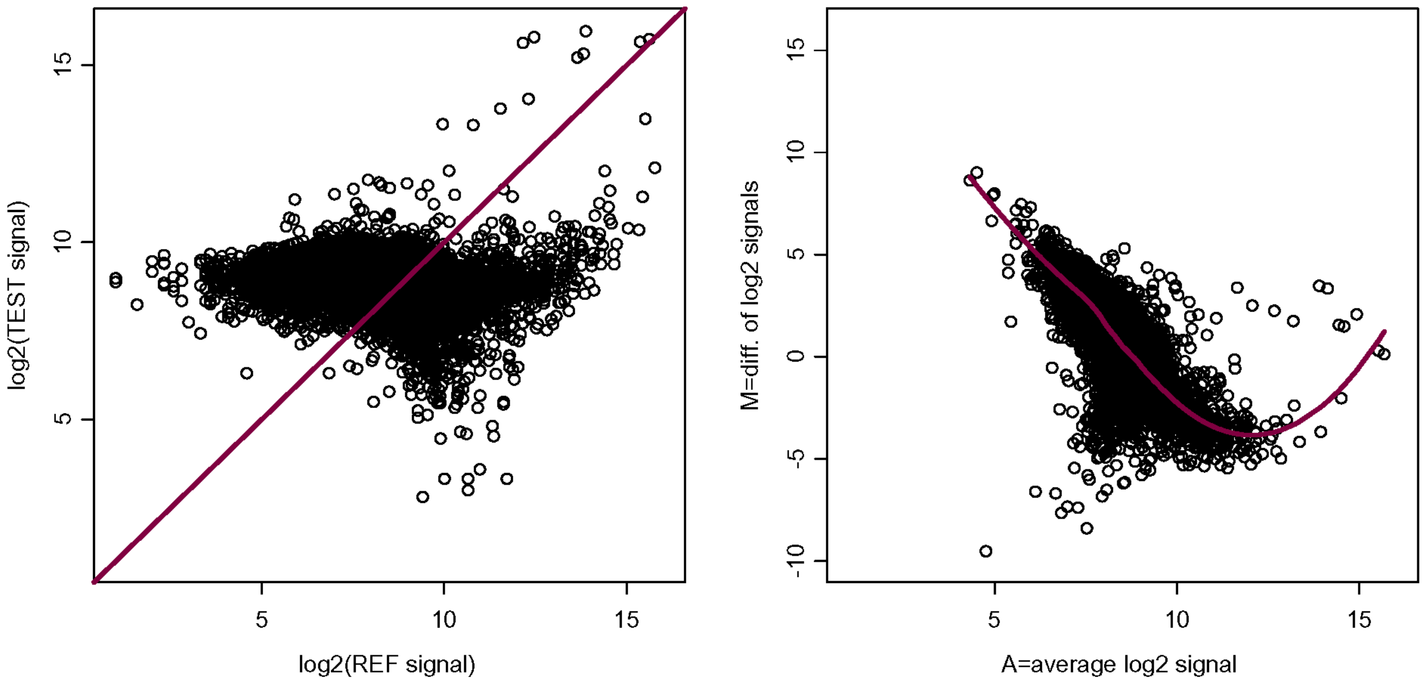
$$M = log(\frac{R}{G})$$



$$A = \frac{1}{2}(log(R) + log(G))$$

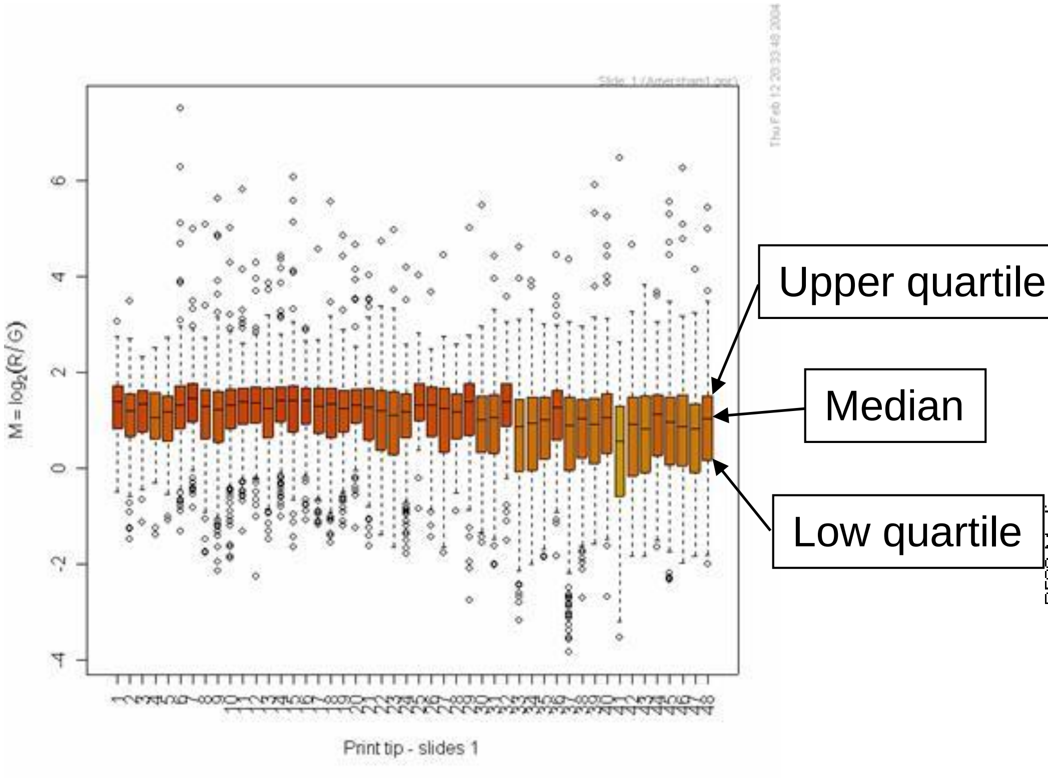
Scatter plots of the Microarrays

Bad Array Example



(McShane, NCI)

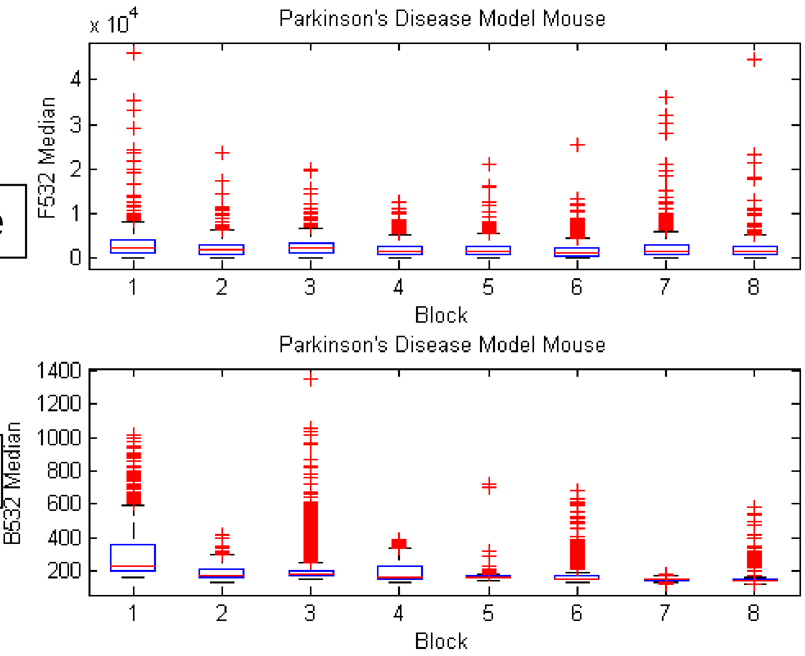
Box plot



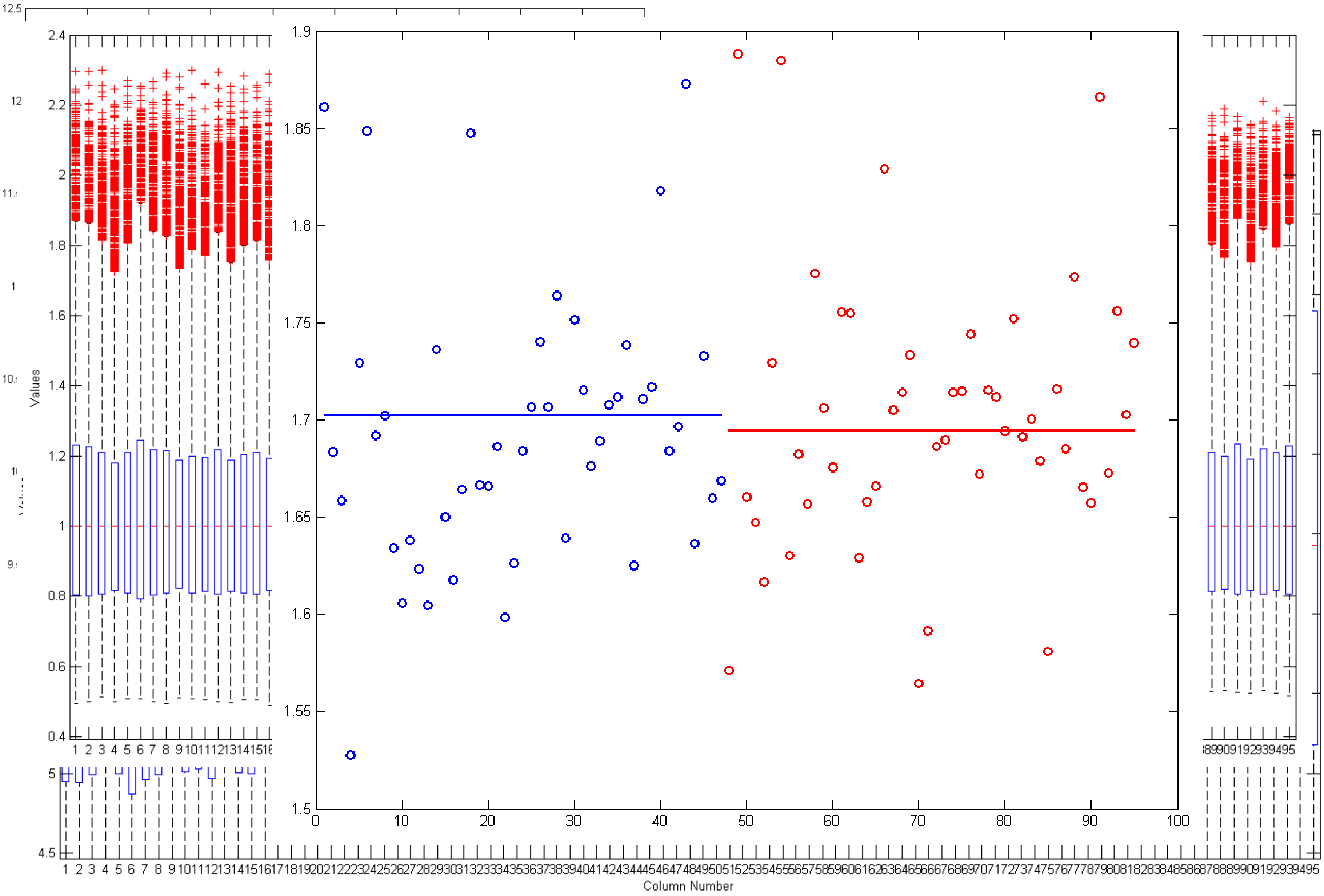
Upper quartile

Median

Low quartile



Example:



12.5
12
11.5
1
10.5
10
9.5
5
4.5
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
Values
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
1.9
1.85
1.8
1.75
1.7
1.65
1.6
1.55
1.5
0
10
20
30
40
50
60
70
80
90
100
89 90 91 92 93 94 95
Column Number

Normalization – microarray data is highly noisy

- **Intensity imbalance between RNA samples**
- **Affect all genes**
- **Not due to biology of samples, but due to technical reasons**
- **Reasons include difference in the settings of the photodetector voltage, imbalance in total amount of RNA in each sample, difference in uptaking of the dyes, etc.**
- **The objective is is to adjust the gene expression values of all genes so that the ones that are not really differentially expressed have similar values across the**

Two major issues to consider

- Which genes to use for normalization
- Which normalization algorithm to use

Which genes to use for normalization

- Housekeeping genes
  - **Genes involved in essential activities of cell maintenance and surviva**l, but not in cell function and proliferation
  - These genes will be similarly expressed in all samples.
  - **Difficult to identify – need to be confirmed**
  - **Affymetrix GeneChip provides a set of house keeping genes** based on a large set of tests on different tissues and were found to have low variability in these samples (but still no guarantee).

Which genes to use for normalization

- Spiked controls
  - Genes that are not usually found in the samples (both control and test sample). E.g., yeast gene in human tissue samples.

Which genes to use for normalization

- Using all genes
  - Simplest approach – use all adequately expressed genes for normalization
  - **The assumption is that the majority of genes on the array are housekeeping genes and the proportion of over expressed genes is similar to that of the under expressed genes.**
  - If the genes on the chip are specially selected, then this method will not work.

Two-color array normalization

- Intra-slide normalization
- Inter-slide for cDNA arrays

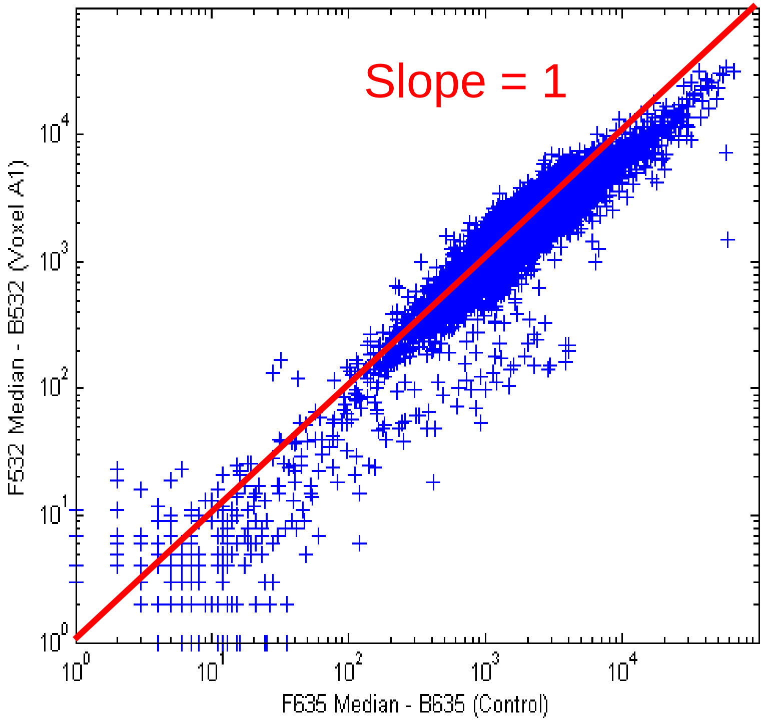
Normalization

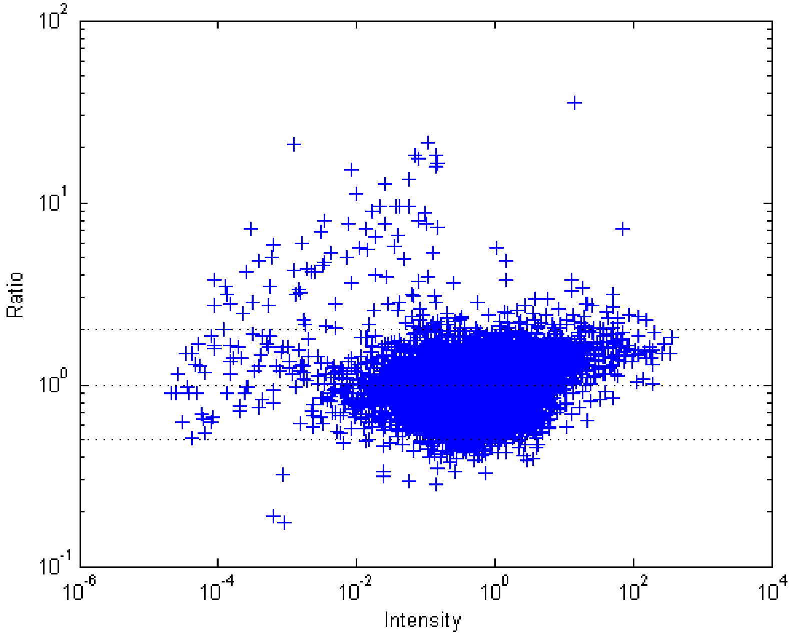
- **Linear (global) normalization**
  - Simplest but most consistent
  - Move the median to zero (slope 1 in scatter plot, this only changes the intersection)
  - No clear nonliearity or slope in MA plot

$$X_i^{norm} = k * X_i$$

$$c = log(k)$$

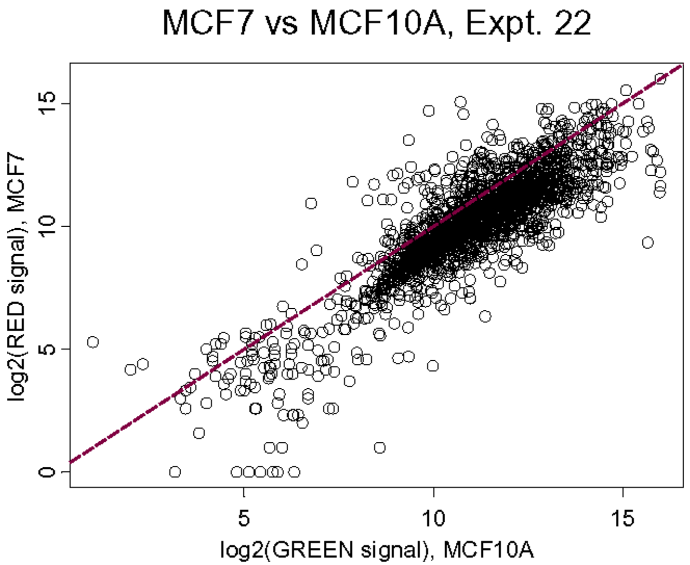
$$M_i^{norm} = log(X_i^{norm}) = c + M_i$$

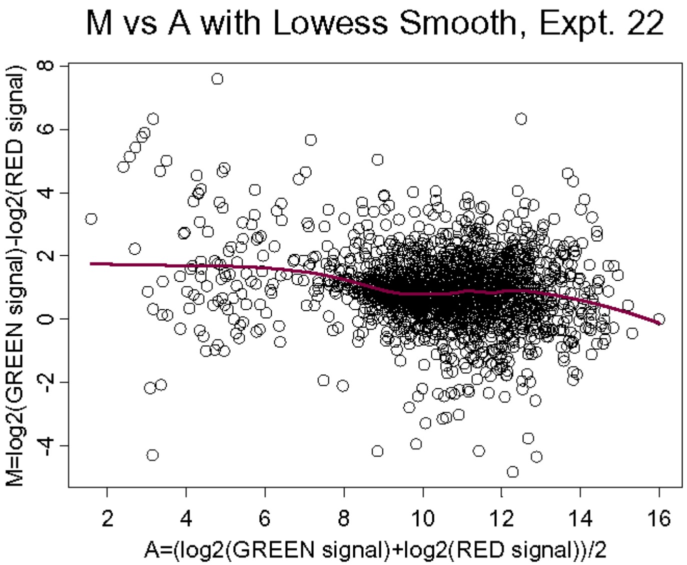




Normalization

- **Intensity-based (Loess/Lowess) normalization**
  - **Loess/Lowess fit**
  - **Overall magnitude of the spot intensity has an impact**

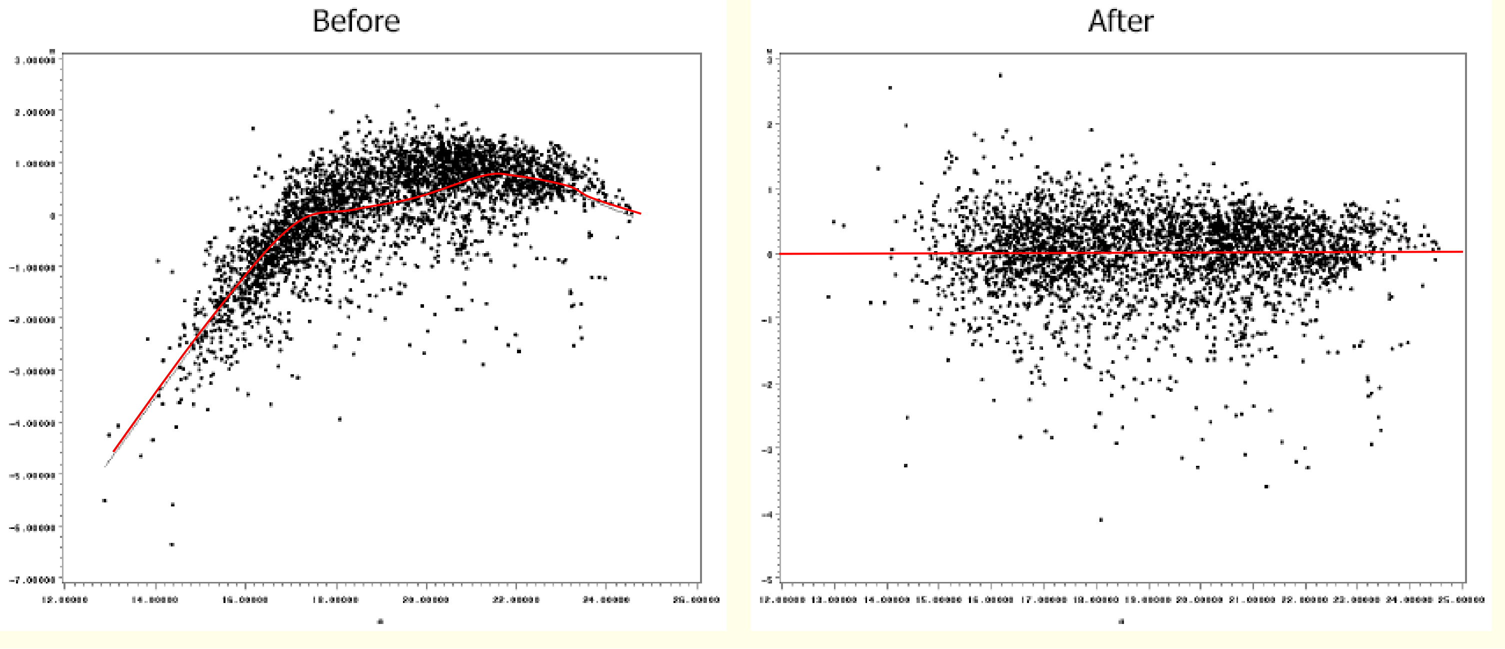




(McShane, NCI)

Normalization

- **Intensity-based normalization**
  - **“Straighten” the Lo(w)ess fit line in MA plot to horizontal line and move it to zero**



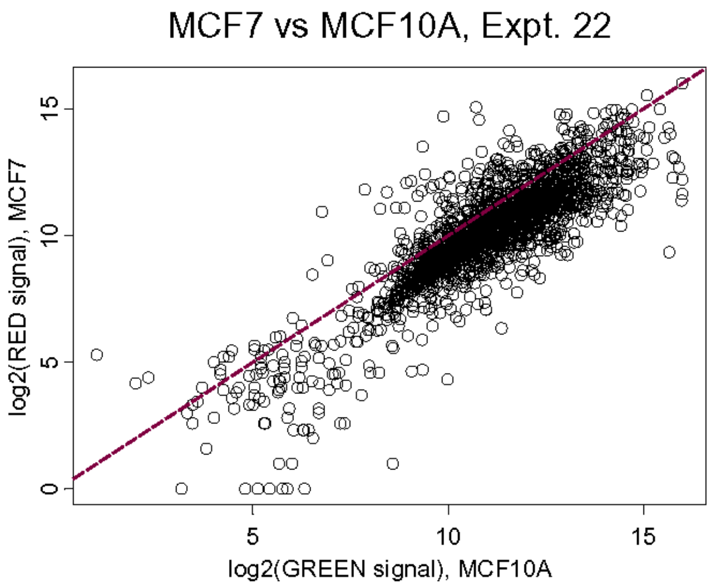
$$X_i^{norm} = k(A) * X_i$$

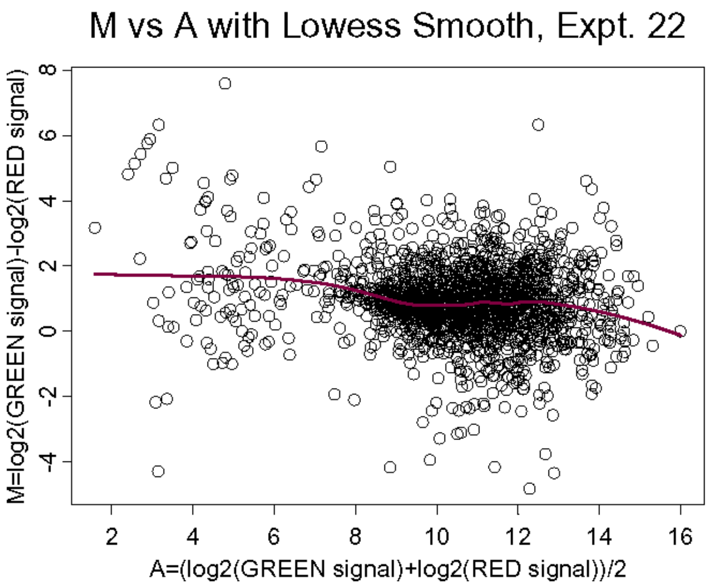
$$c(A) = log(k(A))$$

$$M_i^{norm} = log(X_i^{norm}) = c(A) + M_i$$

Normalization

- **Intensity-based (Lowess) normalization**
  - Nonlinear
  - Gene-by-gene, could introduce bias
  - Use only when there is a compelling

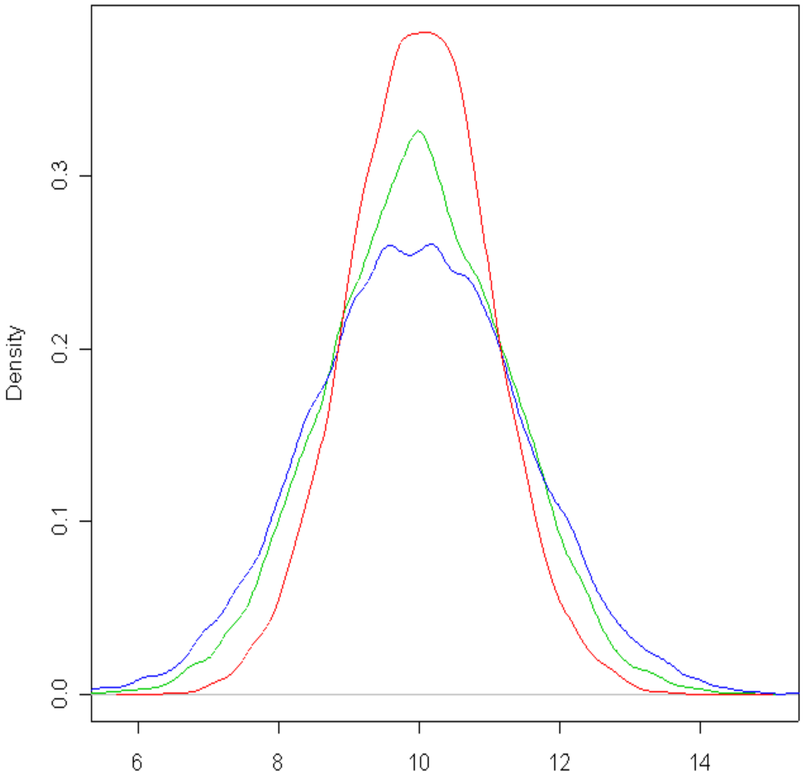

MCF7 vs MCF10A, Expt. 22


M vs A with Lowess Smooth, Expt. 22

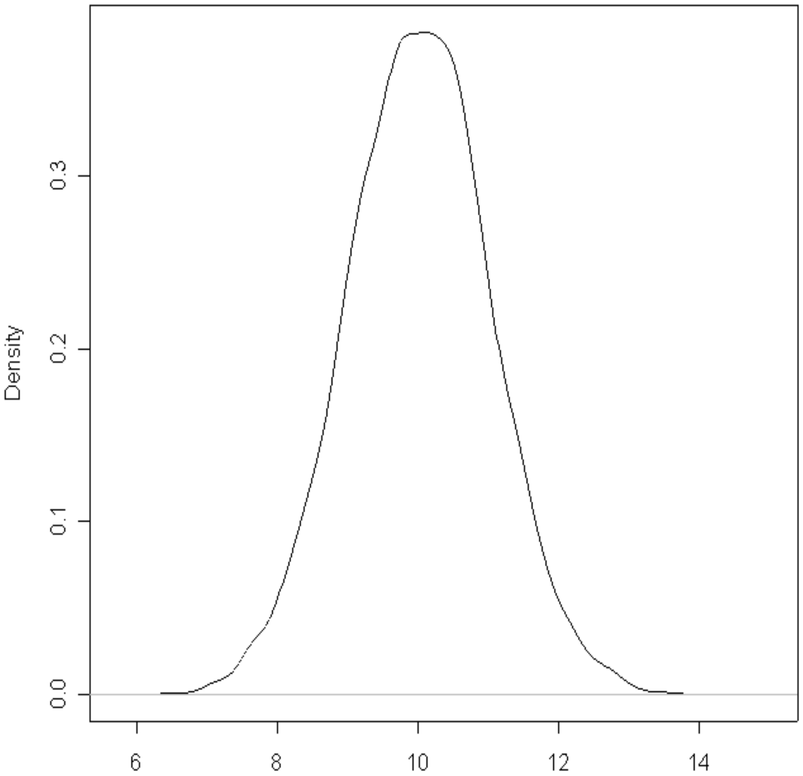
(McShane, NCI)

Normalization

- **Quantile normalization**
  - Nonlinear
  - Same intensity distribution



After Lowess normalization



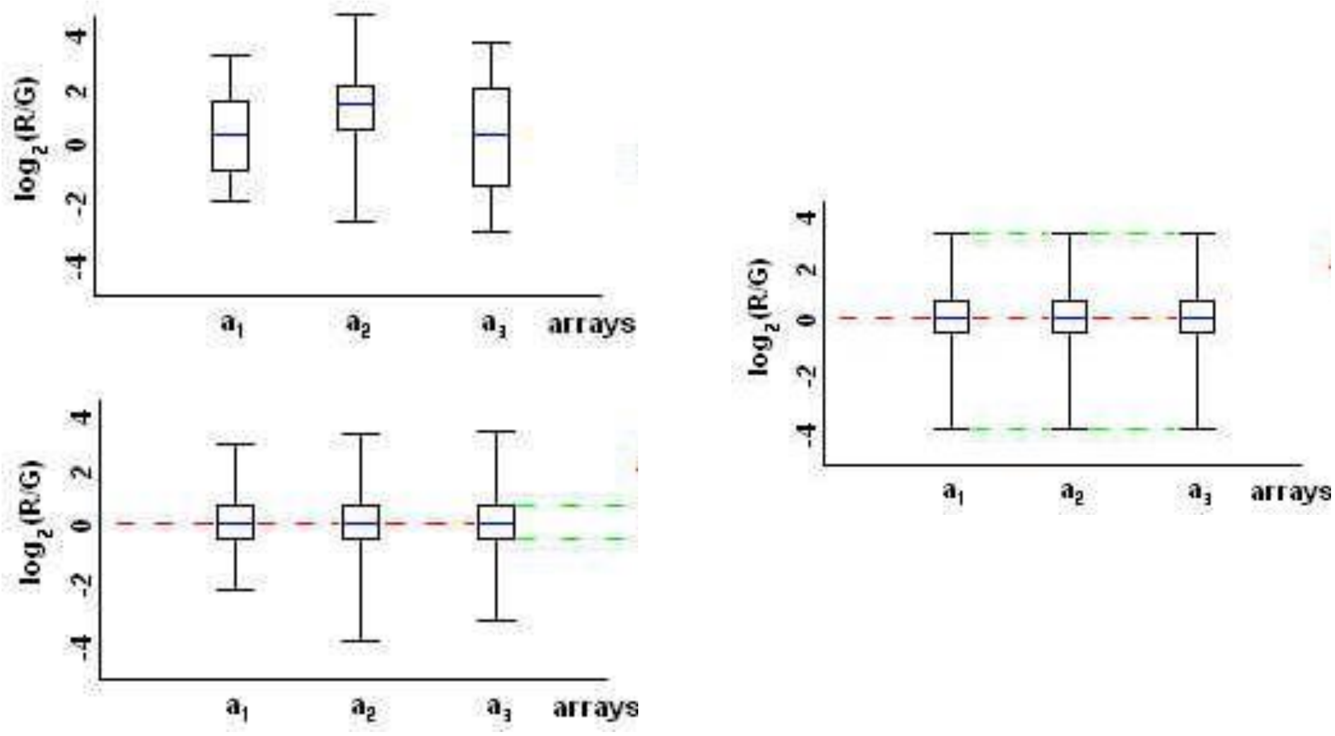
After quantile normalization

Normalization

- Location-based normalization
  - Background subtracted ratios on the array may vary in a predicable manner.
  - Sample uniformly across the chip
  - Nonlinear
  - Gene-by-gene, could introduce bias
  - Use only when there is a compelling reason
- Other normalization method
  - Combination of location and intensity-based normalization

Normalization

- **Which normalization algorithm to use**
  - Inter-slide normalization
  - Not just for Affymetrix arrays



Normalization

- Linear (global) – the chips have equal median (or mean) intensity
- Intensity-based (Lowess) – the chips have equal medians (means) at all intensity values
- Quantile – the chips have identical intensity distribution
- Quantile is the “best” in term of normalizing the data to desired distribution, however it also changes the gene expression level individually
- Potential issues - overfitting

Affymetrix array normalization

- Inter-slide normalization only
- Probe-level normalization
- Affymetrix MicroArray Suite (MAS) 5.0
- Robust Multiarray Average (RMA)

- Quantile
- GC-RMA

Affymetrix array normalization

- **Inter-slide normalization only**
- **Probe-level normalization**
- **Affymetrix MicroArray Suite (MAS) 4.0**
    - Simple subtraction of MM from PM
    - Use only probes within 3 times of SD of PM-MM to exclude outliers
    - Not robust
- **MAS 5.0**
    - Use weight (Turkey Biweight Estimate) for each probe based on its intensity difference from the mean
    - Log transformed data for mean (geometric mean)
    - Robust

Affymetrix array normalization

- **Robust Multiarray Average (RMA)**
  - Background correction on each chip.
    - Assuming strictly positive distribution. No negative numbers
    - Do NOT use MM information
  - Normalization (inter-chip).
    - Quantile
  - Probe level intensity calculation.
    - Linear model for signal, affinity, and noise.
  - Probe set summarization.
    - Combine probes for one probeset into a single number
    - Median polishing (chip to its median, gene to its median, iterate and converge)

Affymetrix array normalization

- **GC-Robust Multiarray Average (GC-RMA)**
  - Correct back ground noise and non-specific binding
  - Affinity computed from position specific base effect
  - MM information is used (subtracted from PM after correction)

Affymetrix array normalization

- **RMA/GCRMA pros and cons (comparing to MAS5.0)**
  - Less variance at low expression values
  - Less false positives
  - Consistent fold change estimates
  - More false negatives, especially for low-expression level probes
  - Quality control after normalization is difficulty
  - Quantile normalization may overfit and hide real differences

- Introduction to gene expression microarray
  - A middle-man’s approach
  - Applications of microarray
- Microarray data processing/analysis workflow
  - Data format and visualization
  - Data normalization
    - Two-color array
    - Affymetrix array
- **Software and databases**

Microarray analysis software

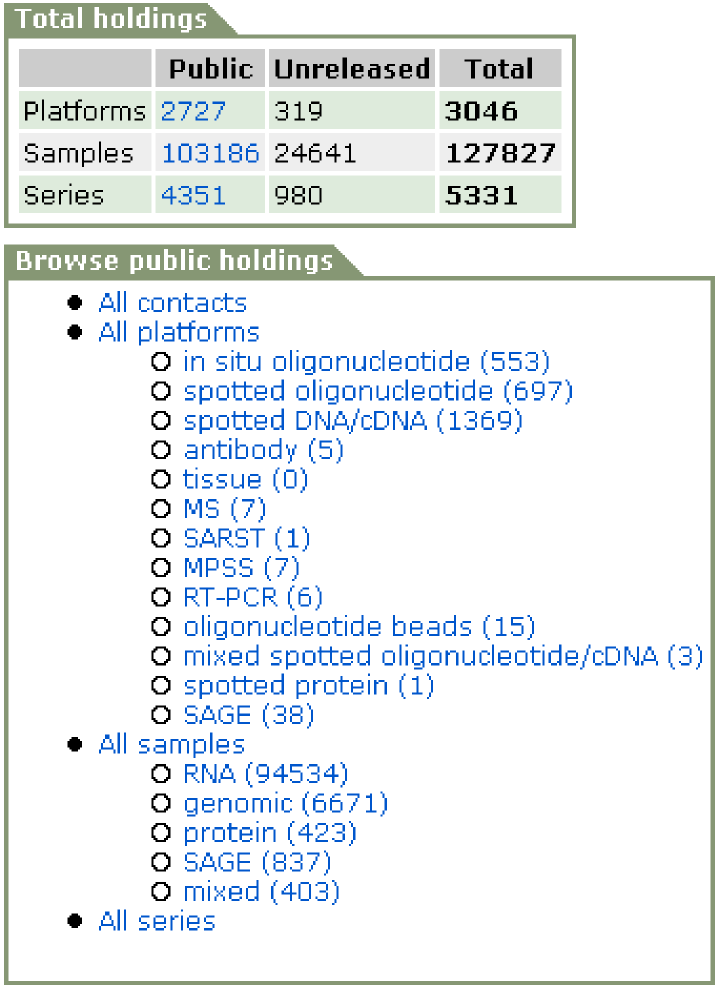
- Open source R
- Bioconductor
- BRBArray tools (NCI biometric research branch)
- Matlab Bioinformatics Toolbox
- Affymetrix Expression Console
- DChip
- GeneSpring
- Partek
- …

Microarray Databases

- **Gene Expression Ominbus (GEO) database – NCBI**
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
- **EMBL-EBI microarray database (ArrayExpress)**
  - http://www.ebi.ac.uk/Databases/microarray.html
- **Stanford Microarray Database (SMD)**
  - http://genome-www5.stanford.edu/
- **caARRAY sites**
- **The Cancer Genome Atlas (TCGA)**
- **Other specialized, regional and aggregated databases**
  - http://psi081.ba.ars.usda.gov/SGMD/
  - http://www.oncomine.org/main/index.jsp
  - http://ihome.cuhk.edu.hk/~b400559/arraysoft_public.html
  - …

Gene Expression Omnibus (GEO)

- http://www.ncbi.nlm.nih.gov/projects/geo/query/browse.cgi



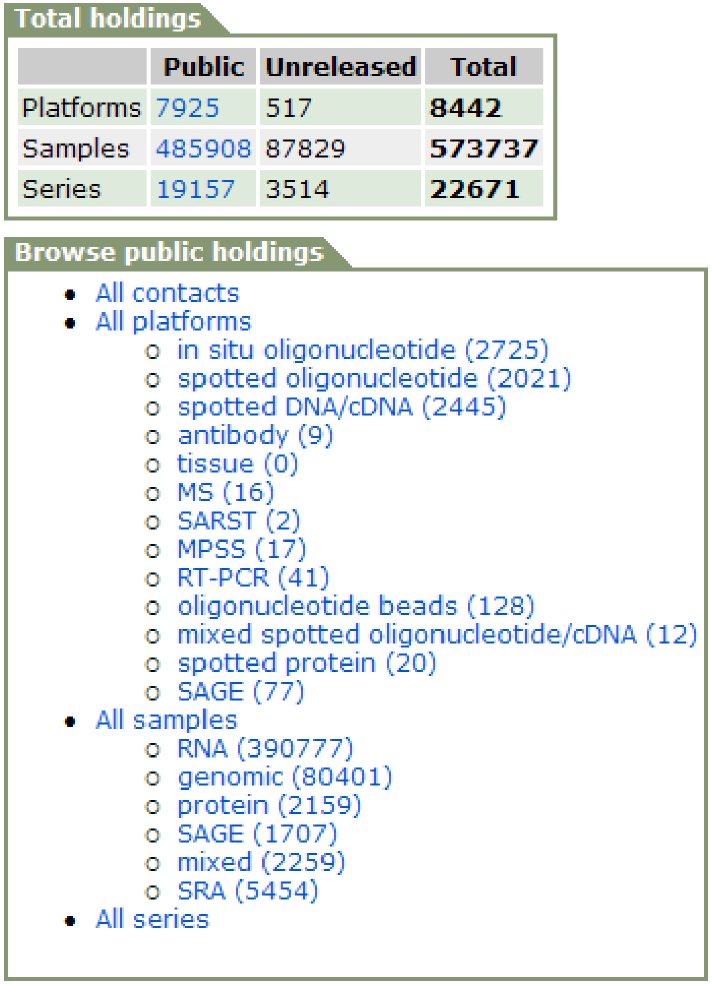
Total holdings

	Public	Unreleased	Total
Platforms	2727	319	**3046**
Samples	103186	24641	**127827**
Series	4351	980	**5331**

Browse public holdings

- All contacts
- All platforms
  - in situ oligonucleotide (553)
  - spotted oligonucleotide (697)
  - spotted DNA/cDNA (1369)
  - antibody (5)
  - tissue (0)
  - MS (7)
  - SARST (1)
  - MPSS (7)
  - RT-PCR (6)
  - oligonucleotide beads (15)
  - mixed spotted oligonucleotide/cDNA (3)
  - spotted protein (1)
  - SAGE (38)
- All samples
  - RNA (94534)
  - genomic (6671)
  - protein (423)
  - SAGE (837)
  - mixed (403)
- All series

Oct. 2006



Total holdings

	Public	Unreleased	Total
Platforms	7925	517	**8442**
Samples	485908	87829	**573737**
Series	19157	3514	**22671**

Browse public holdings

- All contacts
- All platforms
  - in situ oligonucleotide (2725)
  - spotted oligonucleotide (2021)
  - spotted DNA/cDNA (2445)
  - antibody (9)
  - tissue (0)
  - MS (16)
  - SARST (2)
  - MPSS (17)
  - RT-PCR (41)
  - oligonucleotide beads (128)
  - mixed spotted oligonucleotide/cDNA (12)
  - spotted protein (20)
  - SAGE (77)
- All samples
  - RNA (390777)
  - genomic (80401)
  - protein (2159)
  - SAGE (1707)
  - mixed (2259)
  - SRA (5454)
- All series

Oct. 2010

Gene Expression Omnibus (GEO)

- **GEO Profiles**
  This database stores individual gene expression and molecular abundance profiles assembled from the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics. GEO Profiles facilitates powerful searching and linking to additional information sources.
- **GEO DataSets**
  This database stores curated gene expression and molecular abundance DataSets assembled from the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

(From GEO website)

Gene Expression Omnibus (GEO)

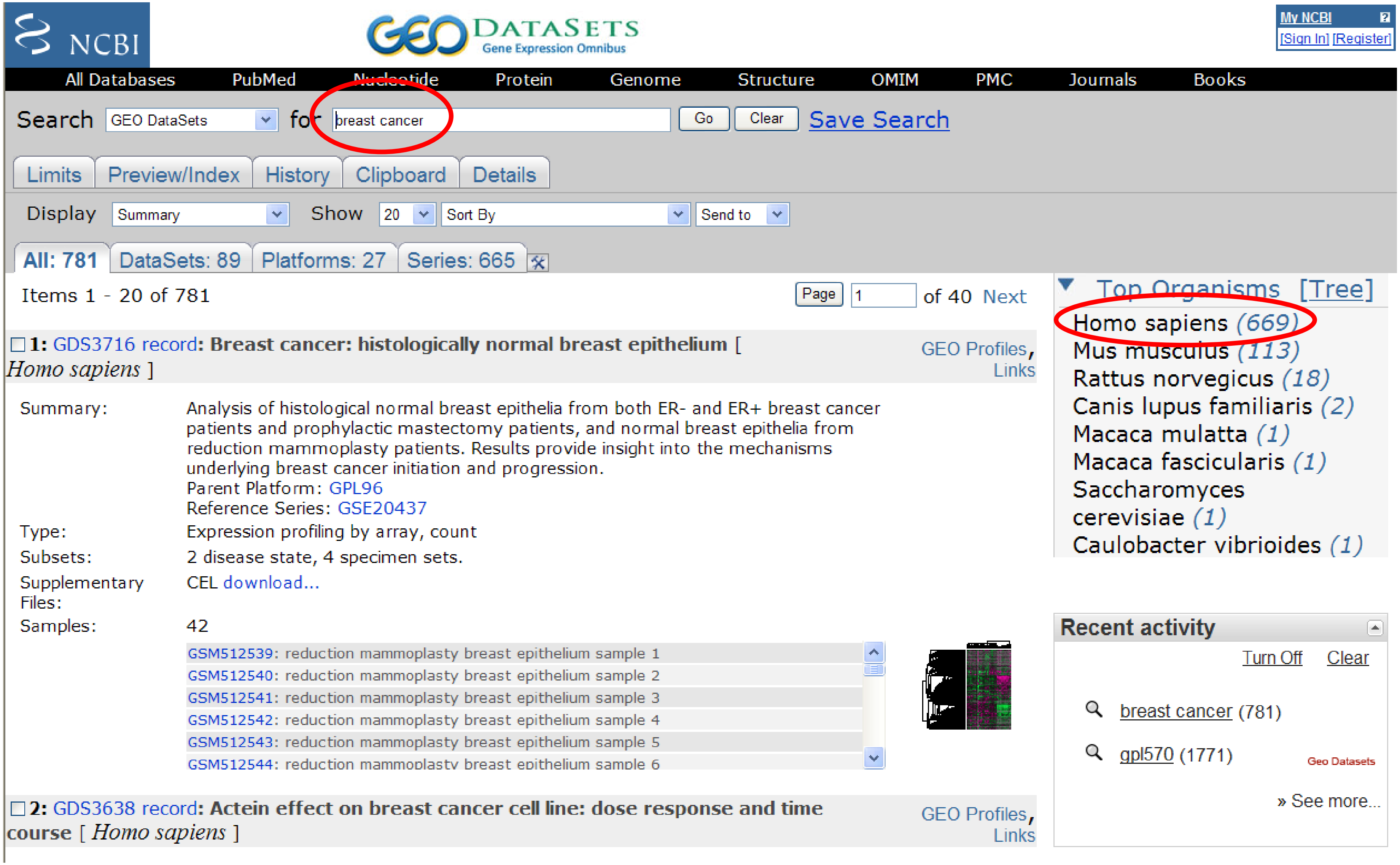
- **GPL**
  - A Platform record describes the list of elements on the array (e.g., cDNAs, oligonucleotide probesets, ORFs, antibodies) or the list of elements that may be detected and quantified in that experiment (e.g., SAGE tags, peptides). Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters.

- **GSM**
  - A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series.

Gene Expression Omnibus (GEO)

- **GSE**
  - A Series record defines a set of related Samples considered to be part of a group, how the Samples are related, and if and how they are ordered. A Series provides a focal point and description of the experiment as a whole. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx).
- **GDS**
  - GEO DataSets (GDS) are curated sets of GEO Sample data. A GDS record represents a collection of biologically and statistically comparable GEO Samples and forms the basis of GEO's suite of data display and analysis tools. Samples within a GDS refer to the same Platform, that is, they share a common set of probe elements. Value measurements for each Sample within a GDS are assumed to be calculated in an equivalent manner, that is, considerations such as background processing and normalization are consistent across the dataset. Information reflecting experimental design is provided through GDS

GEO Datasets



Search GEO DataSets for breast cancer

Display Summary Show 20 Sort By Send to

All: 781 DataSets: 89 Platforms: 27 Series: 665

Items 1 - 20 of 781 Page 1 of 40 Next

Top Organisms [Tree]

Homo sapiens (669)
Mus musculus (113)
Rattus norvegicus (18)
Canis lupus familiaris (2)
Macaca mulatta (1)
Macaca fascicularis (1)
Saccharomyces cerevisiae (1)
Caulobacter vibrioides (1)

□ **1:** GDS3716 record: **Breast cancer: histologically normal breast epithelium** [*Homo sapiens*] GEO Profiles, Links

Summary: Analysis of histological normal breast epithelia from both ER- and ER+ breast cancer patients and prophylactic mastectomy patients, and normal breast epithelia from reduction mammoplasty patients. Results provide insight into the mechanisms underlying breast cancer initiation and progression.
Parent Platform: GPL96
Reference Series: GSE20437

Type: Expression profiling by array, count

Subsets: 2 disease state, 4 specimen sets.

Supplementary Files: CEL download...

Samples: 42

GSM512539: reduction mammoplasty breast epithelium sample 1
GSM512540: reduction mammoplasty breast epithelium sample 2
GSM512541: reduction mammoplasty breast epithelium sample 3
GSM512542: reduction mammoplasty breast epithelium sample 4
GSM512543: reduction mammoplasty breast epithelium sample 5
GSM512544: reduction mammoplasty breast epithelium sample 6

□ **2:** GDS3638 record: **Actein effect on breast cancer cell line: dose response and time course** [*Homo sapiens*] GEO Profiles, Links

Recent activity

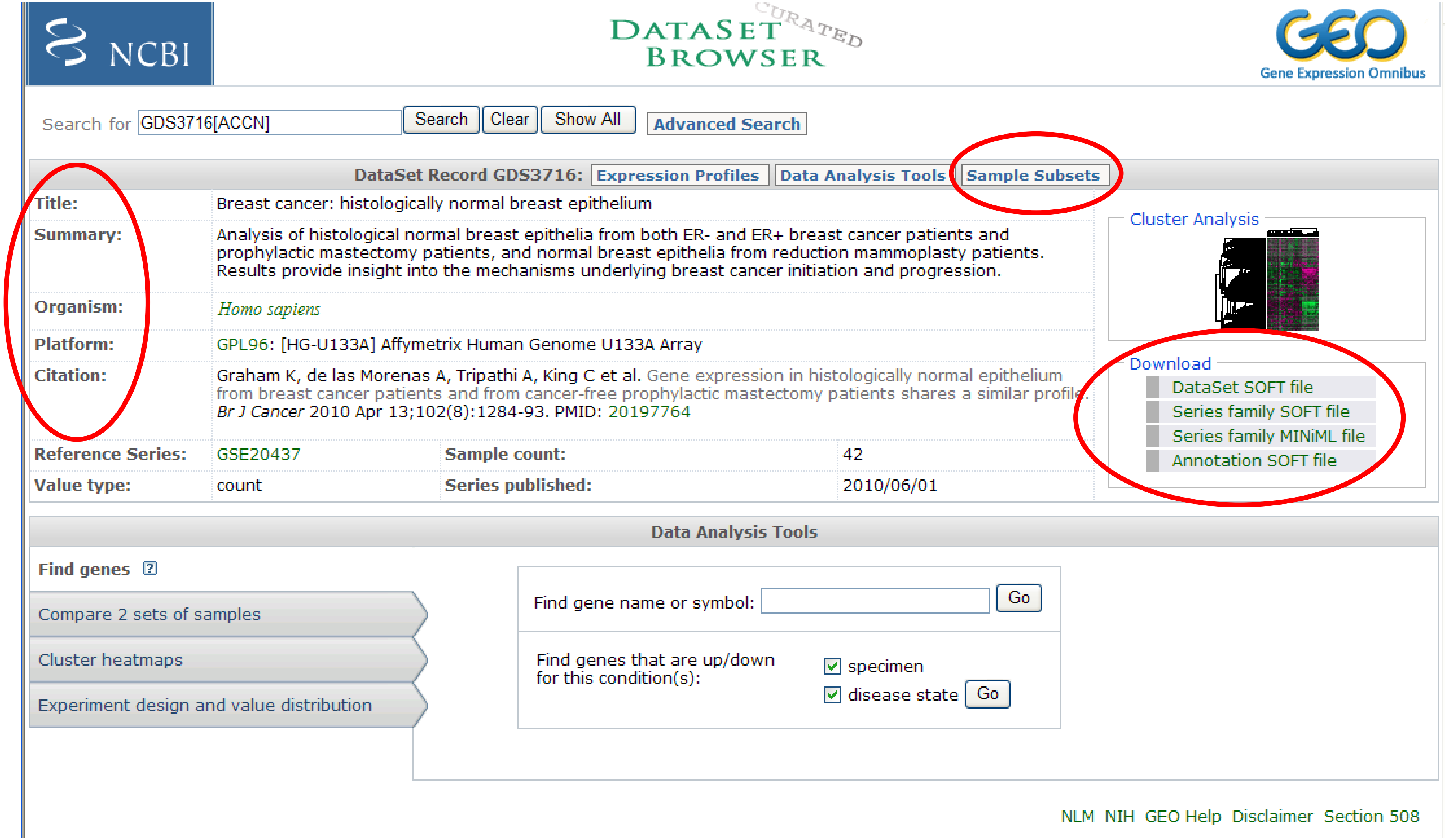
Turn Off Clear

breast cancer (781)

gpl570 (1771) Geo Datasets

» See more...

GEO Datasets



Search for GDS3716[ACCN]

DataSet Record GDS3716:

Title:	Breast cancer: histologically normal breast epithelium		
Summary:	Analysis of histological normal breast epithelia from both ER- and ER+ breast cancer patients and prophylactic mastectomy patients, and normal breast epithelia from reduction mammoplasty patients. Results provide insight into the mechanisms underlying breast cancer initiation and progression.		
Organism:	*Homo sapiens*		
Platform:	GPL96: [HG-U133A] Affymetrix Human Genome U133A Array		
Citation:	Graham K, de las Morenas A, Tripathi A, King C et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *Br J Cancer* 2010 Apr 13;102(8):1284-93. PMID: 20197764		
Reference Series:	GSE20437	**Sample count:**	42
Value type:	count	**Series published:**	2010/06/01

Data Analysis Tools

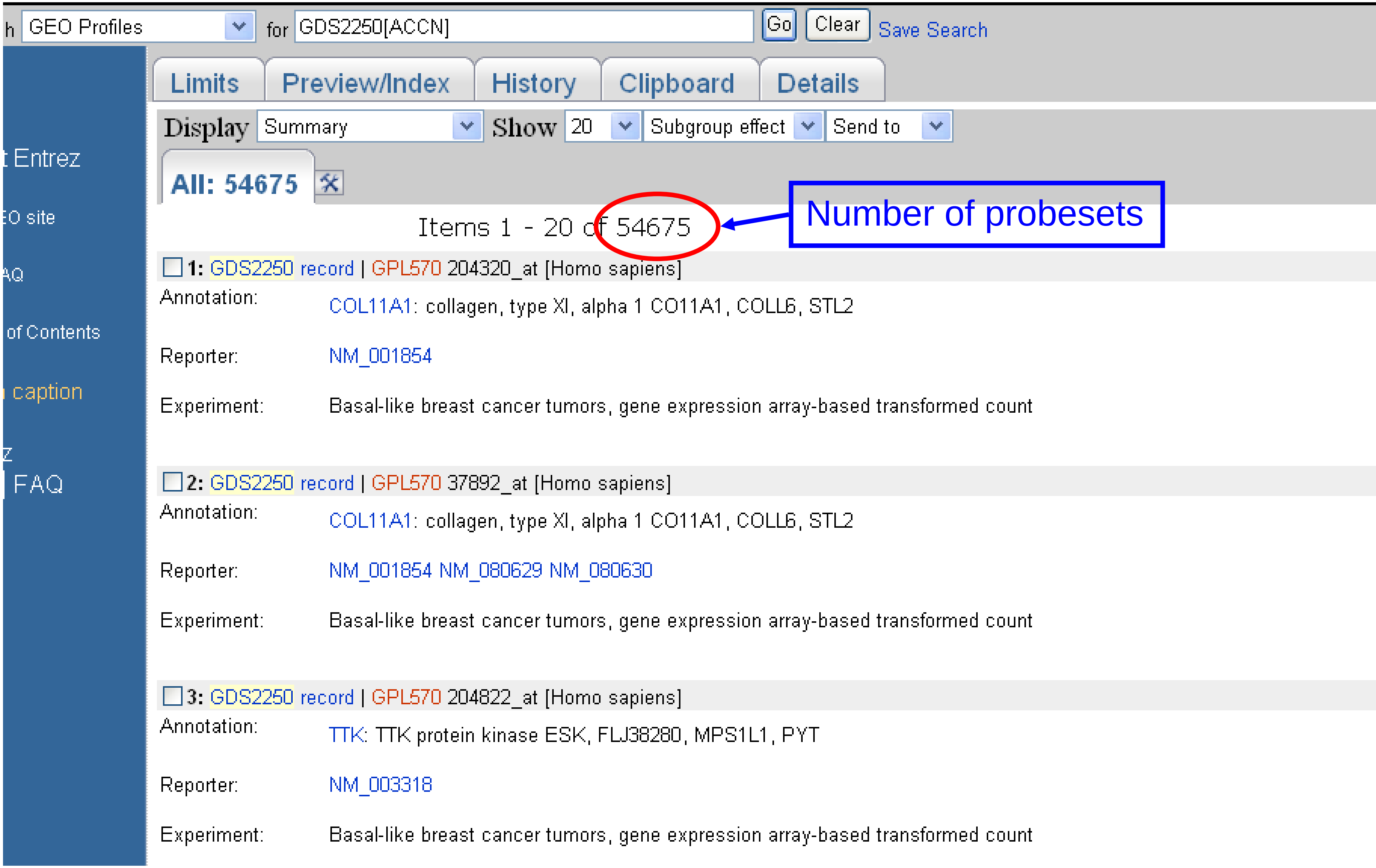
Find genes ?

Find gene name or symbol:

Find genes that are up/down for this condition(s):

☑ specimen
☑ disease state

GEO Profiles



Number of probesets

Items 1 - 20 of 54675

□ **1:** GDS2250 record | GPL570 204320_at [Homo sapiens]

Annotation: COL11A1: collagen, type XI, alpha 1 CO11A1, COLL6, STL2

Reporter: NM_001854

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

□ **2:** GDS2250 record | GPL570 37892_at [Homo sapiens]

Annotation: COL11A1: collagen, type XI, alpha 1 CO11A1, COLL6, STL2

Reporter: NM_001854 NM_080629 NM_080630

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

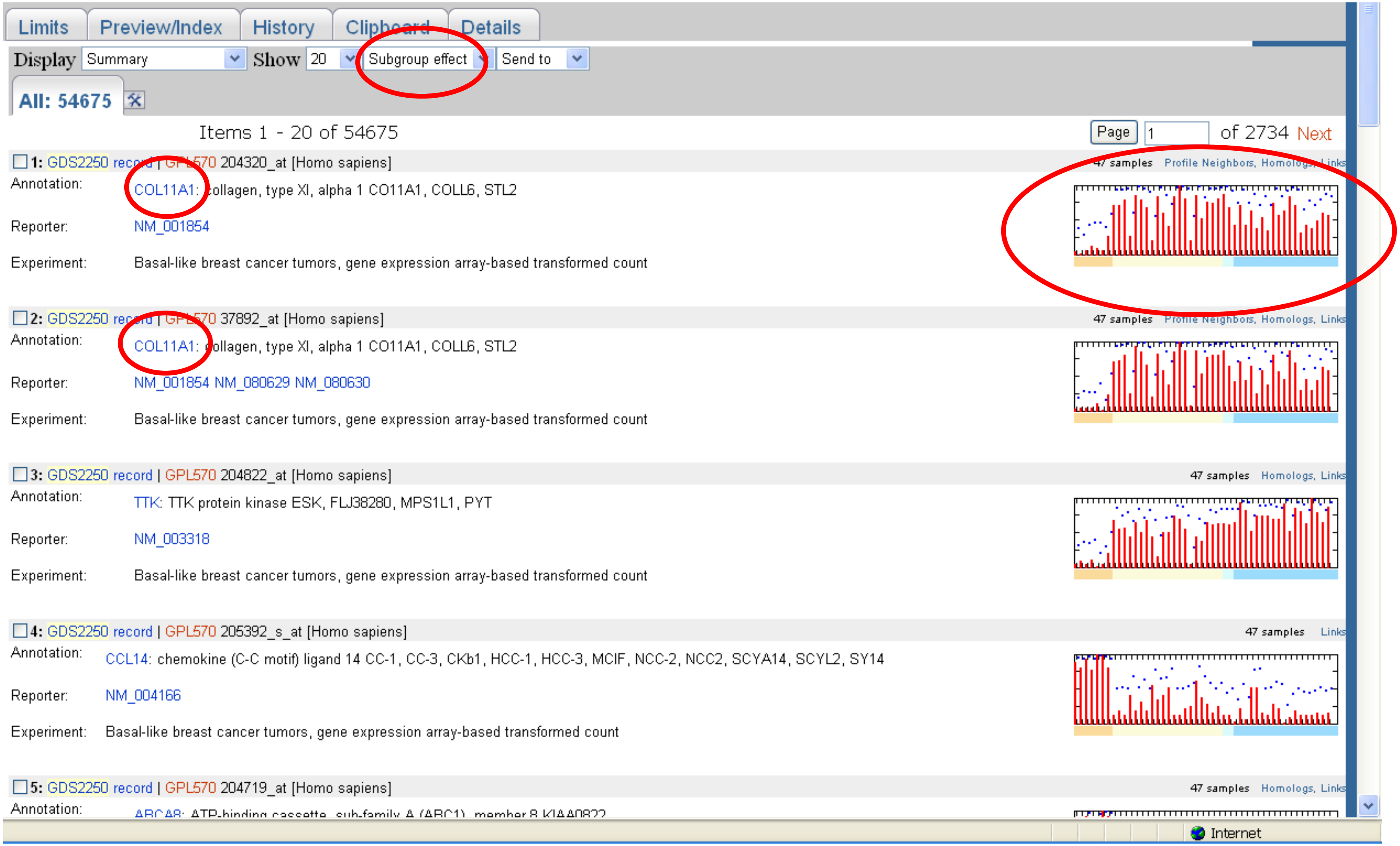
□ **3:** GDS2250 record | GPL570 204822_at [Homo sapiens]

Annotation: TTK: TTK protein kinase ESK, FLJ38280, MPS1L1, PYT

Reporter: NM_003318

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

GEO Profiles



Items 1 - 20 of 54675

☐ 1: GDS2250 record | GPL570 204320_at [Homo sapiens] — 47 samples — Profile Neighbors, Homologs, Links

Annotation: COL11A1: collagen, type XI, alpha 1 CO11A1, COLL6, STL2

Reporter: NM_001854

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

☐ 2: GDS2250 record | GPL570 37892_at [Homo sapiens] — 47 samples — Profile Neighbors, Homologs, Links

Annotation: COL11A1: collagen, type XI, alpha 1 CO11A1, COLL6, STL2

Reporter: NM_001854 NM_080629 NM_080630

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

☐ 3: GDS2250 record | GPL570 204822_at [Homo sapiens] — 47 samples — Homologs, Links

Annotation: TTK: TTK protein kinase ESK, FLJ38280, MPS1L1, PYT

Reporter: NM_003318

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

☐ 4: GDS2250 record | GPL570 205392_s_at [Homo sapiens] — 47 samples — Links

Annotation: CCL14: chemokine (C-C motif) ligand 14 CC-1, CC-3, CKb1, HCC-1, HCC-3, MCIF, NCC-2, NCC2, SCYA14, SCYL2, SY14

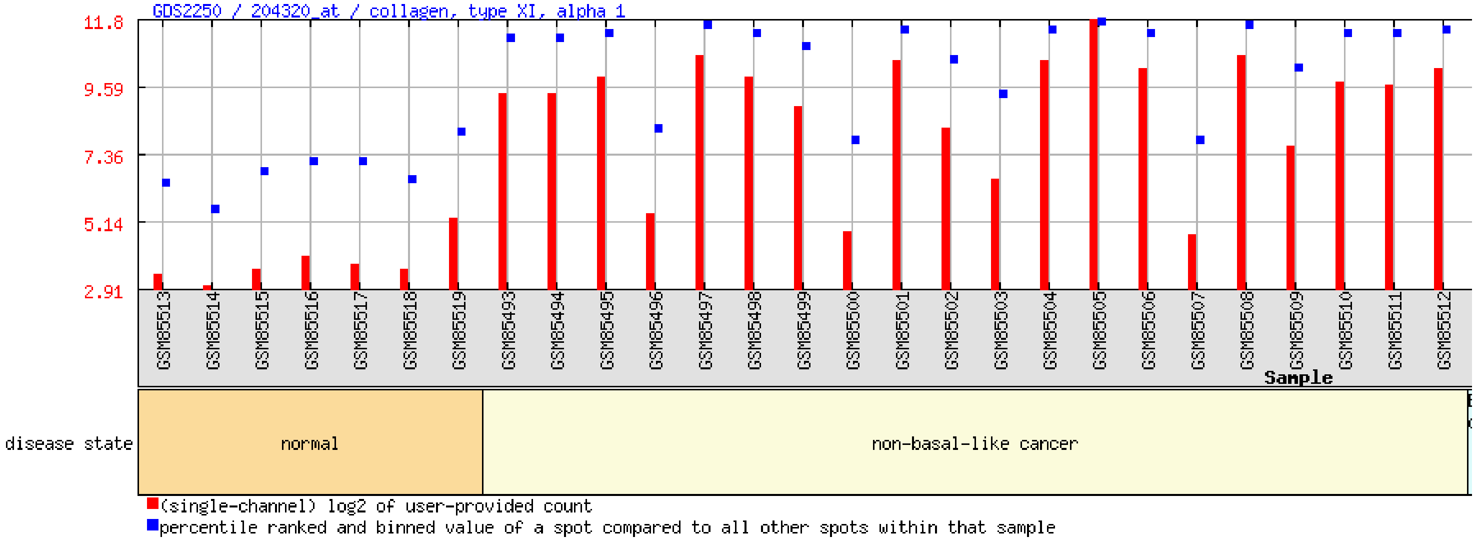
Reporter: NM_004166

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

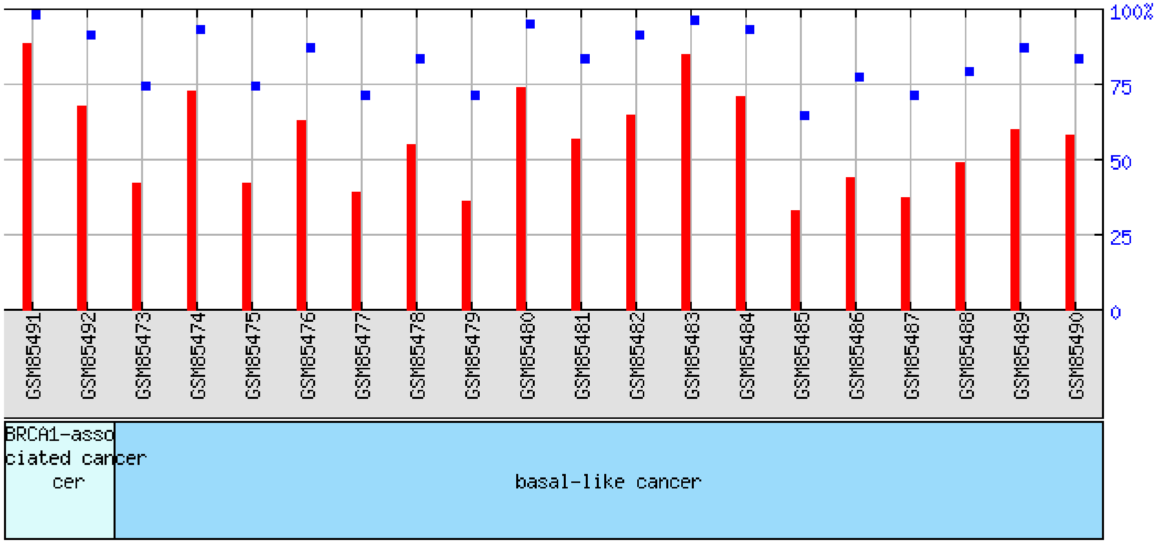
☐ 5: GDS2250 record | GPL570 204719_at [Homo sapiens] — 47 samples — Homologs, Links

Annotation: ABCA8: ATP-binding cassette, sub-family A (ABC1), member 8 KIAA0822

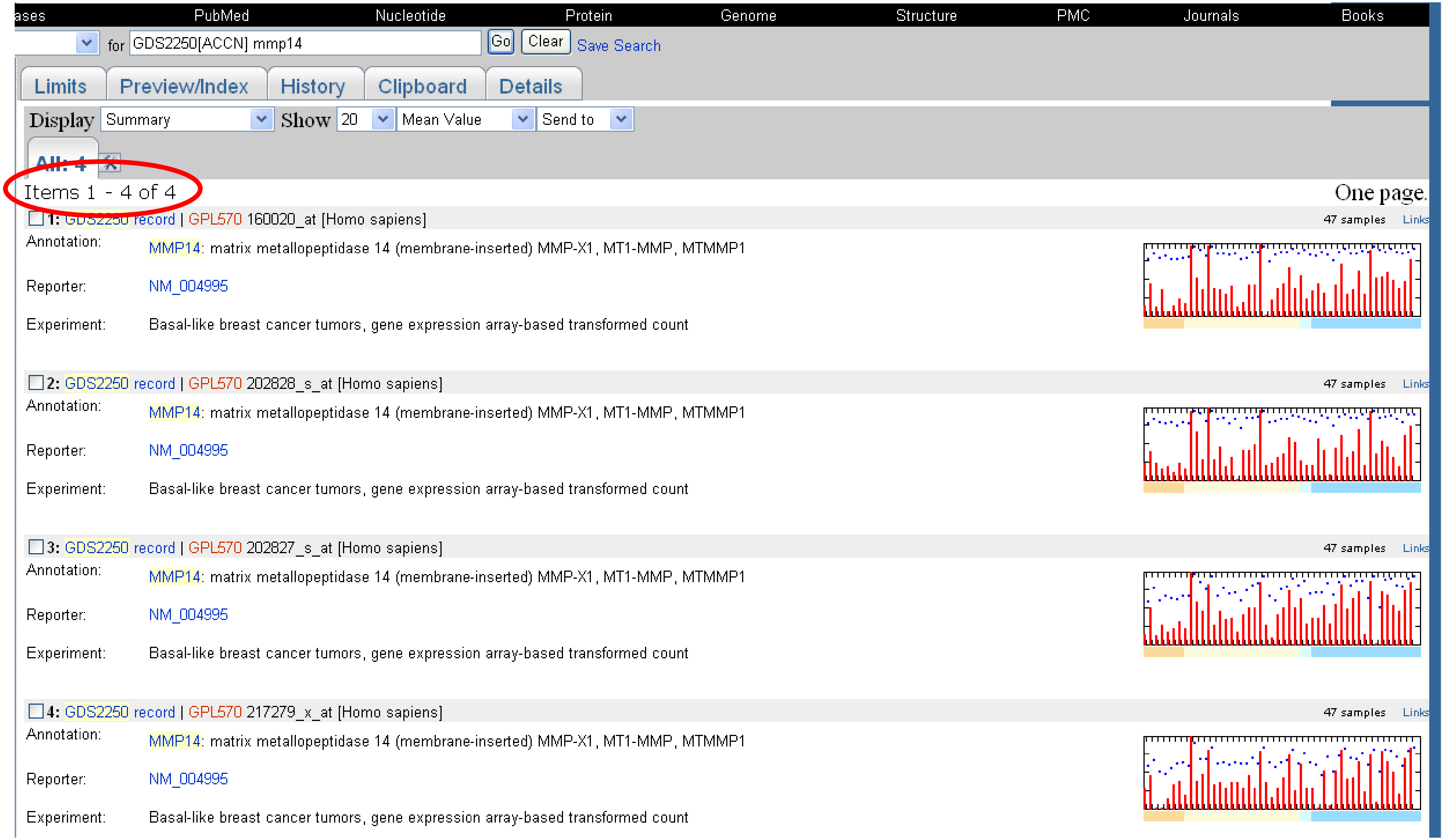
GEO Profiles



- Left y-axis is (supposed to be) log two based (must check to verify) expression level.
- Right y-axis is the percentile of this expression level in the entire chip.
- All the chips are normalized.



GEO Profiles



for GDS2250[ACCN] mmp14

Display Summary Show 20 Mean Value Send to

Items 1 - 4 of 4

One page.

□ 1: GDS2250 record | GPL570 160020_at [Homo sapiens] 47 samples Links

Annotation: MMP14: matrix metallopeptidase 14 (membrane-inserted) MMP-X1, MT1-MMP, MTMMP1

Reporter: NM_004995

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

□ 2: GDS2250 record | GPL570 202828_s_at [Homo sapiens] 47 samples Links

Annotation: MMP14: matrix metallopeptidase 14 (membrane-inserted) MMP-X1, MT1-MMP, MTMMP1

Reporter: NM_004995

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

□ 3: GDS2250 record | GPL570 202827_s_at [Homo sapiens] 47 samples Links

Annotation: MMP14: matrix metallopeptidase 14 (membrane-inserted) MMP-X1, MT1-MMP, MTMMP1

Reporter: NM_004995

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

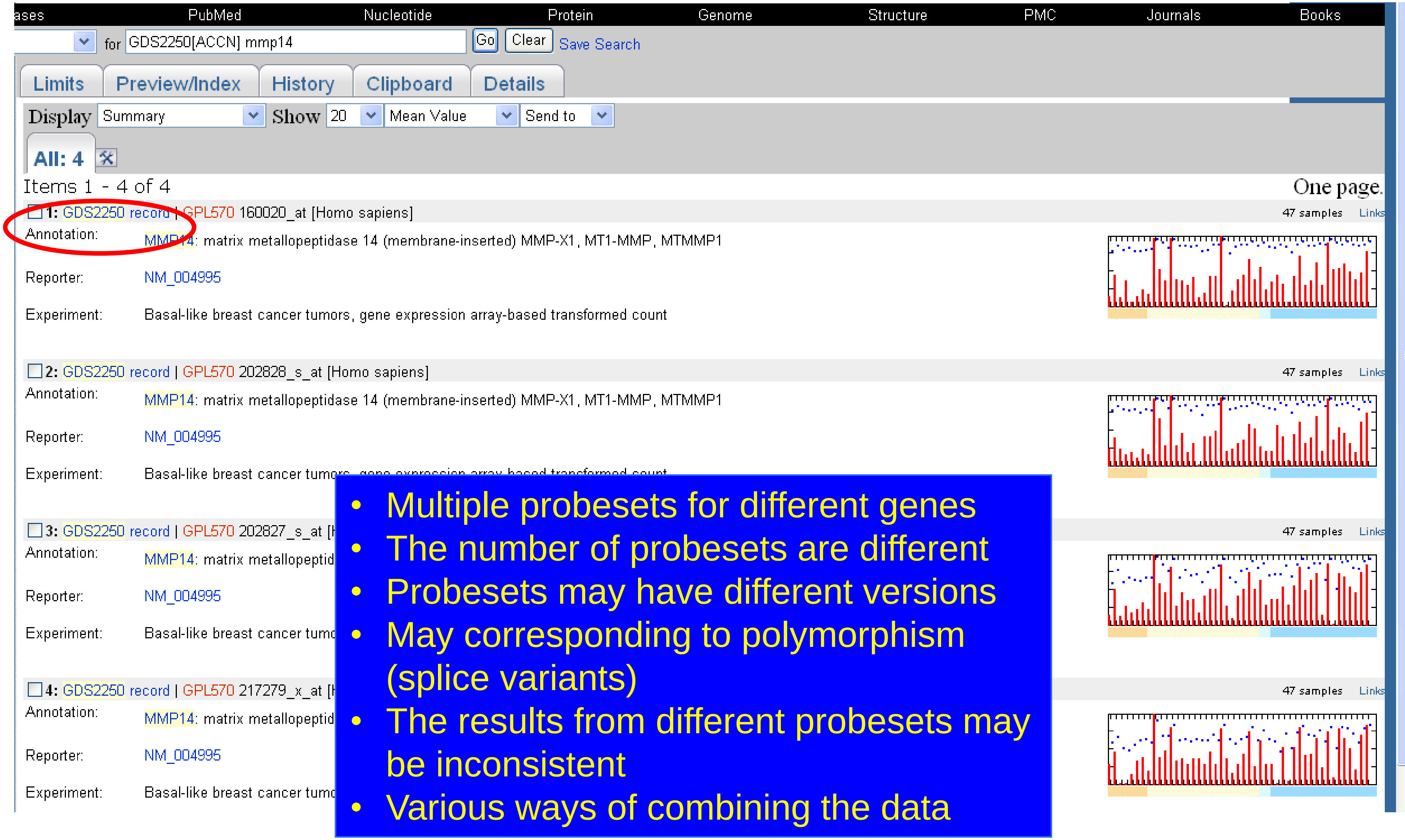
□ 4: GDS2250 record | GPL570 217279_x_at [Homo sapiens] 47 samples Links

Annotation: MMP14: matrix metallopeptidase 14 (membrane-inserted) MMP-X1, MT1-MMP, MTMMP1

Reporter: NM_004995

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

GEO Profiles



for GDS2250[ACCN] mmp14

Display Summary Show 20 Mean Value

Items 1 - 4 of 4

One page.

□ 1: GDS2250 record | GPL570 160020_at [Homo sapiens] 47 samples Links

Annotation: MMP14: matrix metallopeptidase 14 (membrane-inserted) MMP-X1, MT1-MMP, MTMMP1

Reporter: NM_004995

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

□ 2: GDS2250 record | GPL570 202828_s_at [Homo sapiens] 47 samples Links

Annotation: MMP14: matrix metallopeptidase 14 (membrane-inserted) MMP-X1, MT1-MMP, MTMMP1

Reporter: NM_004995

Experiment: Basal-like breast cancer tumors, gene expression array-based transformed count

□ 3: GDS2250 record | GPL570 202827_s_at [H 47 samples Links

Annotation: MMP14: matrix metallopeptid

Reporter: NM_004995

Experiment: Basal-like breast cancer tumo

□ 4: GDS2250 record | GPL570 217279_x_at [H 47 samples Links

Annotation: MMP14: matrix metallopeptid

Reporter: NM_004995

Experiment: Basal-like breast cancer tumo

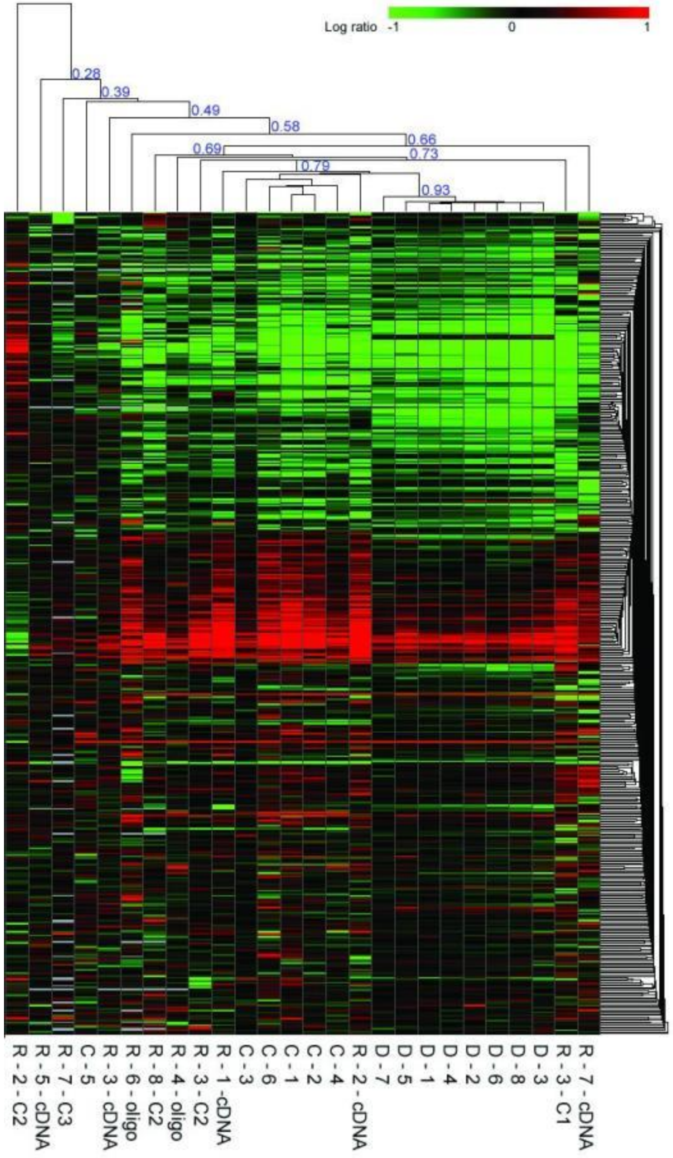
- Multiple probesets for different genes
- The number of probesets are different
- Probesets may have different versions
- May corresponding to polymorphism (splice variants)
- The results from different probesets may be inconsistent
- Various ways of combining the data

GEO Profiles

- Most new datasets are deposited as GSE series datasets instead of GDS datasets and cannot be visualized directly.
- Users need to download them for further processing.
- A simple way is to download the Data Matrix.

How do we use microarray?

- Profiling
- Comparative study
- Clustering
- Network inference



Supplementary Figure 1: Clustering of laboratory/platform combinations using log ratio values of commone genes

Moving beyond microarray?

- Cut the middleman
  - Next generation sequencing
  - Single-molecule sequencing
- Where will microarray go?
  - Diagnosis
  - Specialized quick testing kit