BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu

http://www.cs.fiu.edu/~giri/teach/BSC4934_Su11.html

July 2011

Gene Expression

- Process of transcription and/or translation of a gene is called gene expression.
- Every cell of an organism has the same genetic material, but different genes are expressed at different times.
- Patterns of gene expression in a cell is indicative of its state.

Hybridization

- ❑ If two complementary strands of DNA or mRNA are brought together under the right experimental conditions they will hybridize.
- ❑ A hybridizes to B ⇒
  - ● A is reverse complementary to B, or
  - ● A is reverse complementary to a subsequence of B.
- ❑ It is possible to experimentally verify whether A hybridizes to B, by labeling A or B with a radioactive or fluorescent tag, followed by excitation by laser.

Measuring gene expression

- Gene expression for a single gene can be measured by extracting mRNA from the cell and doing a simple hybridization experiment.
- Given a sample of cells, gene expression for every gene can be measured using a single microarray experiment.

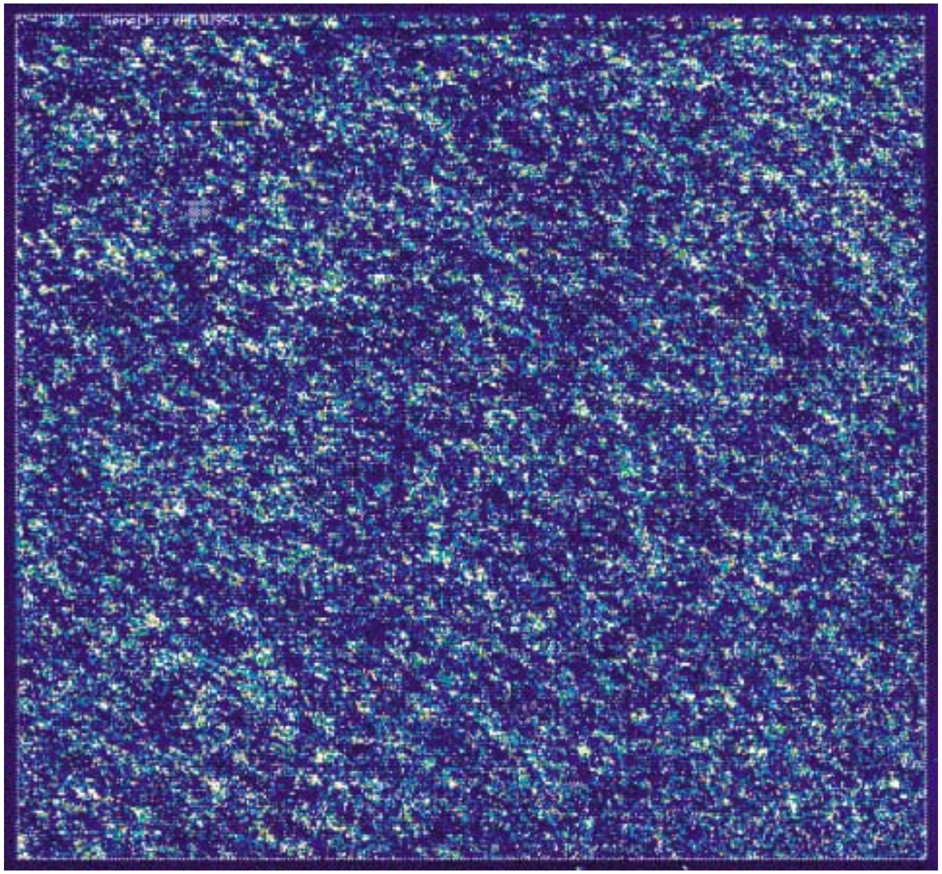
Microarray/DNA chip technology

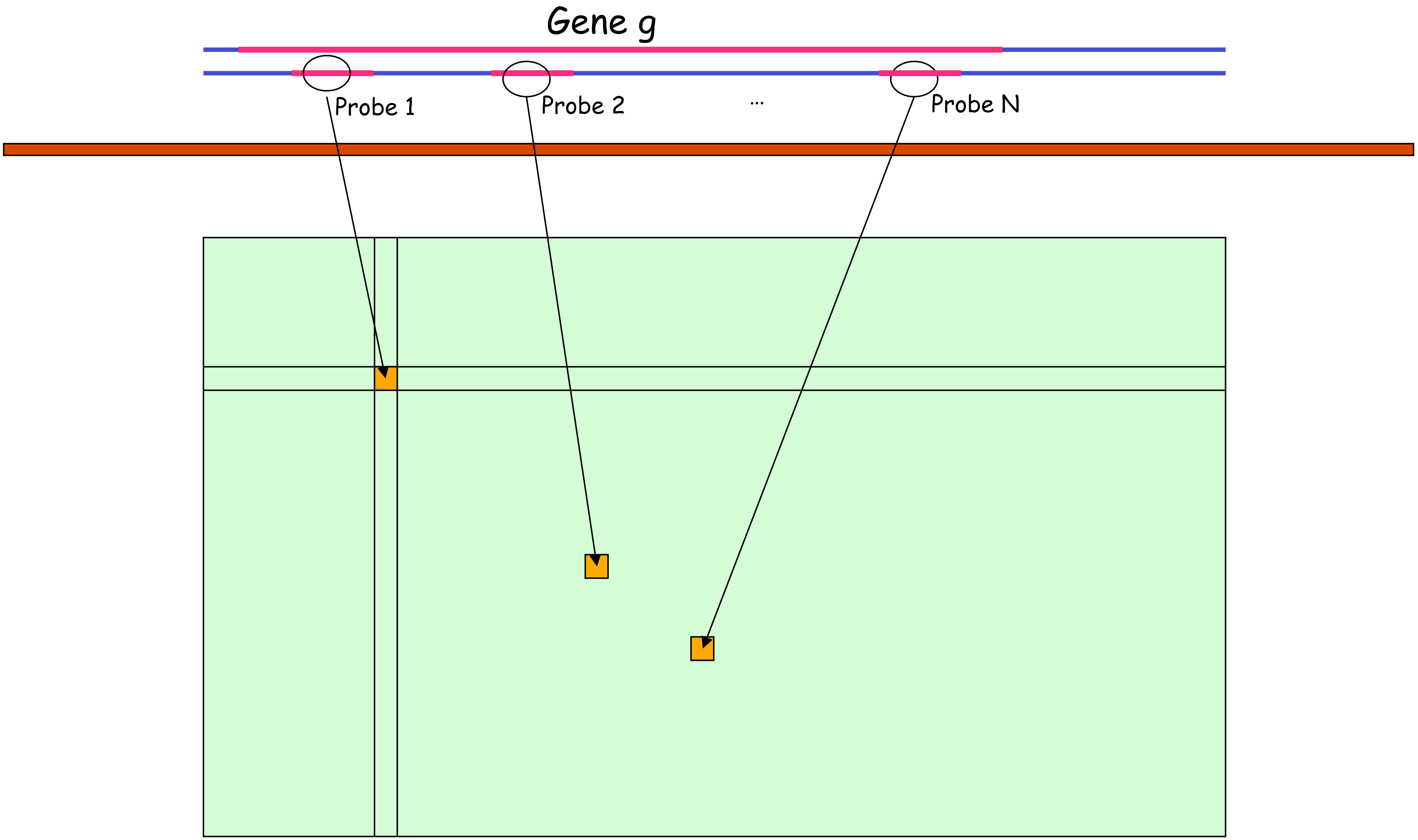
- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states

Microarray Data

Gene	Expression Level
Gene1	
Gene2	
Gene3	
…	

Gene Chips

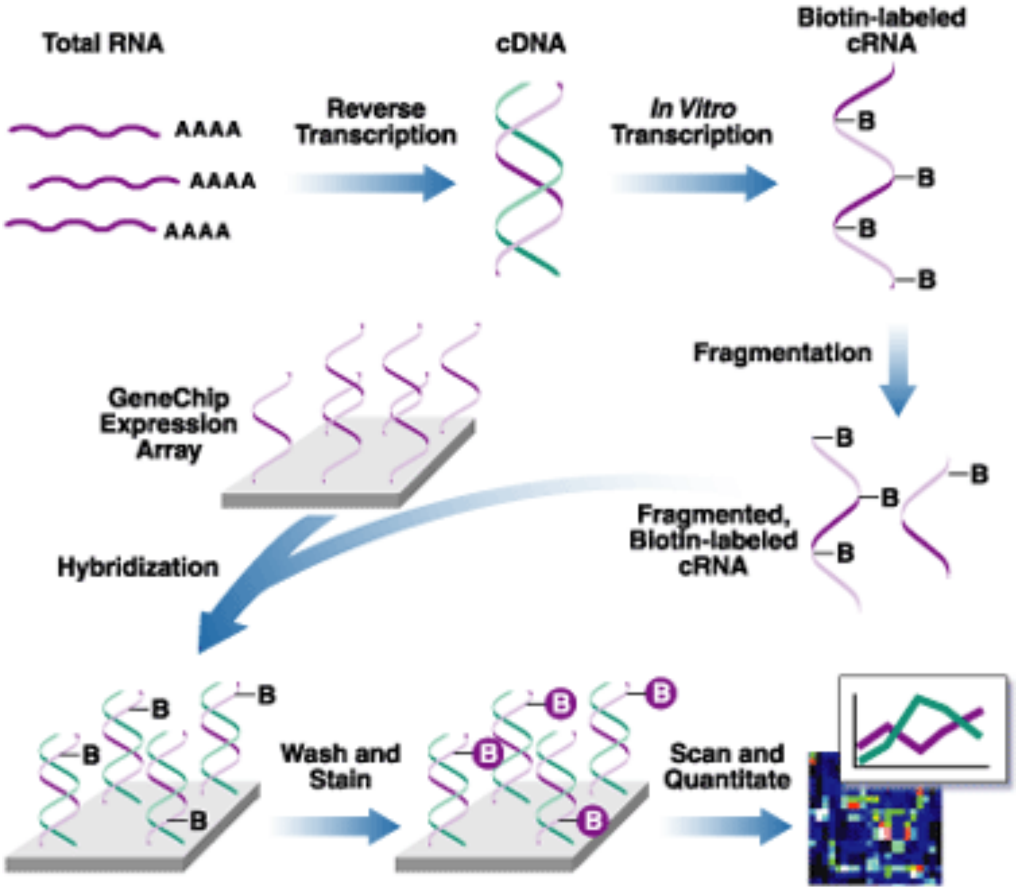



Gene g
Probe 1
Probe 2
…
Probe N

Microarray/DNA chips (Simplified)

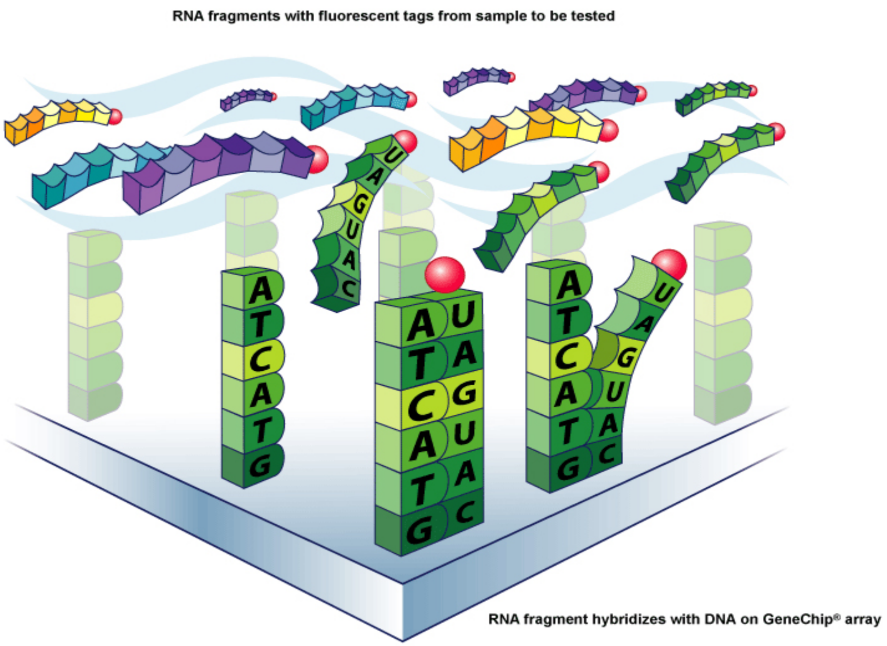
- ❑ Construct probes corresponding to reverse complements of genes of interest.
- ❑ Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- ❑ Extract mRNA from sample cells and label them.
- ❑ Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.
- ❑ Wash off unhybridized material.
- ❑ Use optical detector to measure amount of fluorescence from each spot.

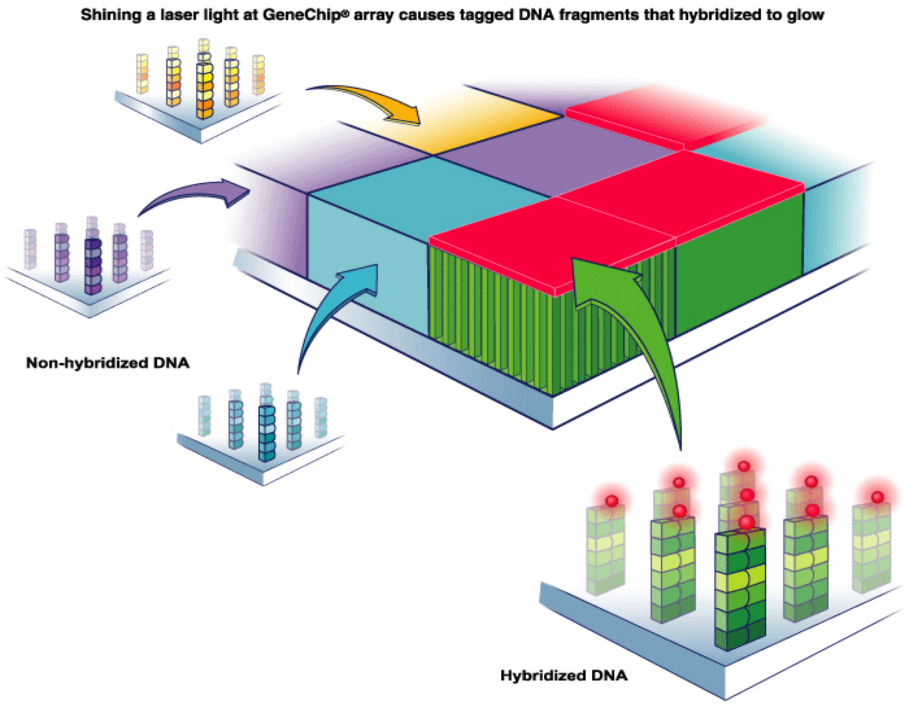
Affymetrix DNA chip schematic



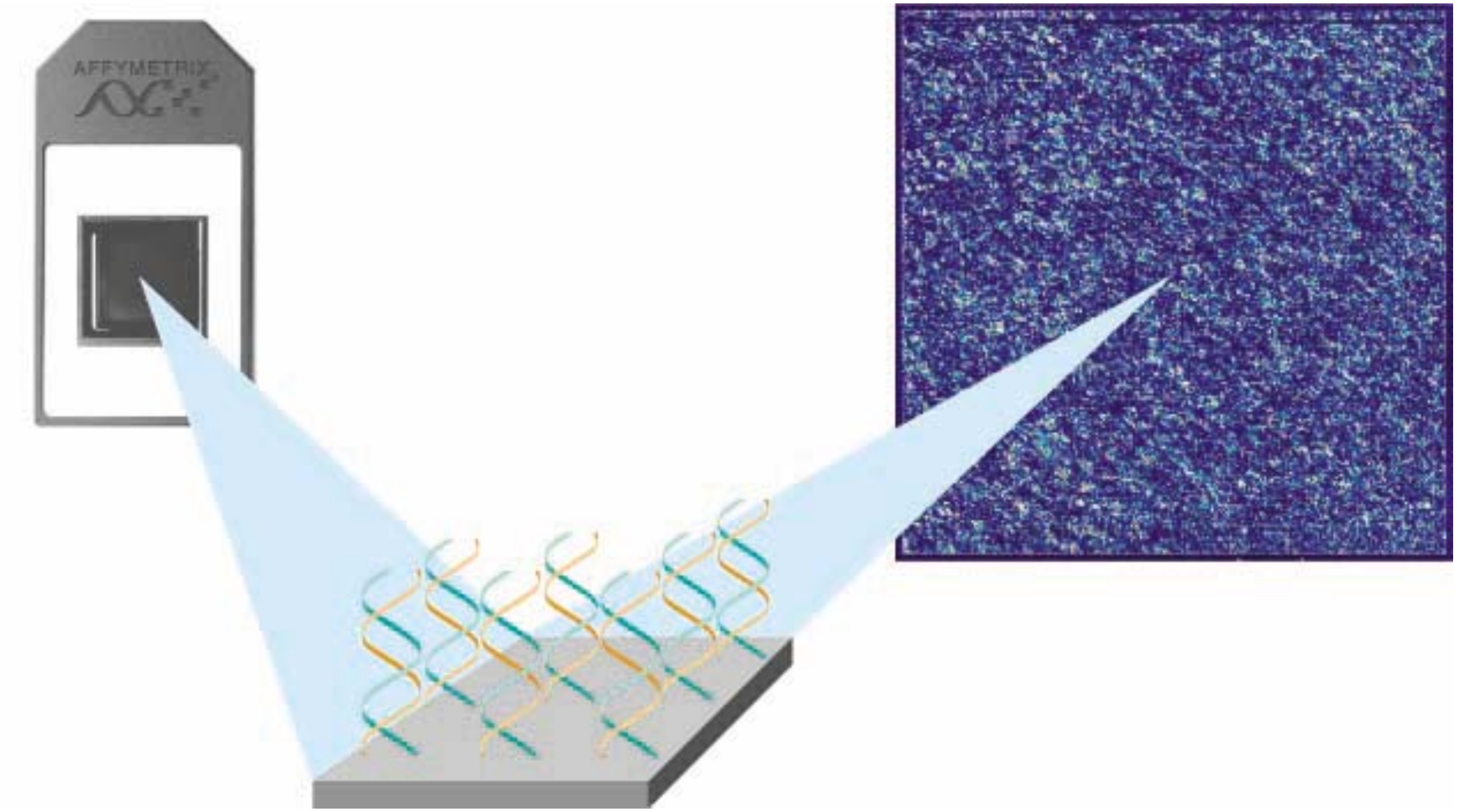
www.affymetrix.com

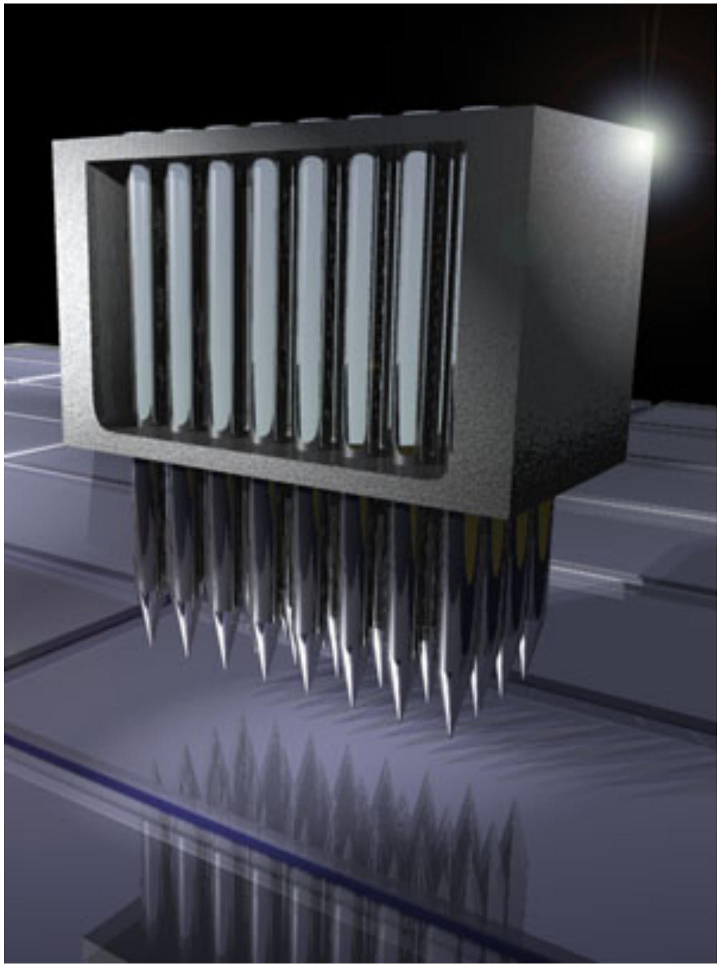
What’s on the slide?


RNA fragments with fluorescent tags from sample to be tested
A T C A T G
A T C A T G
U A G U A C
A T C A T G
U A G U A C
U A G U A C
RNA fragment hybridizes with DNA on GeneChip® array


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

DNA Chips & Images

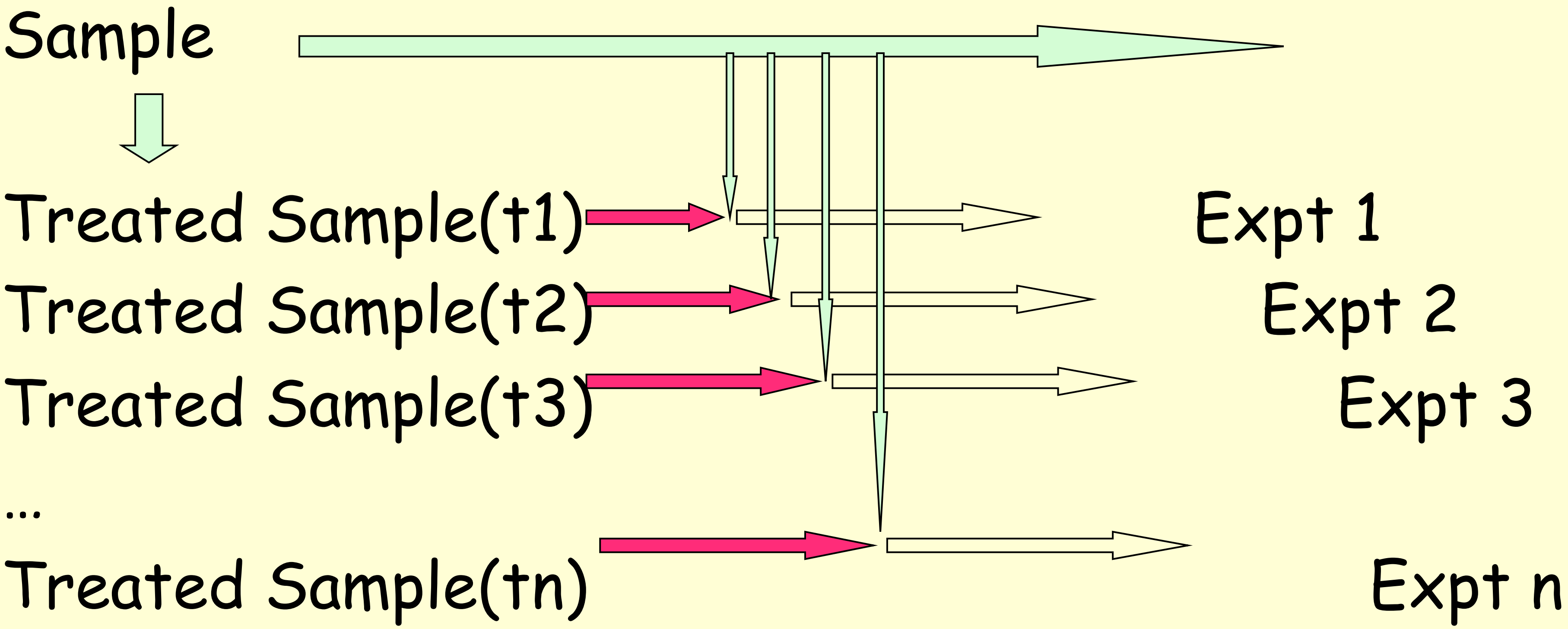




Microarrays: competing technologies

- Affymetrix & Agilent
- Differ in:
  - method to place DNA: Spotting vs. photolithography
  - Length of probe
  - Complete sequence vs. series of fragments

Study effect of treatment over time



Sample

Treated Sample(t1)

Treated Sample(t2)

Treated Sample(t3)

…

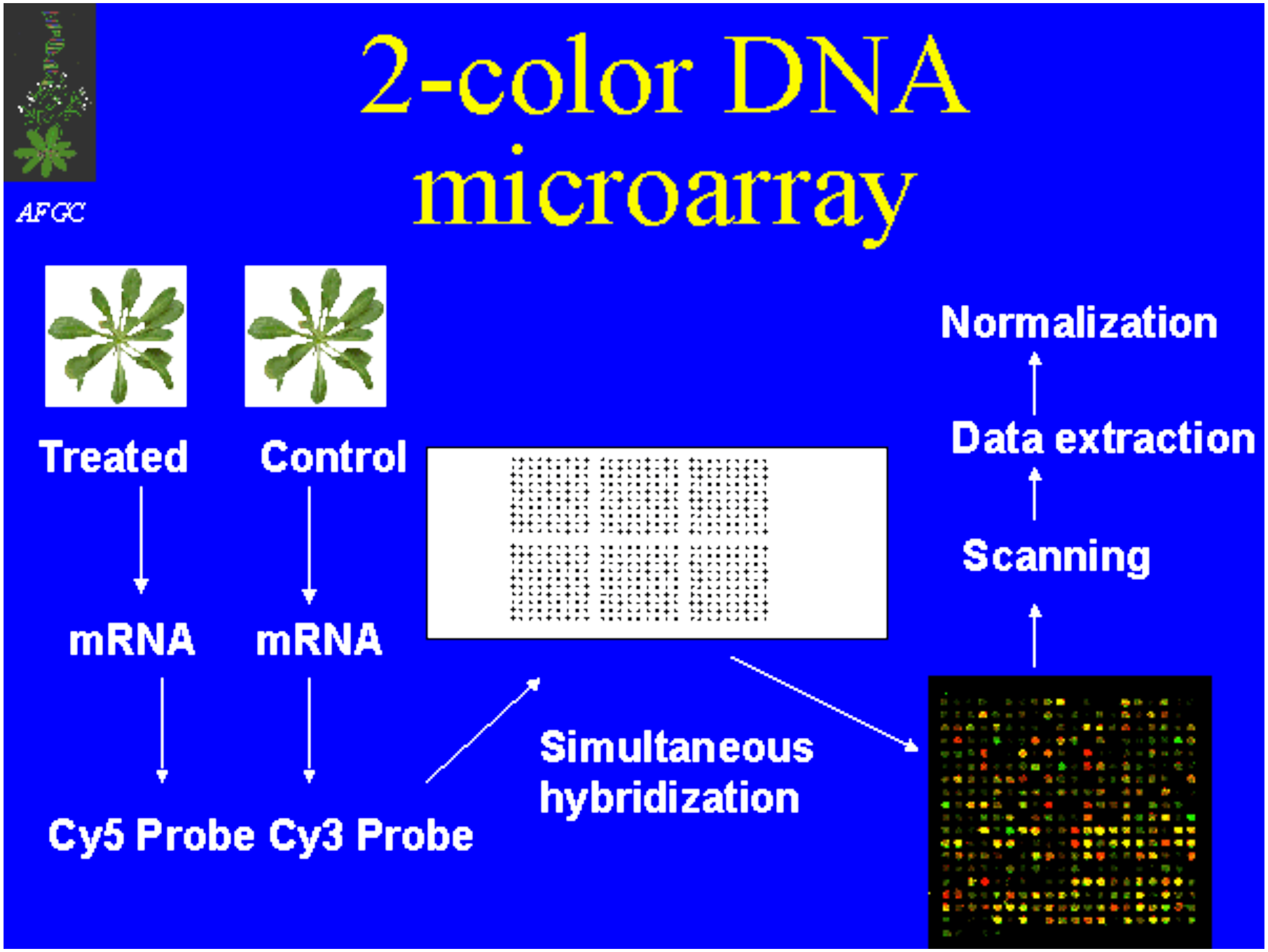
Treated Sample(tn)

Expt 1

Expt 2

Expt 3

Expt n



AFGC

2-color DNA microarray

Treated

mRNA

Cy5 Probe

Control

mRNA

Cy3 Probe

Simultaneous hybridization

Scanning

Data extraction

Normalization

http://www.arabidopsis.org/info/2010_projects/comp_proj/AFGC/RevisedAFGC/Friday/

How to compare 2 cell samples with Two-Color Microarrays?

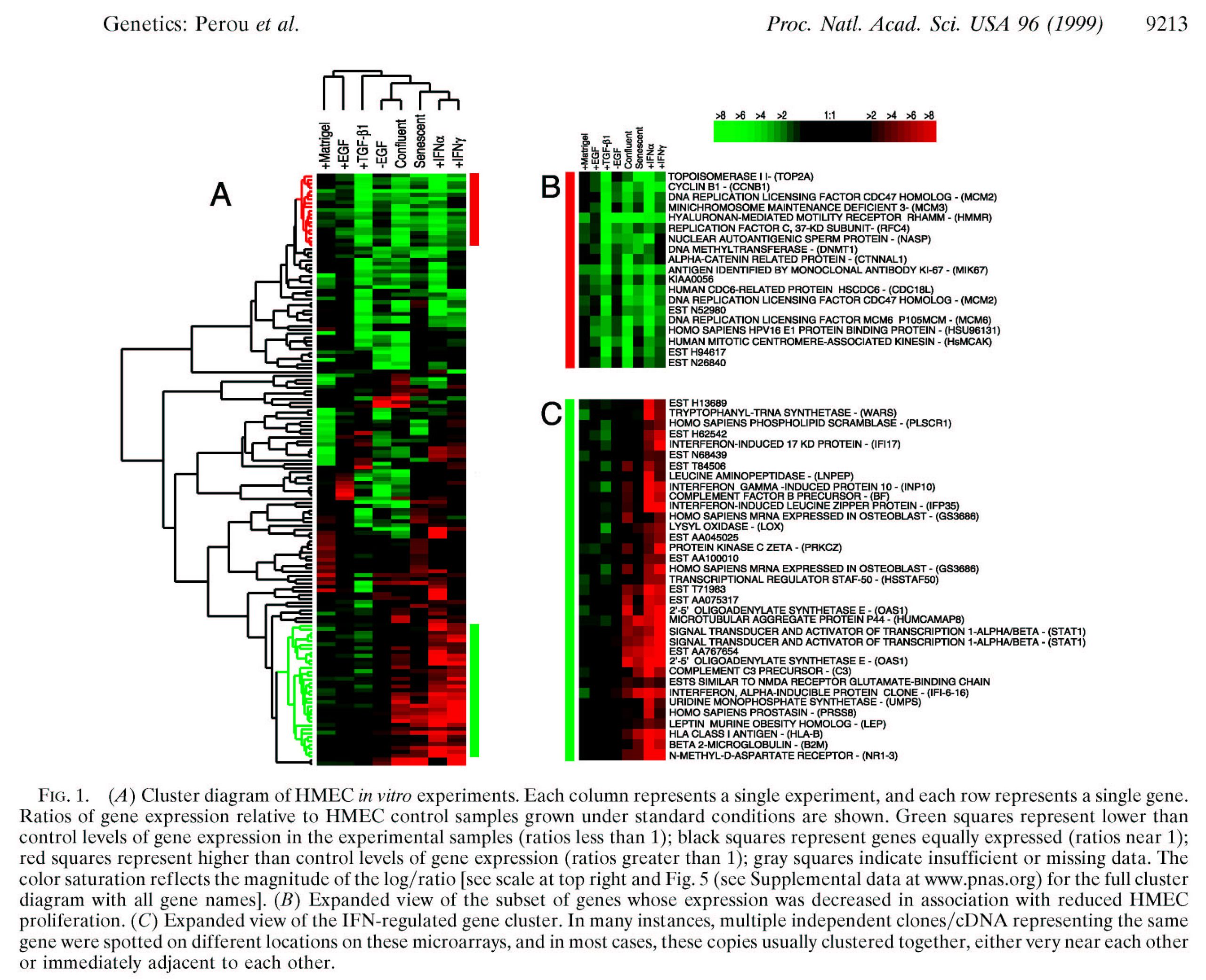
- ❑ mRNA from sample 1 is extracted and labeled with a red fluorescent dye.
- ❑ mRNA from sample 2 is extracted and labeled with a green fluorescent dye.
- ❑ Mix the samples and apply it to every spot on the microarray. Hybridize sample mixture to probes.
- ❑ Use optical detector to measure the amount of green and red fluorescence at each spot.

Sources of Variations & Experimental Errors

- ❑ Variations in cells/individuals
- ❑ Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference
- ❑ Variations in probe concentration, probe amounts, substrate surface characteristics
- ❑ Variations in hybridization conditions and kinetics
- ❑ Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- ❑ Cross-hybridization of sequences with high sequence identity
- ❑ Limit of factor 2 in precision of results
- ❑ Variation changes with intensity: larger variation at low or high expression levels

Need to Normalize data

Analyzing Microarray Data



Genetics: Perou *et al.* *Proc. Natl. Acad. Sci. USA 96 (1999)* 9213

FIG. 1. (*A*) Cluster diagram of HMEC *in vitro* experiments. Each column represents a single experiment, and each row represents a single gene. Ratios of gene expression relative to HMEC control samples grown under standard conditions are shown. Green squares represent lower than control levels of gene expression in the experimental samples (ratios less than 1); black squares represent genes equally expressed (ratios near 1); red squares represent higher than control levels of gene expression (ratios greater than 1); gray squares indicate insufficient or missing data. The color saturation reflects the magnitude of the log/ratio [see scale at top right and Fig. 5 (see Supplemental data at www.pnas.org) for the full cluster diagram with all gene names]. (*B*) Expanded view of the subset of genes whose expression was decreased in association with reduced HMEC proliferation. (*C*) Expanded view of the IFN-regulated gene cluster. In many instances, multiple independent clones/cDNA representing the same gene were spotted on different locations on these microarrays, and in most cases, these copies usually clustered together, either very near each other or immediately adjacent to each other.

Microarray Data Analysis: Subtyping

MOLECULAR CLASSIFICATION OF HUMAN CARCINOMAS

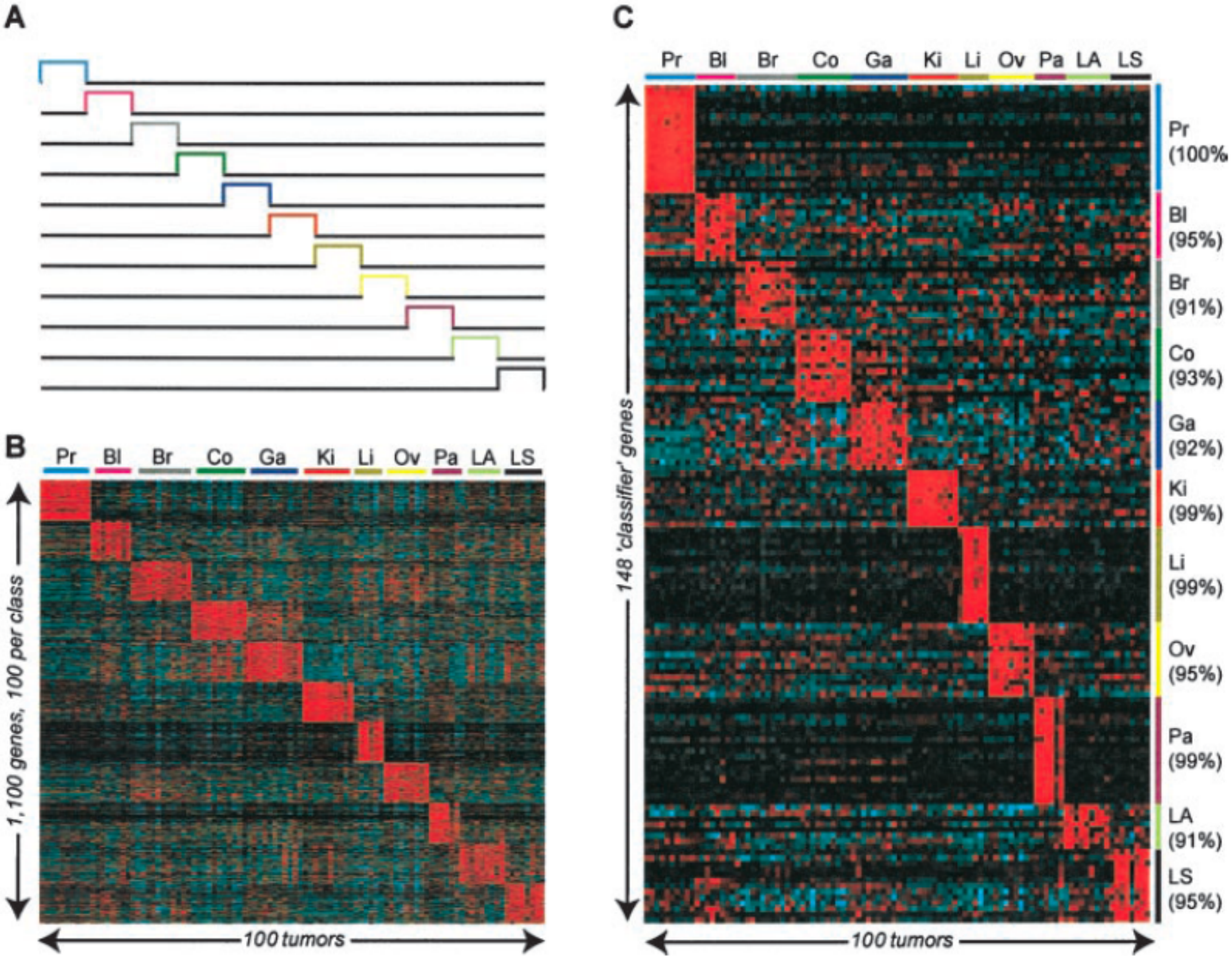


Fig. 1. Selection of tumor-specific genes for cancer class prediction. *A*, schematic diagram depicting the idealized expression profile of tumor-specific genes that the method selects as classifiers. The shape of each profile represents genes that are highly expressed in each cancer type relative to all other tumors in the training set. *B*, 100 genes per tumor class (total, 1100) with the most significant scores in a Wilcoxon rank-sum test for equality were selected as likely candidates for tumor classifiers. *Pr*, prostate; *Bl*, bladder/ureter; *Br*, breast; *Co*, colorectal; *Ga*, gastroesophagus; *Ki*, kidney; *Li*, liver; *Ov*, ovary; *Pa*, pancreas; *LA*, lung adenocarcinomas; *LS*, lung squamous cell carcinoma. *C*, the final refined set of gene classifiers was generated after the genes in *B* were ranked by SVM/LOOCV accuracy. Annotations of the genes from which 110 "predictor" genes were bootstrapped are provided on our website.[4] For clarity, only 8 of 76 predictor genes for lung adenocarcinomas are depicted here. Levels of gene expression (depicted in each *row*) across all samples (*columns*) were median-centered and normalized by "Cluster" and output in "Treeview" (12). *Red*, increased gene expression; *blue*, decreased expression; *black*, median level of gene expression. The color intensity is proportional to the hybridization intensity of a gene from its median level across all samples.

Differential Analysis

- Determine differentially expressed genes
  - Need for Replication and Normalization
  - Differential Analysis: test statistics
    - Fold-change (Sample vs Control)
    - t-test
    - F-statistic
    - Other Non-parametric rank-based statistics
  - Significance of observed statistic (Permutation test)
  - False Discovery Rate
    - Multiple test corrections
  - Pattern Discovery

Pattern Discovery

- Dimensionality reduction
  - Principal Component Analysis
  - Multidimensional scaling
  - Singular-value decomposition
- Visualization methods

Pattern Discovery

Principal Component Analysis

Clustering

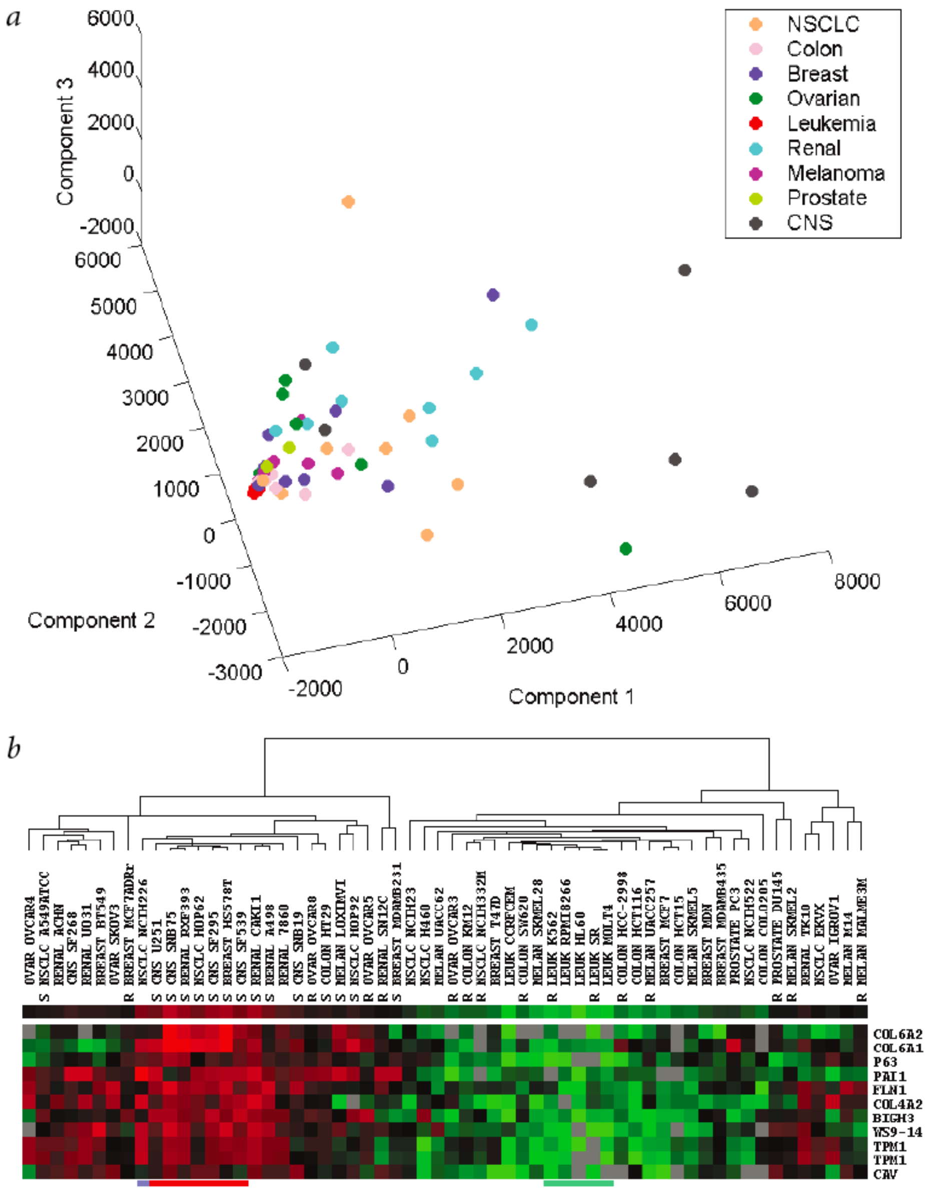


Fig. 2 Two pattern-discovery techniques. Data for both figures measure expression for 11 genes characterizing sensitivity to compound cytochalasin D in 60 cancer cell lines[97]. ***a***, The first three principal components, plotted using Matlab software (Mathworks). Apparent features include a tight cluster of leukemia samples (red dots, nearly superimposed) and the more scattered outlying cluster of CNS tumors (black dots). A single lung cancer sample (NSCLC-NCIH226) also appears as an outlier — the solitary orange dot at the top. ***b***, Hierarchical clustering of the same data, using Cluster/TreeView (http://rana.lbl.gov/EisenSoftware.htm). Names of samples extremely sensitive or resistant to cytochalasin D (see Supplementary information) are prefixed 'S' and 'R' respectively. The samples fall into two main clusters, roughly, but not perfectly, separating the sensitive and resistant samples. As in ***a***, fine structure shows a tight leukemia cluster (underlined in green) and a tight CNS cluster (underlined in red), but does not suggest that the CNS cluster or NSCLC-NCIH226 (underlined in blue) are outliers. Apparent in both ***a*** and ***b*** is the relative heterogeneity of the breast cancer cell lines.

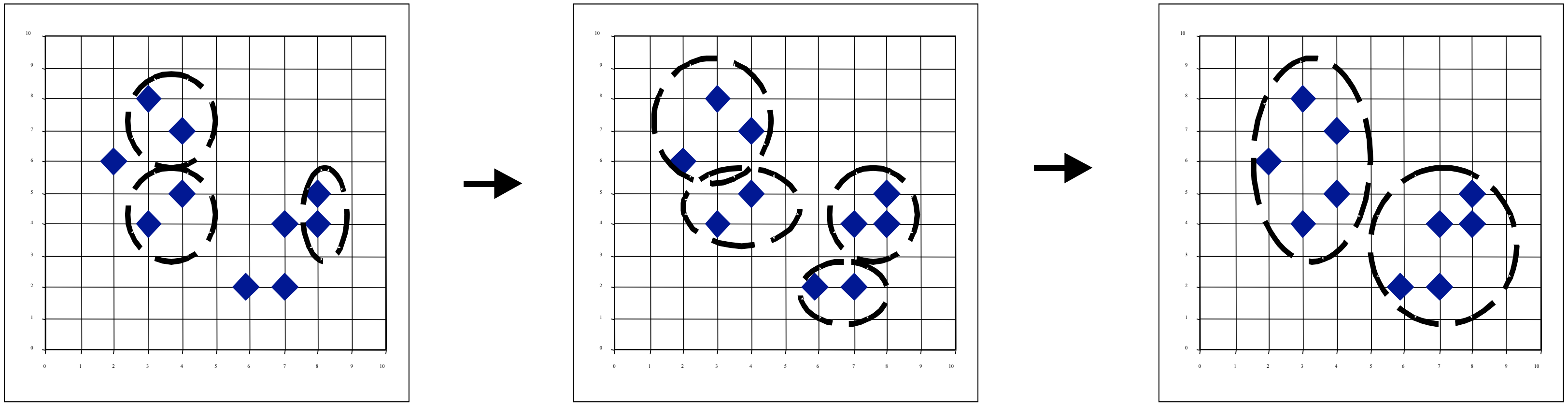
merging the two closest clusters is repeated until a single cluster remains. This arranges the data into a tree structure that can be broken into the desired number of clusters by cutting across the tree at a particular height. Tree structures are easily viewed and understood (Fig. 2*b*), and the hierarchical structure provides potentially useful information about the relationships between clusters. Trees are known to reveal close relationships very well. However, as

07/18/11

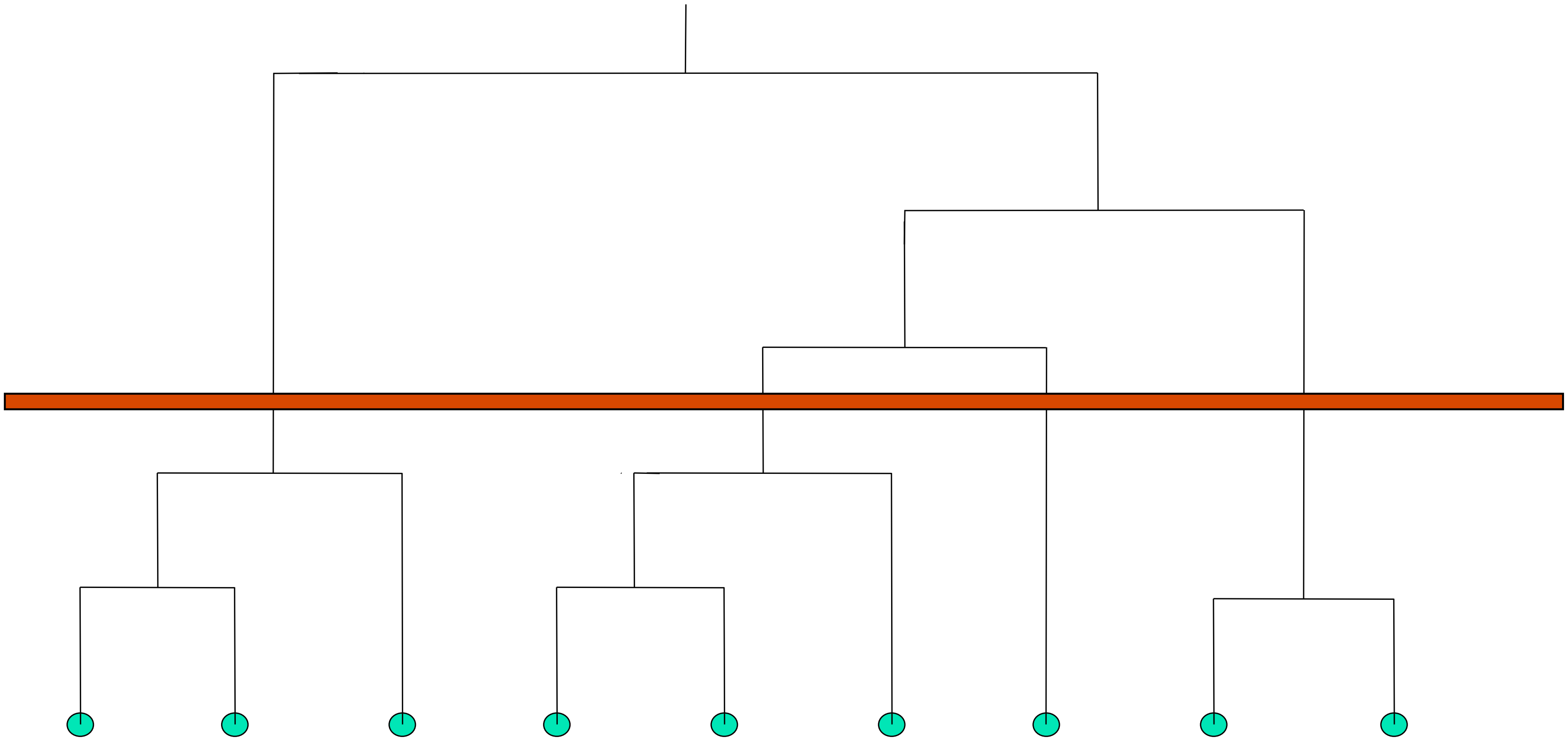
Clustering

- Clustering is a general method to study patterns in gene expressions.
- Several known methods:
  - Hierarchical Clustering (Bottom-Up Approach)
  - K-means Clustering (Top-Down Approach)
  - Self-Organizing Maps (SOM)

Hierarchical Clustering: Example



A Dendrogram



Hierarchical Clustering [Johnson, SC, 1967]

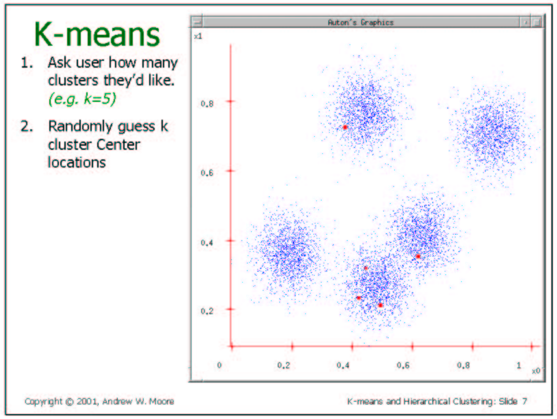
- Given n points in R^d, compute the distance between every pair of points
- While (not done)
  - Pick closest pair of points s_i and s_j and make them part of the same cluster.
  - Replace the pair by an average of the two s_{ij}

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

K-Means Clustering: Example

Example from Andrew Moore’s tutorial on Clustering.

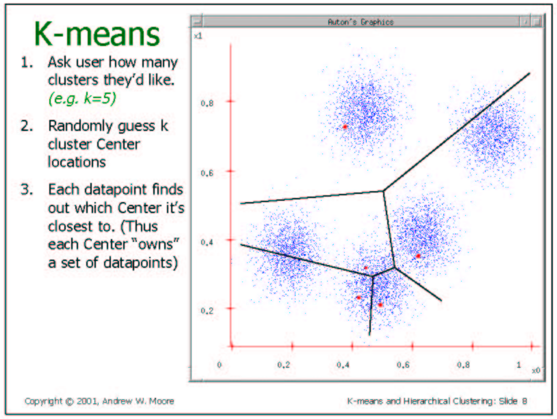
Start



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations

K-means and Hierarchical Clustering: Slide 7

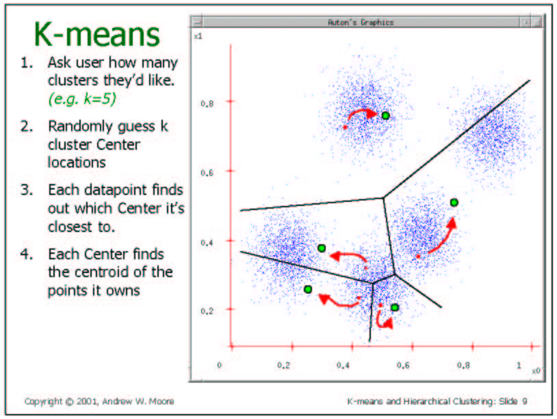


K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)

K-means and Hierarchical Clustering: Slide 8

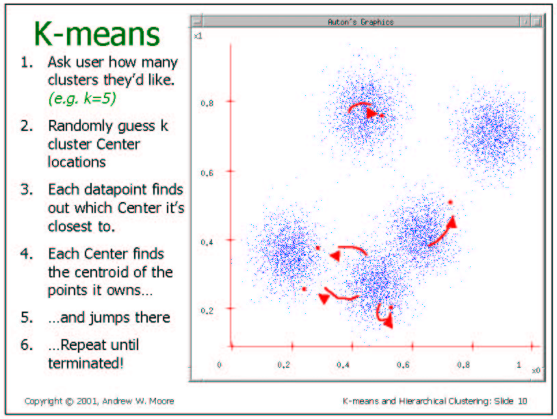
4



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns

K-means and Hierarchical Clustering: Slide 9

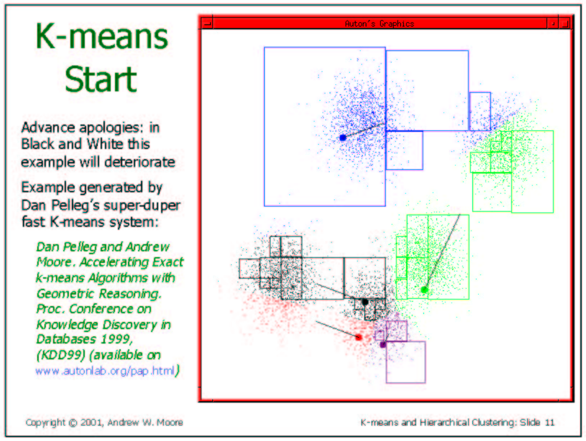


K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns…
5. …and jumps there
6. …Repeat until terminated!

K-means and Hierarchical Clustering: Slide 10

5



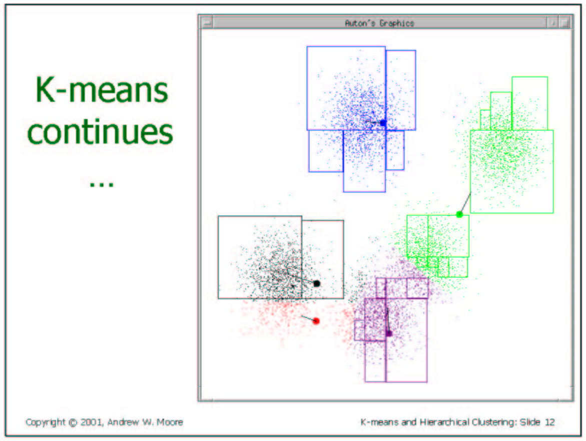
K-means Start

Advance apologies: in Black and White this example will deteriorate

Example generated by Dan Pelleg's super-duper fast K-means system:

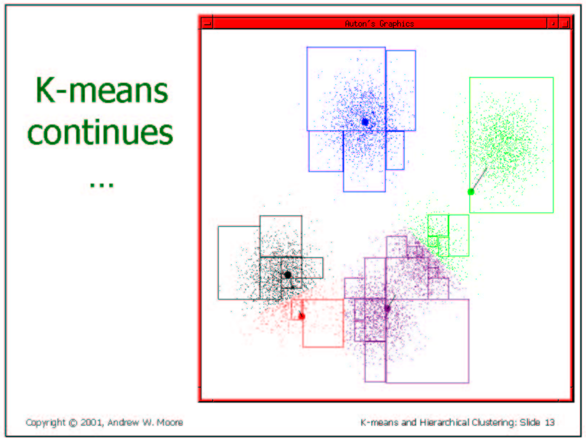
Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html*)*

K-means and Hierarchical Clustering: Slide 11



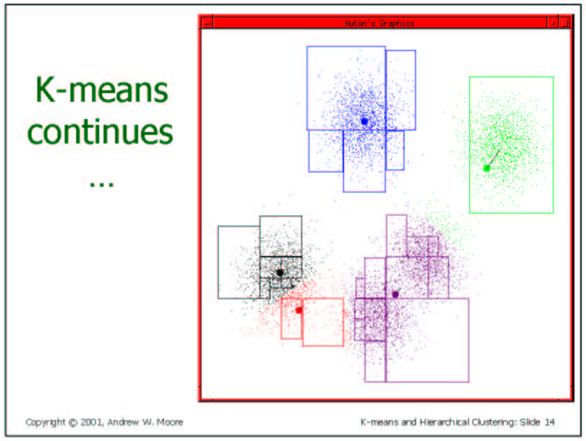
K-means continues …

K-means and Hierarchical Clustering: Slide 12



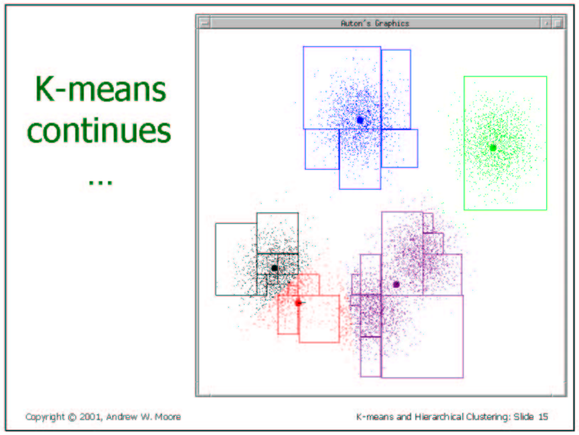
K-means continues …

K-means and Hierarchical Clustering: Slide 13



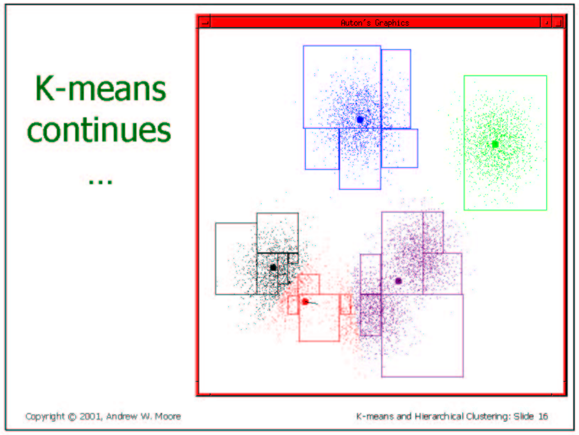
K-means continues …

K-means and Hierarchical Clustering: Slide 14



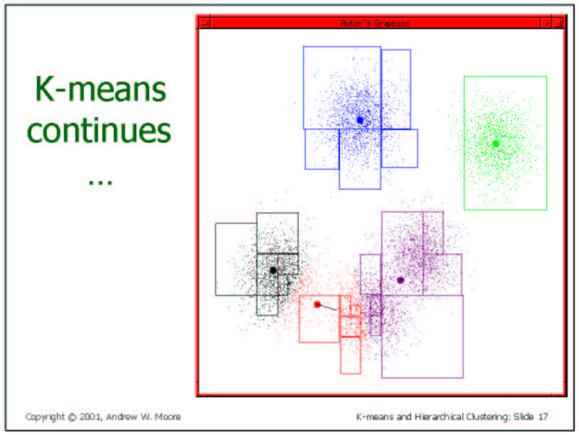
K-means continues …

K-means and Hierarchical Clustering: Slide 15



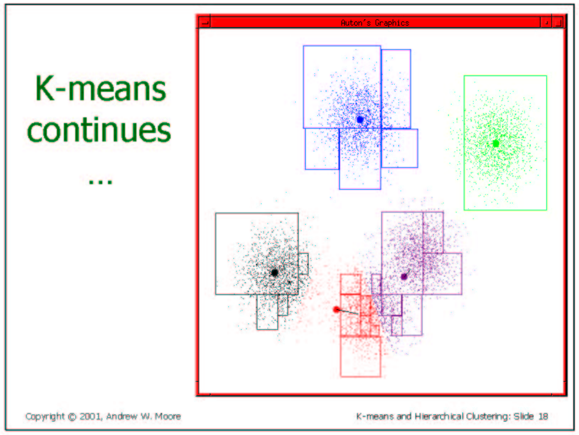
K-means continues …

K-means and Hierarchical Clustering: Slide 16



K-means continues …

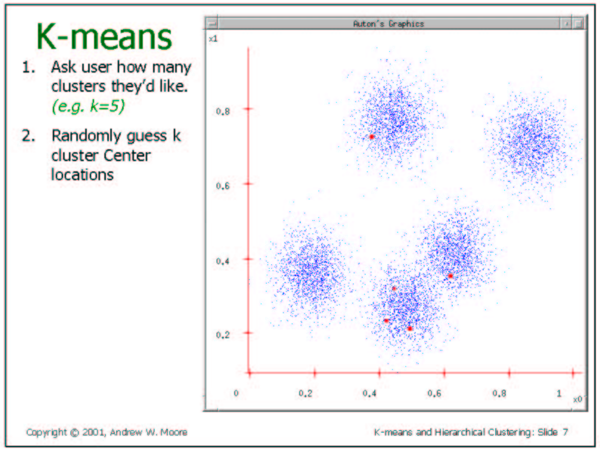
K-means and Hierarchical Clustering: Slide 17



K-means continues …

K-means and Hierarchical Clustering: Slide 18

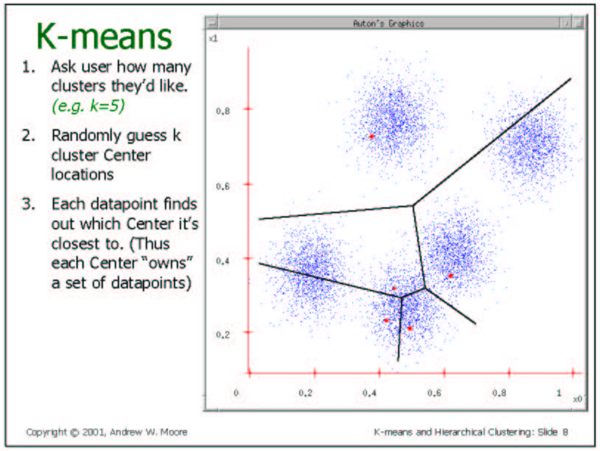
Start



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations

K-means and Hierarchical Clustering: Slide 7

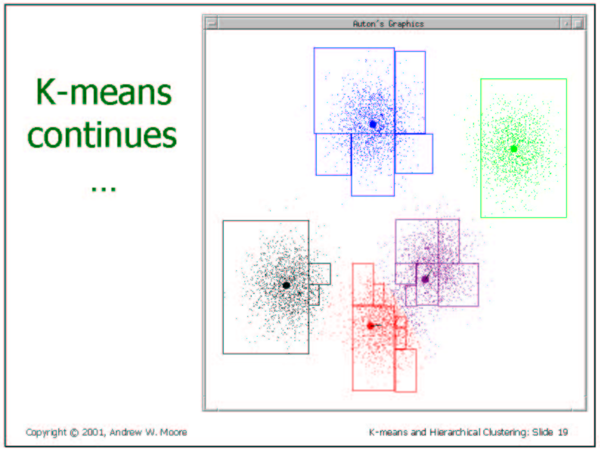


K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)

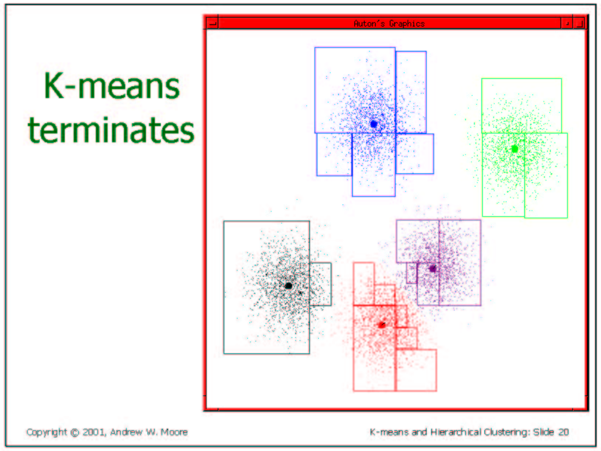
K-means and Hierarchical Clustering: Slide 8

4



K-means continues …

K-means and Hierarchical Clustering: Slide 19



K-means terminates

K-means and Hierarchical Clustering: Slide 20

10

End

K-Means Clustering [McQueen '67]

Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Comparisons

- Hierarchical clustering
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.
- K-Means
  - Need definition of a mean. Categorical data?
  - More efficient and often finds optimum clustering.

Class Prediction

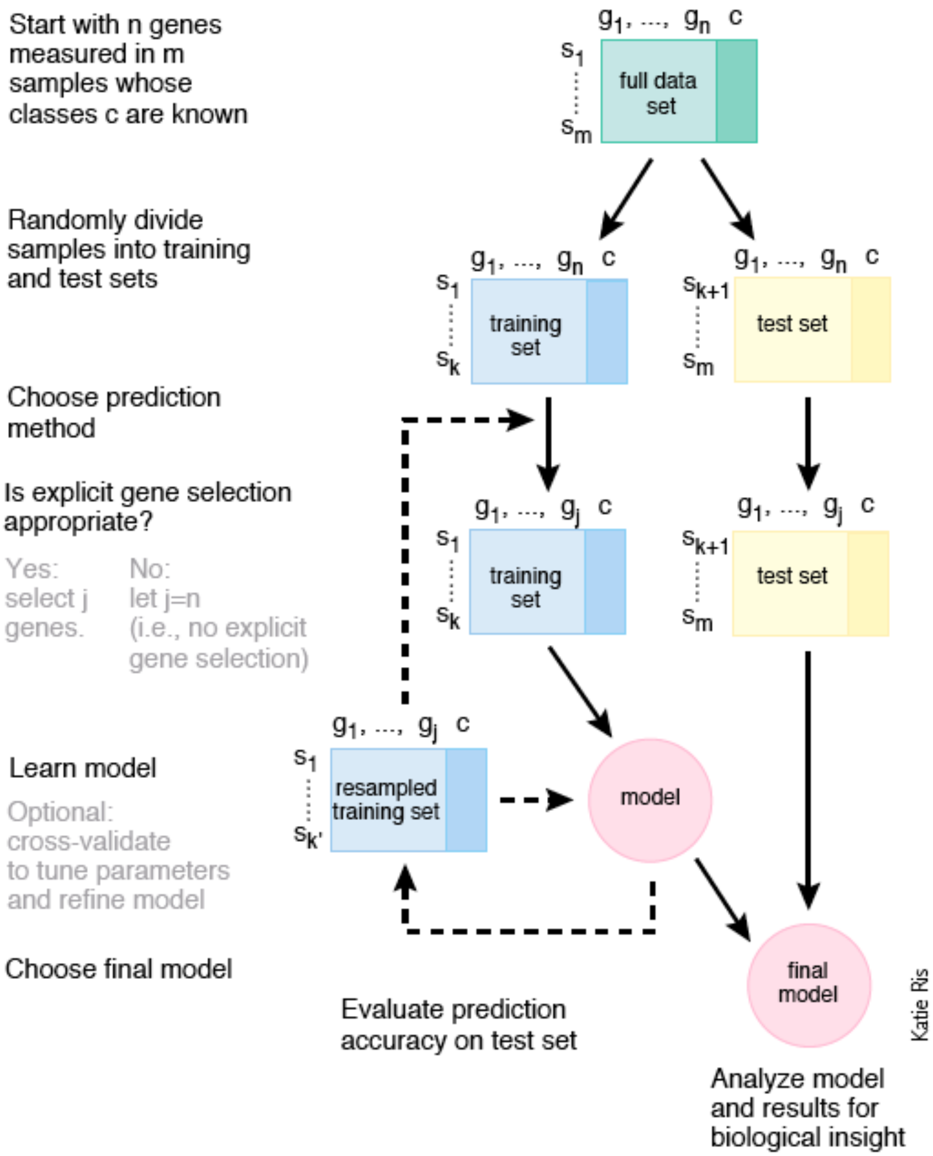

Start with n genes measured in m samples whose classes c are known
Randomly divide samples into training and test sets
Choose prediction method
Is explicit gene selection appropriate?
Yes: select j genes.
No: let j=n (i.e., no explicit gene selection)
Learn model
Optional: cross-validate to tune parameters and refine model
Choose final model
Evaluate prediction accuracy on test set
Analyze model and results for biological insight
g_1, ..., g_n c
s_1
s_m
full data set
training set
s_k
s_{k+1}
test set
g_1, ..., g_j c
resampled training set
$s_{k'}$
model
final model
Katie Ris

Fig. 3 An overview of the process for building a prediction model to classify samples. The partition into training and test data is ideally chosen at random across the entire set of samples. Many prediction methods require tuning some parameter (such as the number of genes, the number of nearest-neighbors to consider, or the number of decision trees built). This choice is often evaluated by cross-validation — the process of repeatedly removing smaller test sets from the training set, building new models (starting with the gene selection process) with the remaining data, and evaluating performance across all the different models built. For example, "leave-one-out cross validation" (also called "*n*-way") builds *n* models, each using *n*–1 training examples and evaluated on the remaining one; the accuracy for predicting all *n* samples is reported. Observing that predictors may succeed by chance even in cross-validation, Radmacher *et al.* suggest using permutation testing to determine the significance of the observed results[98]. Ultimately the final model, perhaps chosen during the cross-validation process, is then tested on entirely new data not used in the model generation process. The model itself, as well as the prediction results and the influential genes, may yield new biological insights.

Class Prediction Methods

- Decision Trees
- Support Vector Machines (SVM)
- k-NN or k-nearest neighbor method
- Fisher's linear discriminant method
- Neural Networks
- Self-Organizing Maps
- Ensemble methods
  - Boosting
  - Bagging

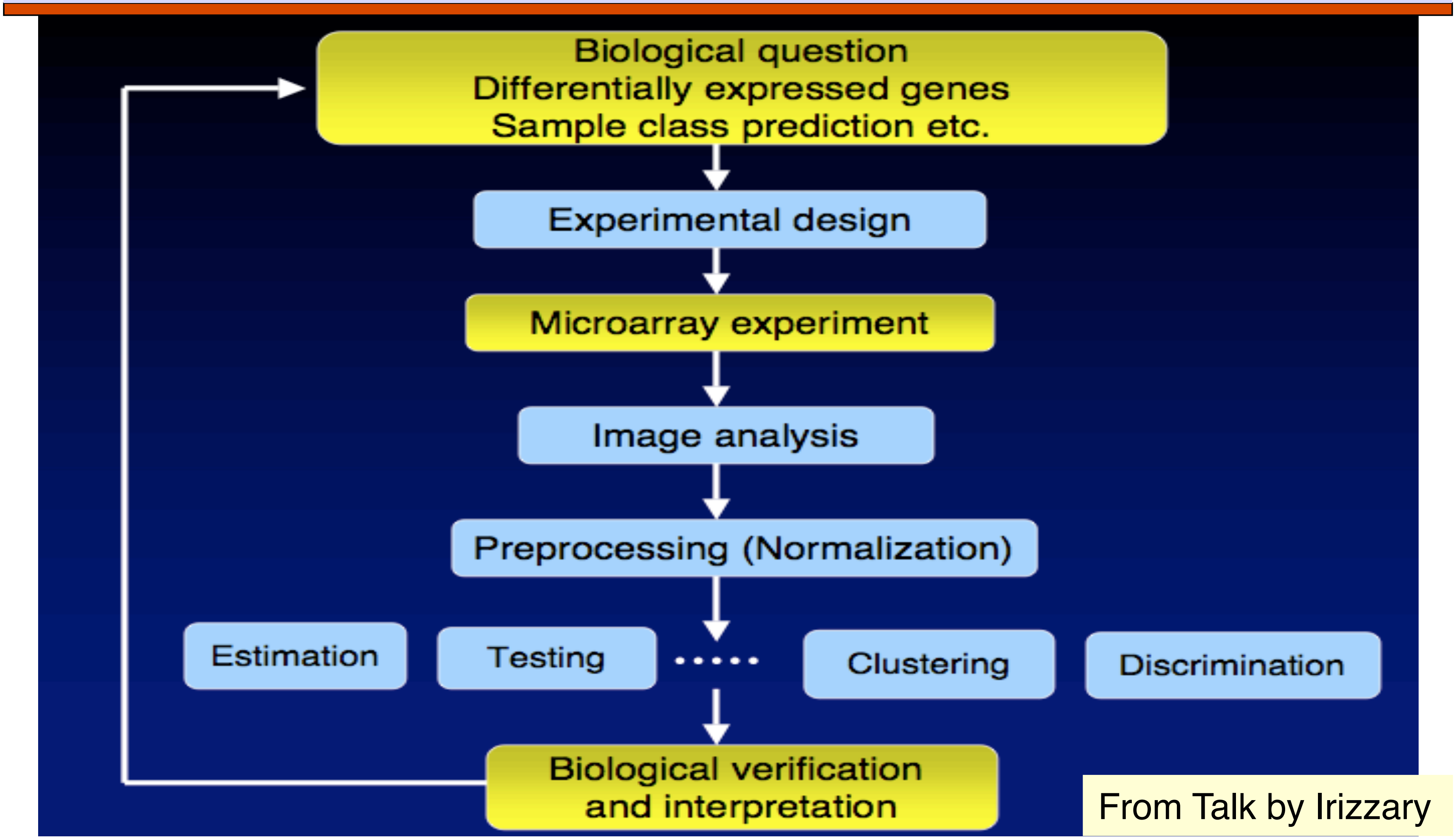
Functional Biases, Pathways & Networks

- ❑ Over/Under-representation of functional groups of genes
- ❑ Over/Under-representation of genes involved in functional pathways
- ❑ Inferring of regulatory relationships
- ❑ Inferring of protein-protein interactions

Reading

- The following slides come from a series of talks by Rafael Irizzary from Johns Hopkins
- Much of the material can be found in detail in the following papers from [http://www.biostat.jhsph.edu/~ririzarr/papers/]
  - Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics. Vol. 4, Number 2: 249-264.
  - Bolstad, B.M., Irizarry RA, Astrand, M, and Speed, TP (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics. 19(2):185-193.

Inference Process



Biological question
Differentially expressed genes
Sample class prediction etc.

Experimental design

Microarray experiment

Image analysis

Preprocessing (Normalization)

Estimation

Testing

.....

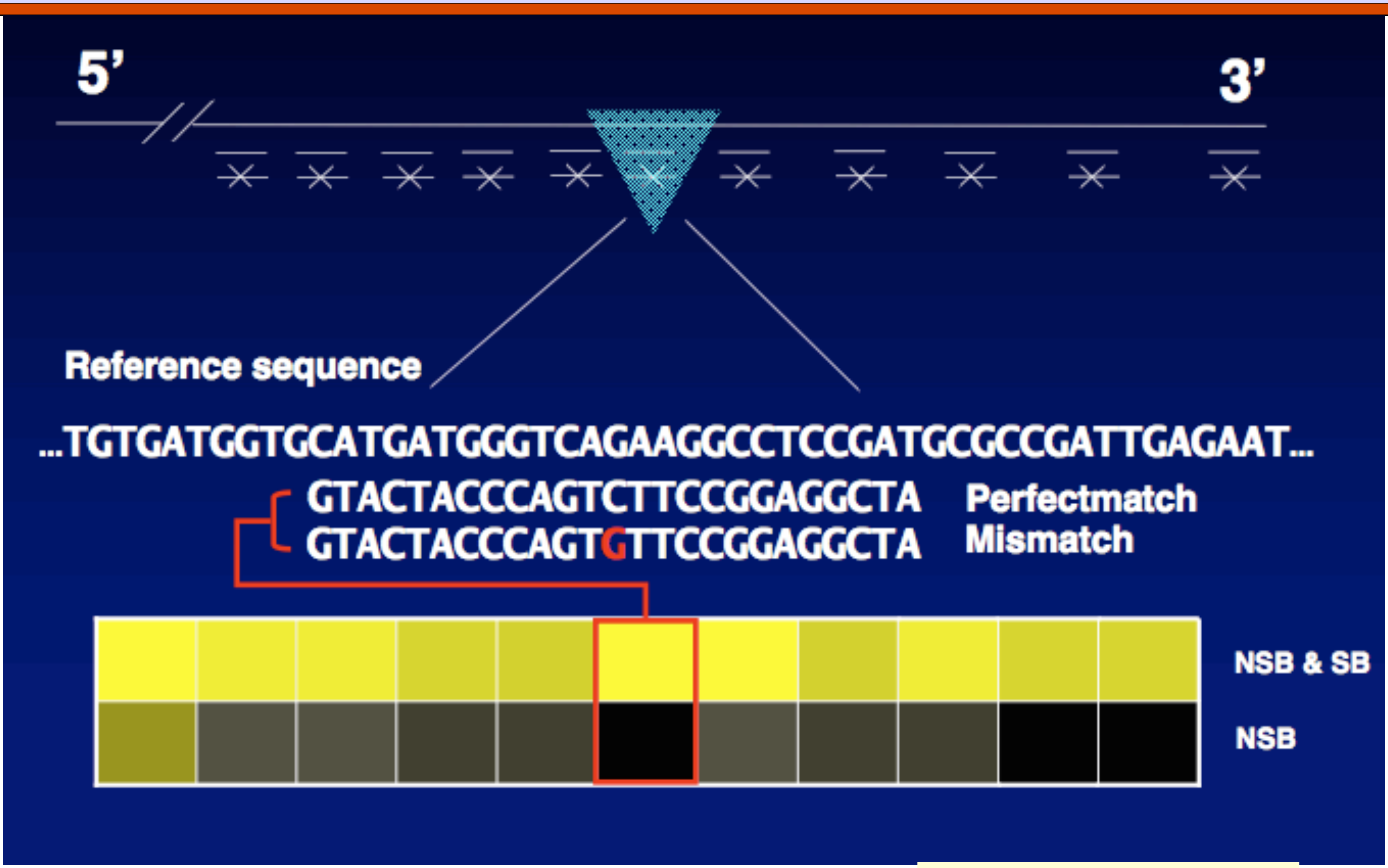
Clustering

Discrimination

Biological verification
and interpretation

From Talk by Irizzary

Affymetrix Genechip Design



5’

3’

Reference sequence

…TGTGATGGTGCATGATGGGTCAGAAGGCCTCCGATGCGCCGATTGAGAAT…

GTACTACCCAGTCTTCCGGAGGCTA Perfectmatch

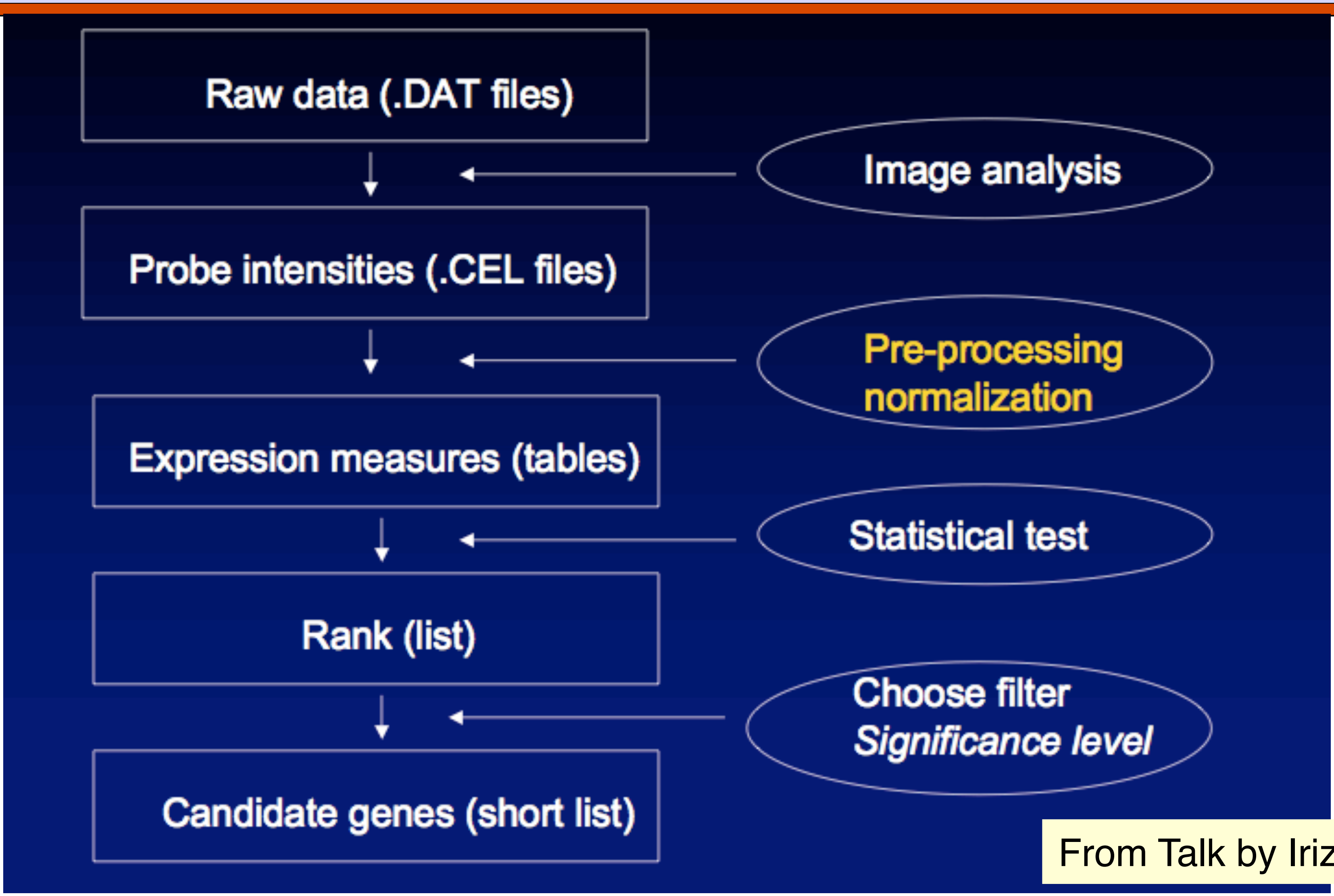
GTACTACCCAGTGTTCCGGAGGCTA Mismatch

NSB & SB

NSB

From Talk by Irizzary

Workflow: Analyzing Affy data



Raw data (.DAT files)

Image analysis

Probe intensities (.CEL files)

Pre-processing
normalization

Expression measures (tables)

Statistical test

Rank (list)

Choose filter
Significance level

Candidate genes (short list)

From Talk by Irizzary

Affy Files

- **DAT** file: image file, about 10 million pixels, 30-50 MB
- **CEL** file: cell intensity file with probe level PM and MM values
- **CDF** file: chip description file describing which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs)

From Talk by Irizzary

Image analysis & Background Correction

- Each probe cell: 10 X 10 pixels
- Gridding estimates location of probe cell centers
- Signal is computed by
  - Ignoring outer 36 pixels leaving a 8 X 8 pixel area
  - Taking the 75 percentile of the signal from the 8 X 8 pixel area
- Background signal is computed as the average of the lowest 2% probe cell values, which is then subtracted from the individual signals

From Talk by Irizzary

Standard Normalization Procedure

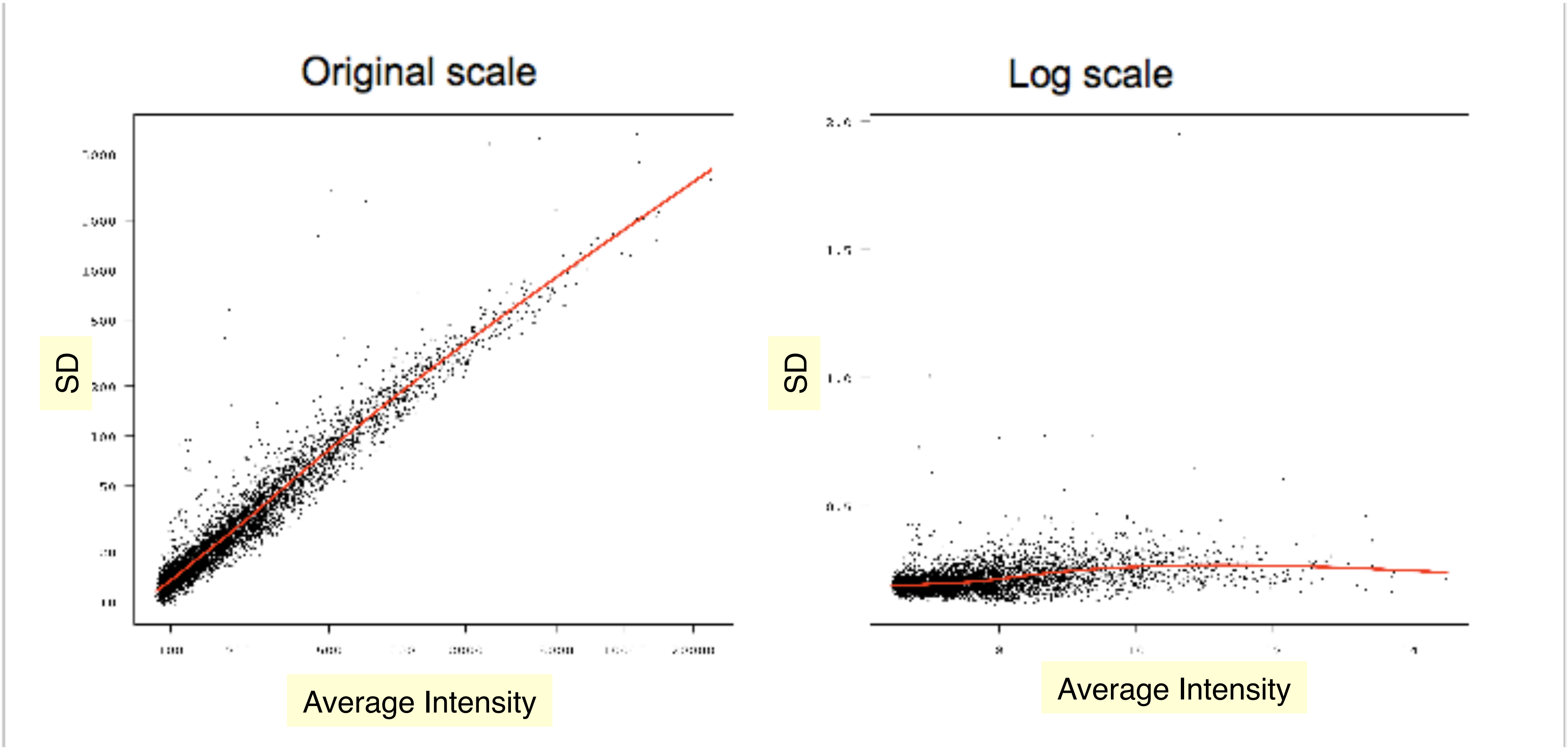
- ❑ Log-transform the data
- ❑ Ensure that the average intensity and the standard deviation are the same across all arrays.
- ❑ This requires the choice of a baseline array, which may or may not be obvious.

Analyzing Affy data

- MAS 4.0
  - Works with PM-MM
  - Negative values result very often
  - Very noisy for low expressed genes
  - Averages without log-transformation
- dChip [Li & Wong, PNAS **98**(1):31-36]
  - Accounts for probe effect
  - Uses non-linear normalization
  - Multi-chip analysis reveals outliers
- MAS 5.0
  - Improves on problems with MAS 4.0

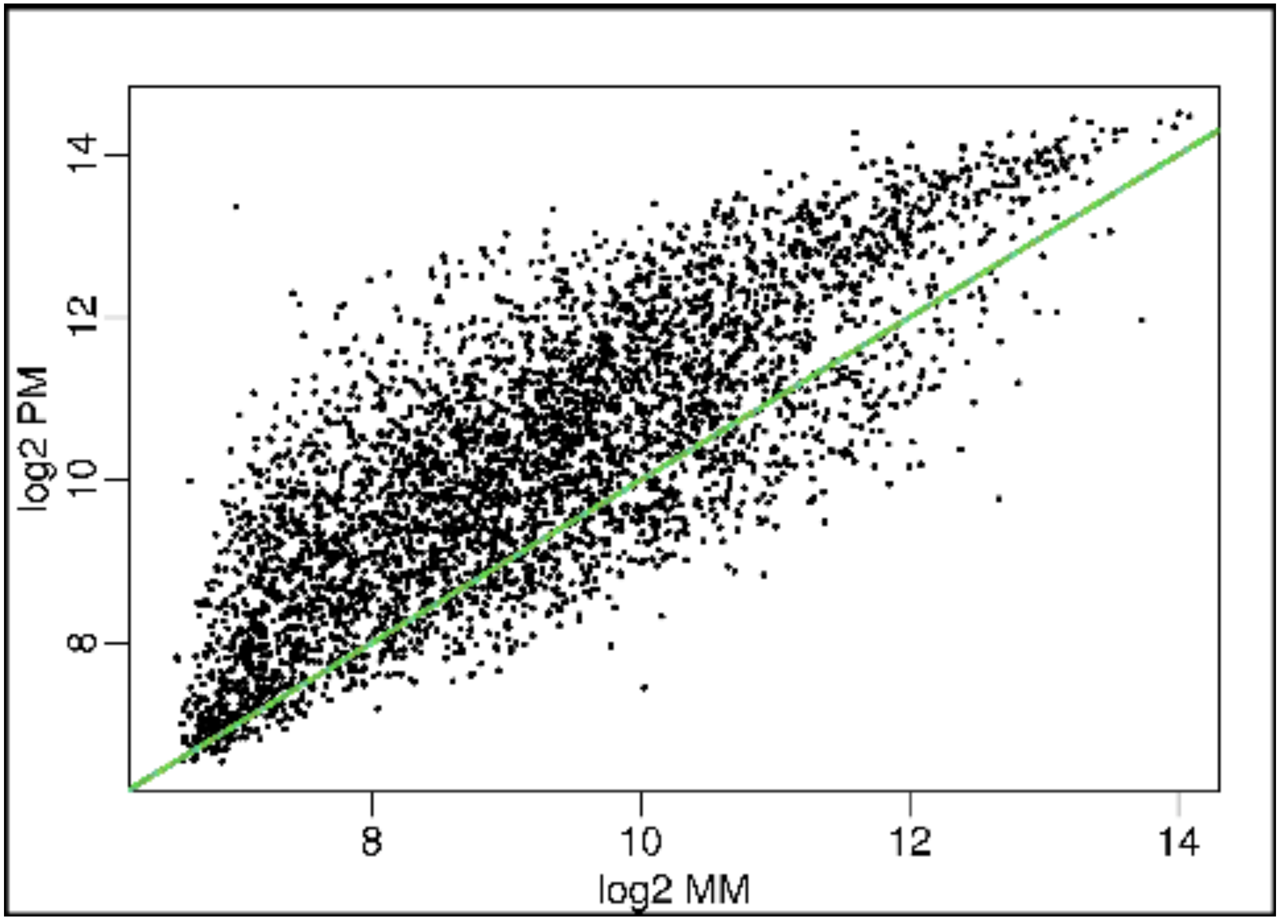
From Talk by Irizzary

Why you use log-transforms?



From Talk by Irizzary

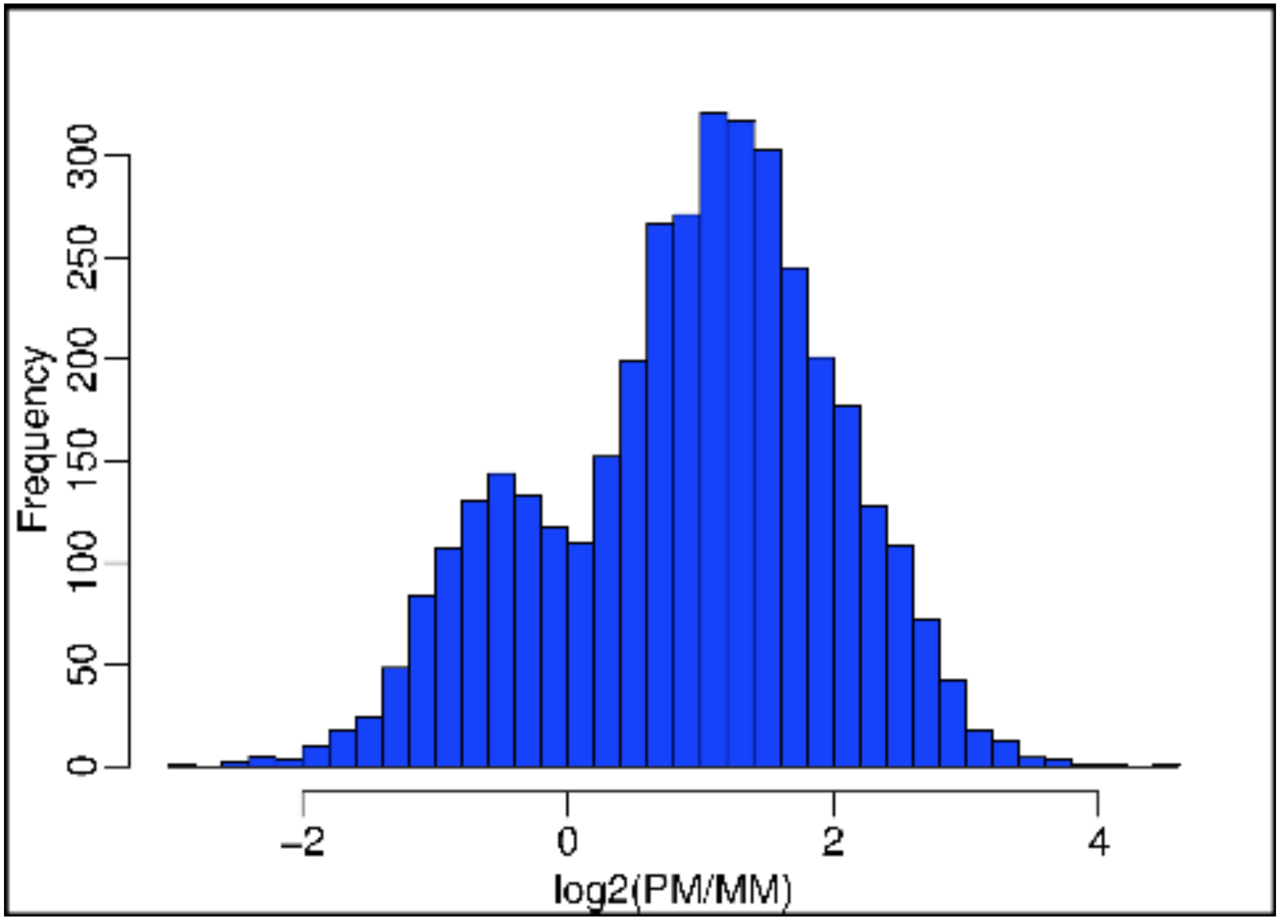
Problem with using (transformed) PM-MM



Sometimes MM is larger than PM!

From Talk by Irizzary

Bimodality for large expression values



From Talk by Irizzary

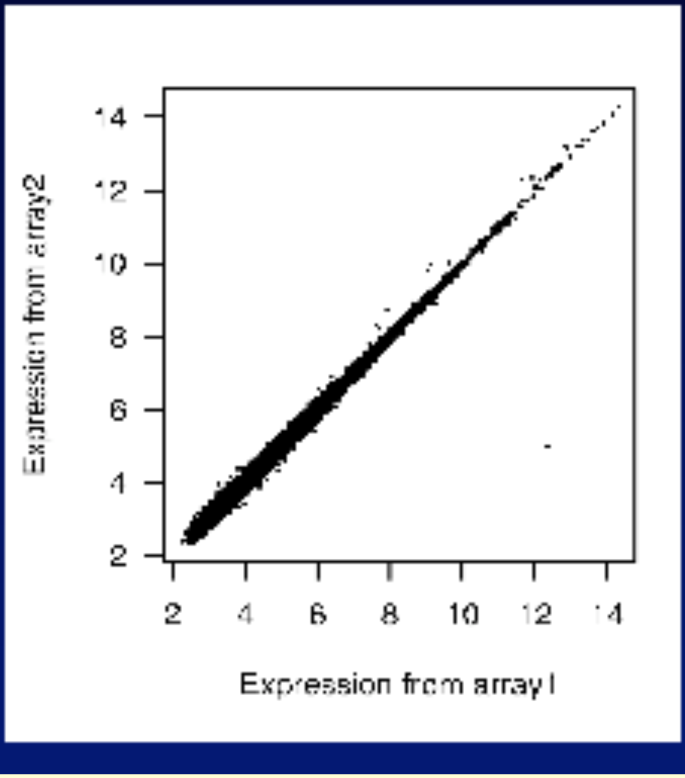
MAS 5.0

- **MAS 5.0** is Affymetrix software for microarray data analysis.
- Ad hoc background procedure used
- Summarization: Averaging over multiple probes
- For summarization, MAS 5.0 uses:
  - **Signal = TukeyBiweight**{$\log(PM_j - MM_j{}^*)$}
  - Tukey Biweight: $B(x) = (1 - (x/c)^2)^2$, if $x<c$

    $= 0$ otherwise
- Ad hoc scale normalization used

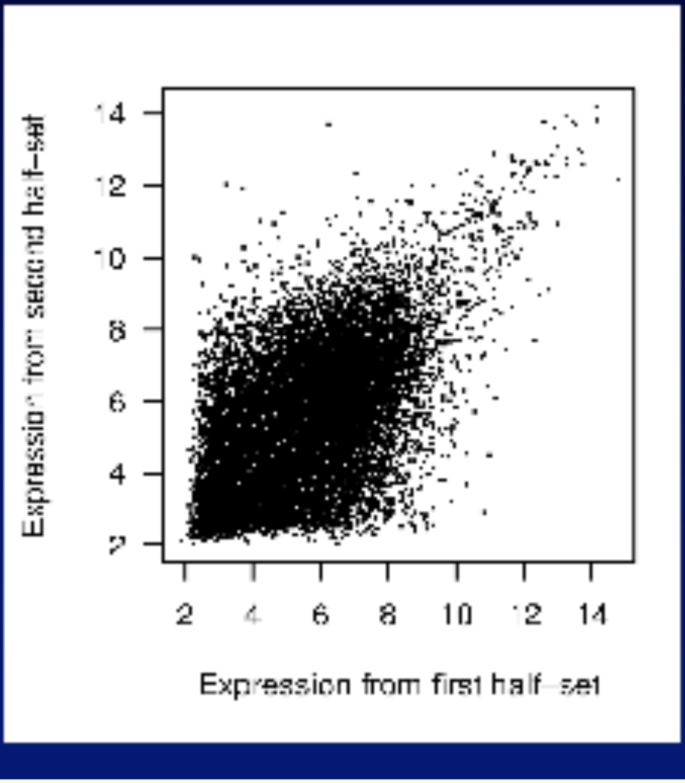
From Talk by Irizzary &
PhD thesis by Astrand

2 replicate arrays



Expression from corresponding probes are highly correlated

Correlation is higher than 0.99

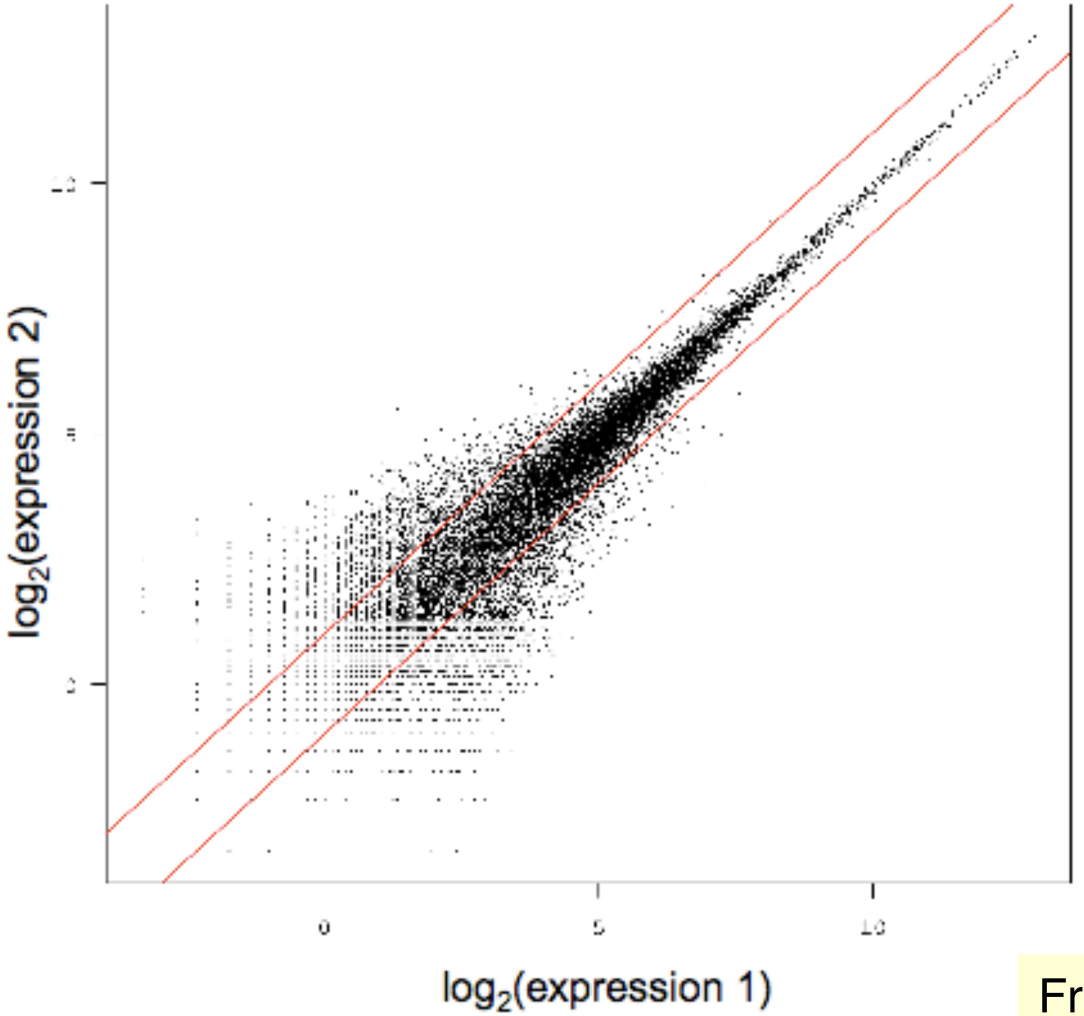


Expression not correlated when probes randomly partitioned

Correlation drops to 0.55

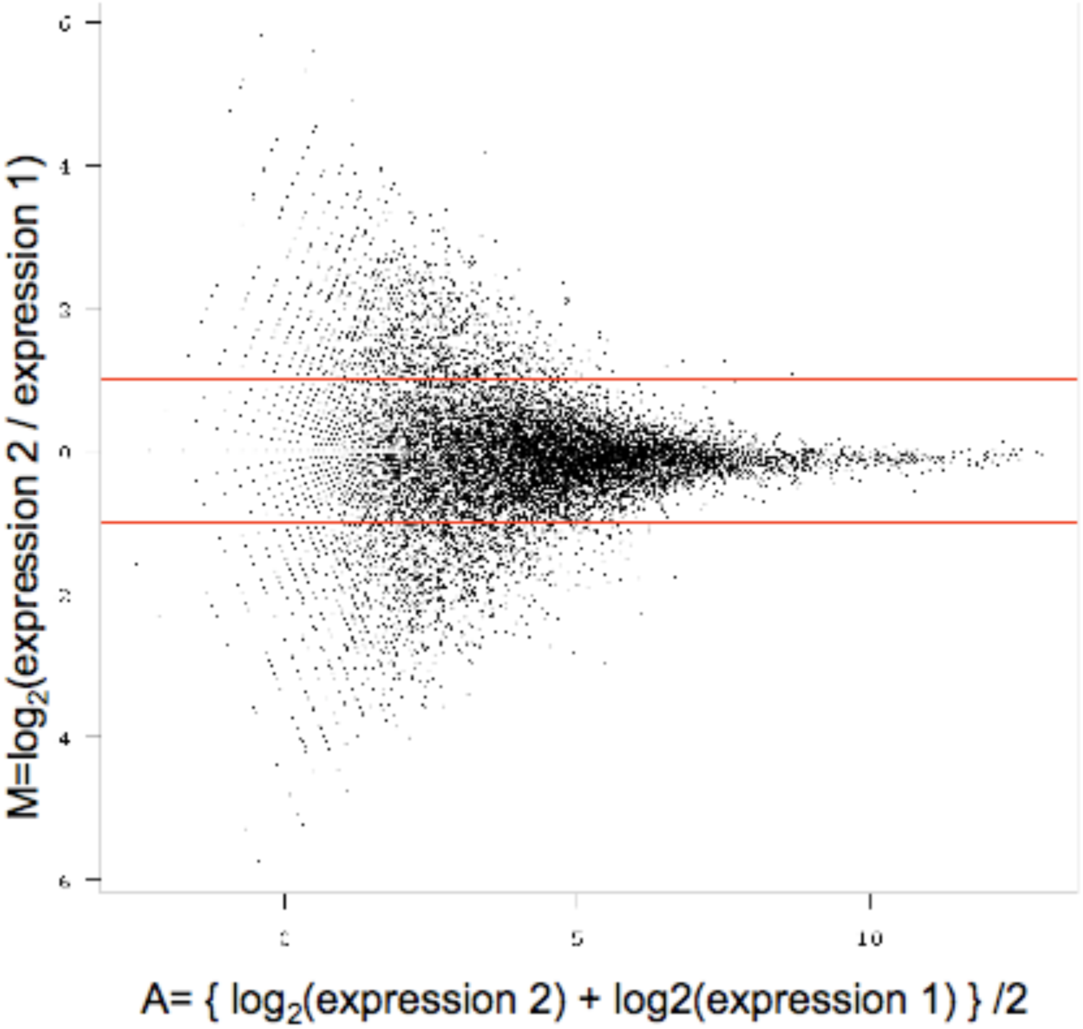
From Talk by Irizzary

We have to deal with **variations**!



From Talk by Irizzary

MvA Plots



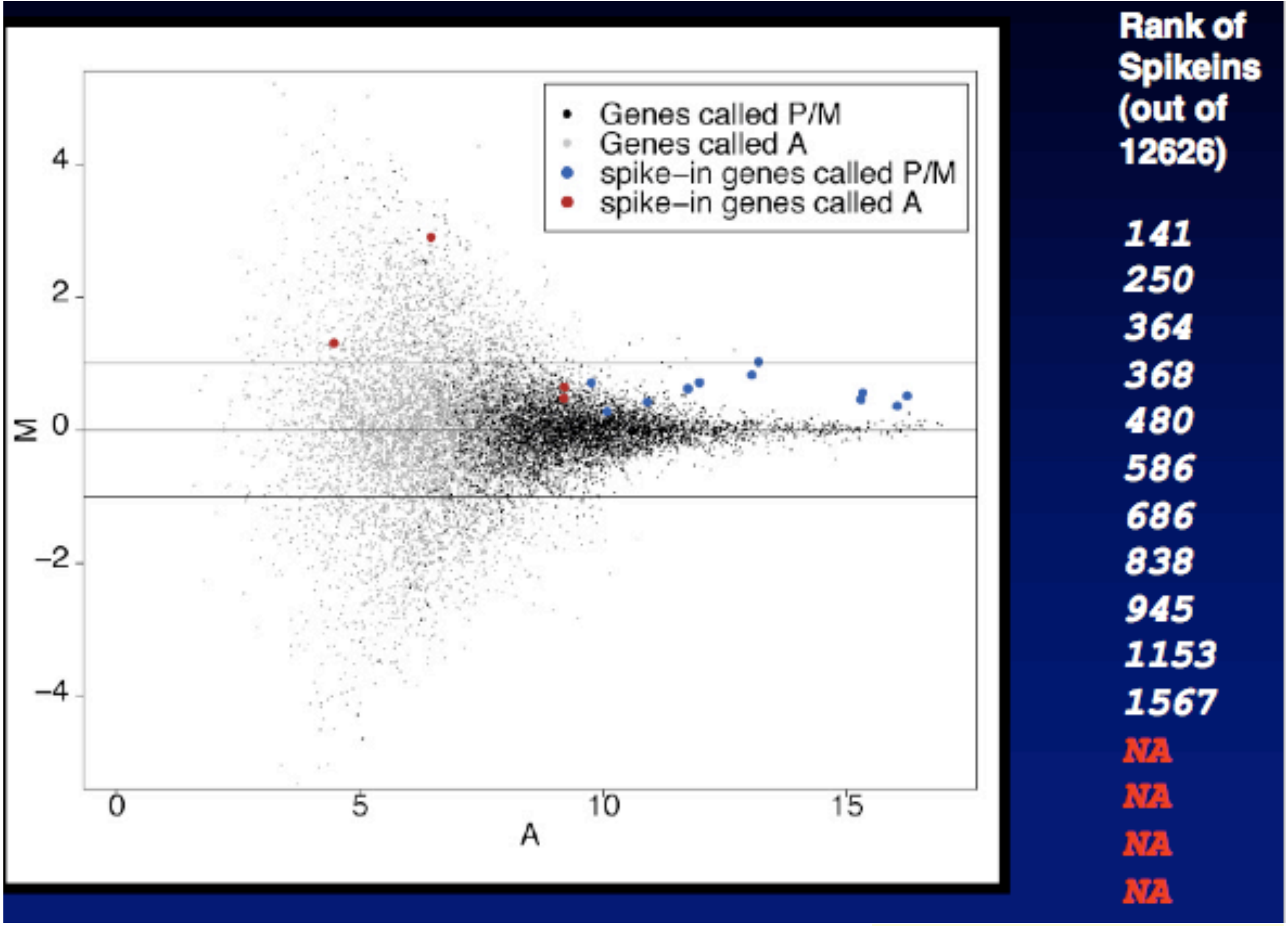
From Talk by Irizzary

Spike-in Experiment

- ❑ Replicate RNA samples were hybridized to various arrays
- ❑ Some probe sets were spiked in at different concentrations across the different arrays
- ❑ Goal was to see if these spiked probe sets “stood out” as differentially expressed

From Talk by Irizzary

Analyzing Spike-in data with MAS 5.0



Rank of Spikeins (out of 12626)

141
250
364
368
480
586
686
838
945
1153
1567
NA
NA
NA
NA

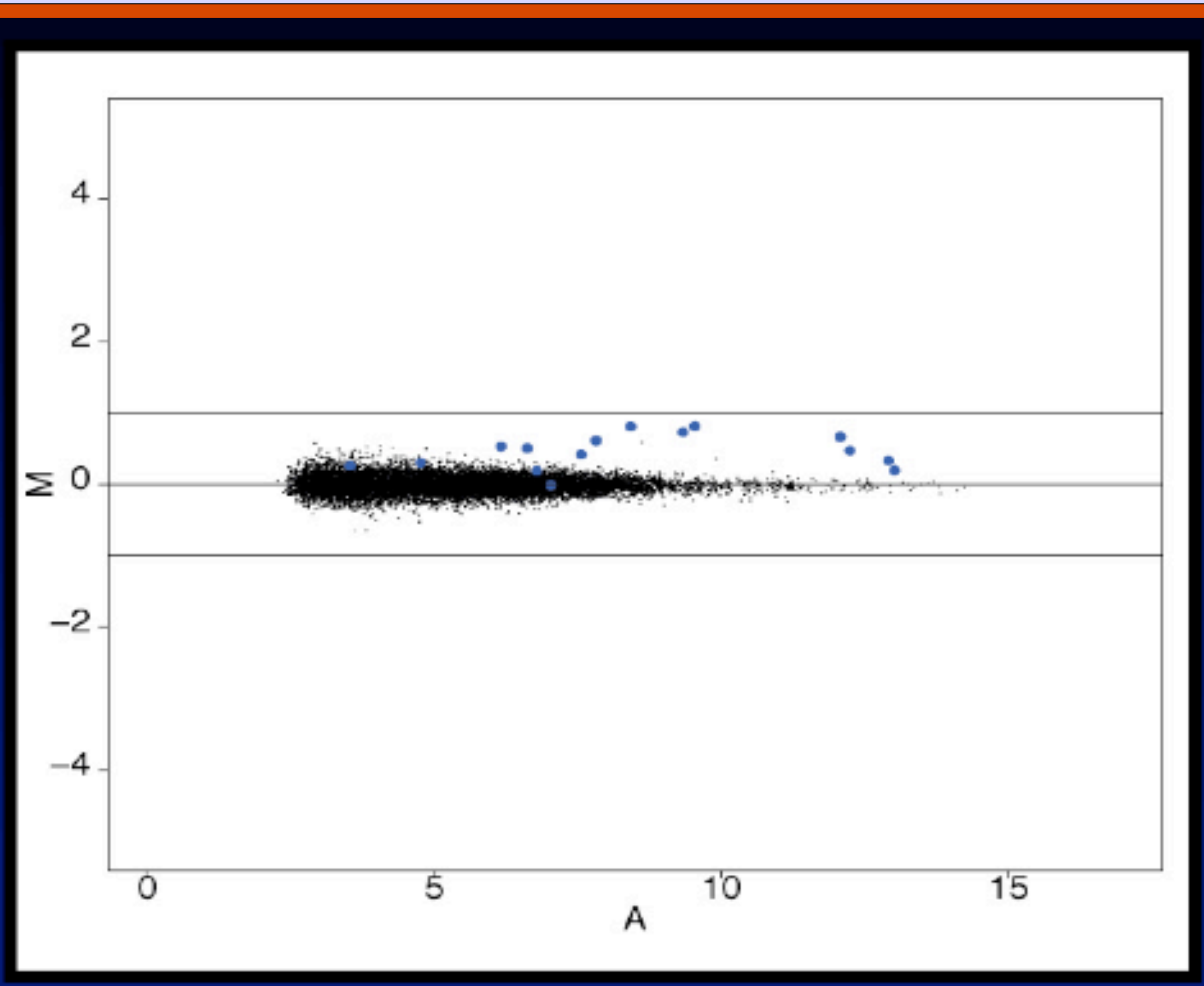
From Talk by Irizzary

Robust Multiarray normalization (RMA)

- Background correction separately for each array
  - Find E{Sig | Sig+Bgd = PM}
  - Bgd is normal and Sig is exponential
- Uses quantile normalization to achieve "identical empirical distributions of intensities" on all arrays
- Summarization: Performed separately for each probe set by fitting probe level additive model
- Uses median polish algorithm to robustly estimate expression on a specific chip
- Also see GCRMA [Wu, Irizzary et al., 2004]

From Talk by Irizzary &
PhD thesis by Astrand

Analyzing Spike-in data with RMA



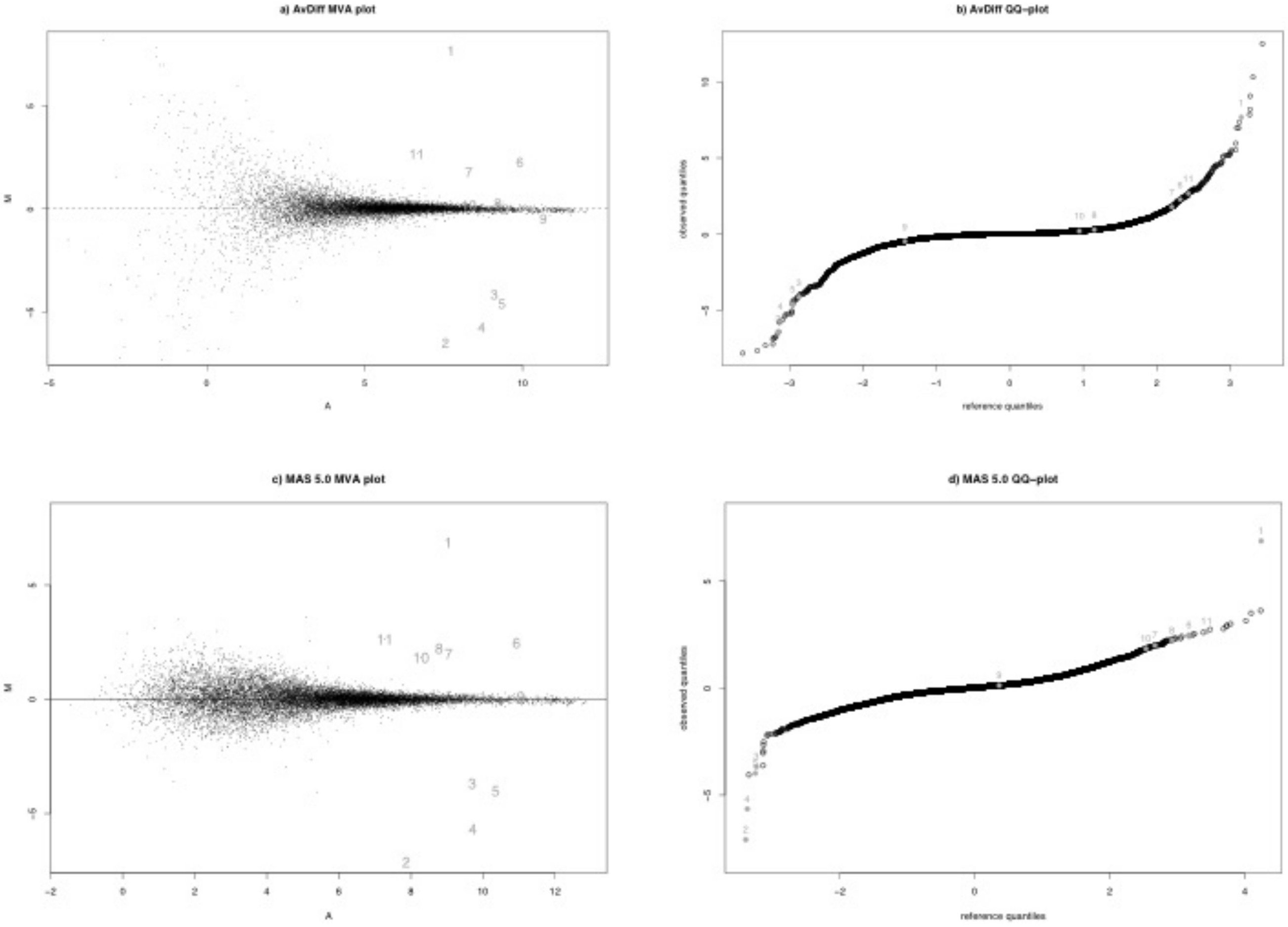
Rank of Spikeins (out of 12626)

1
2
3
4
7
11
15
21
35
122
1182
230
450
1380
11700

Irizarry et al. (2003) *NAR* 31:e15

From Talk by Irizzary

MvA and q-q plots

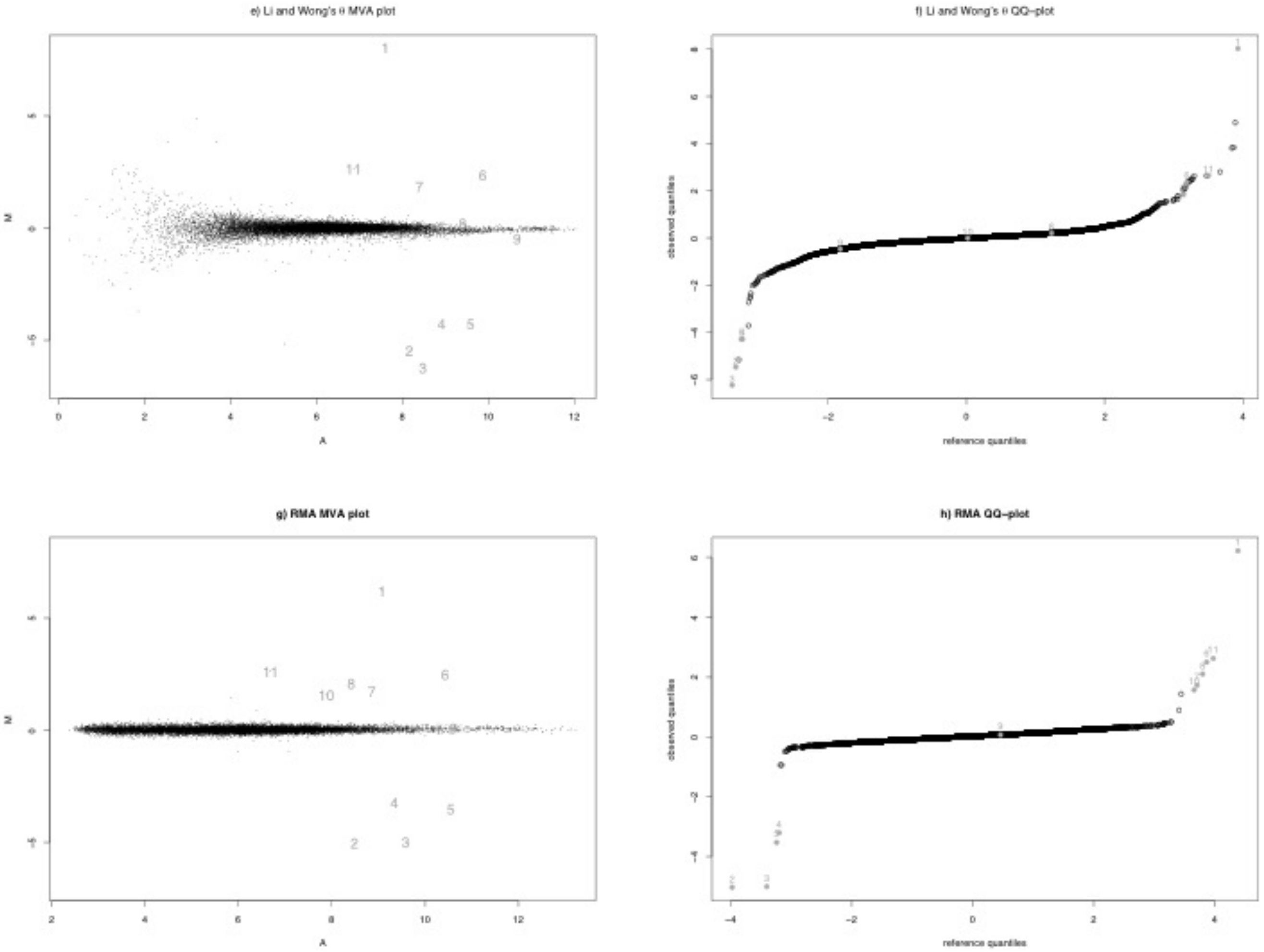


MAS 4.0

MAS 5.0

From Talk by Irizzary

MvA and q-q Plots



MBEI

RMA

From Talk by Irizzary

Before and after quantile normalization

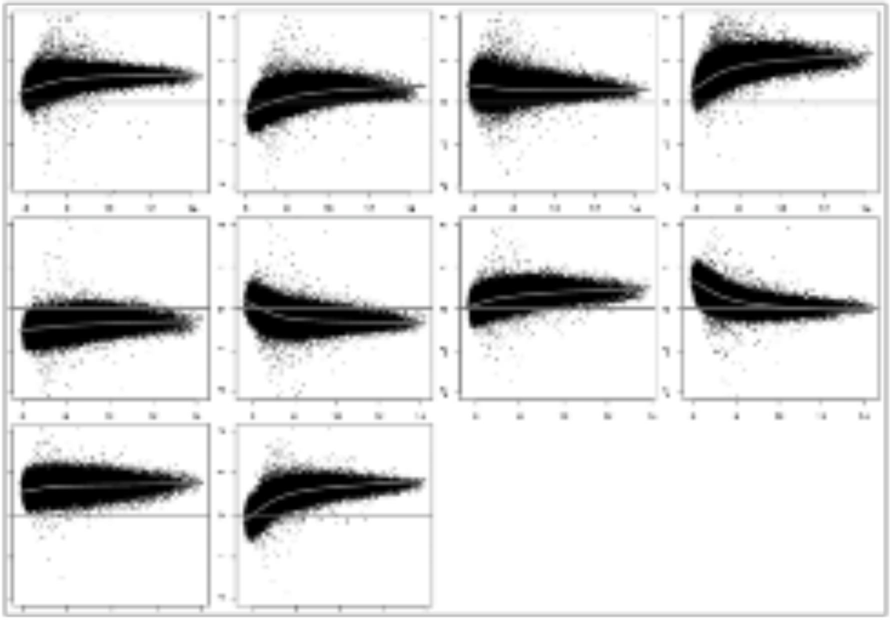


Fig. 2. 10 pairwise *M* versus *A* plots using liver (at concentration 10) dilution series data for unadjusted data.

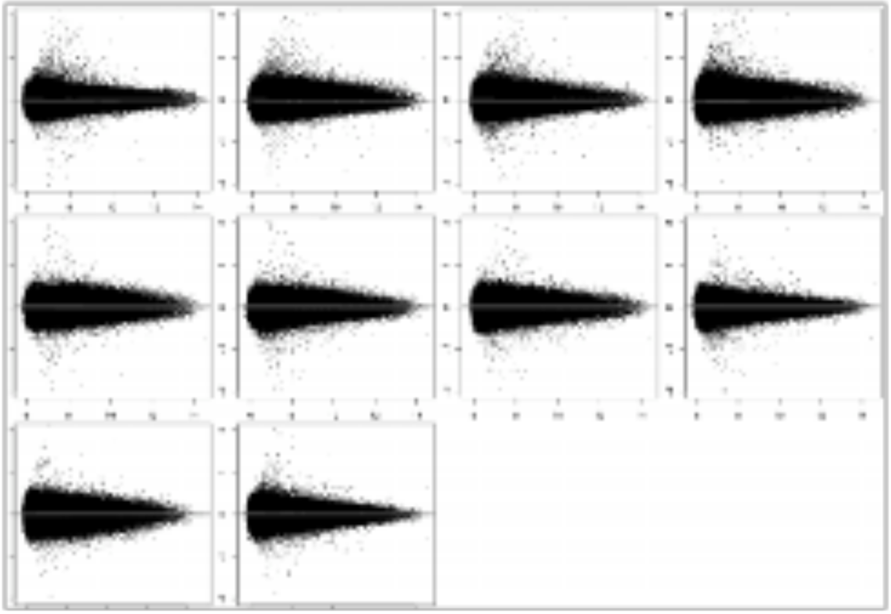


Fig. 3. 10 pairwise *M* versus *A* plots using liver (at concentration 10) dilution series data after quantile normalization.

From Talk by Irizzary

Bioconductor

- **Bioconductor** is an **open source** and open development software project for the analysis of biomedical and genomic data.
- World-wide project started in 2001
- **R** and the **R package system** are used to design and distribute software
- Commercial version of Bioconductor software called ***ArrayAnalyzer***

From Talk by Irizzary

R: A Statistical Programming Language

- ❑ Try the tutorial at: [http://www.cyclismo.org/tutorial/R/]
- ❑ Also at: [http://www.math.ilstu.edu/dhkim/Rstuff/Rtutor.html]