Measuring Gene Expression Part 3

David Wishart

Bioinformatics 301

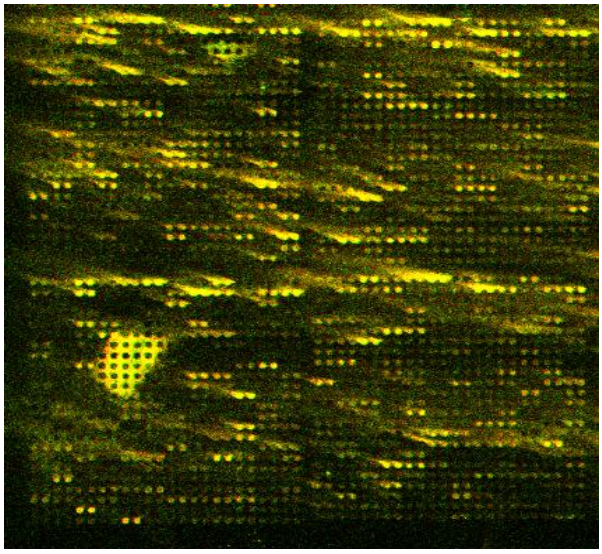
david.wishart@ualberta.ca

Key Steps in Microarray Analysis

- **Quality Control (checking microarrays for errors or problems)**
- **Image Processing**
  - **Gridding**
  - **Segmentation (peak picking)**
  - **Data Extraction (intensity, QC)**
- **Data Analysis and Data Mining**

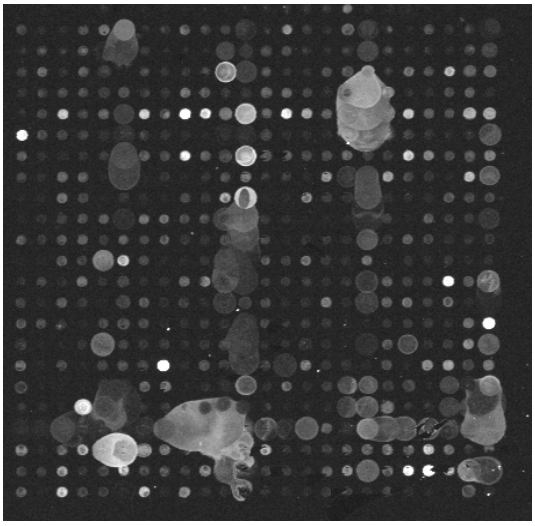
Comet Tailing

- Often caused by insufficiently rapid immersion of the slides in the succinic anhydride blocking solution.



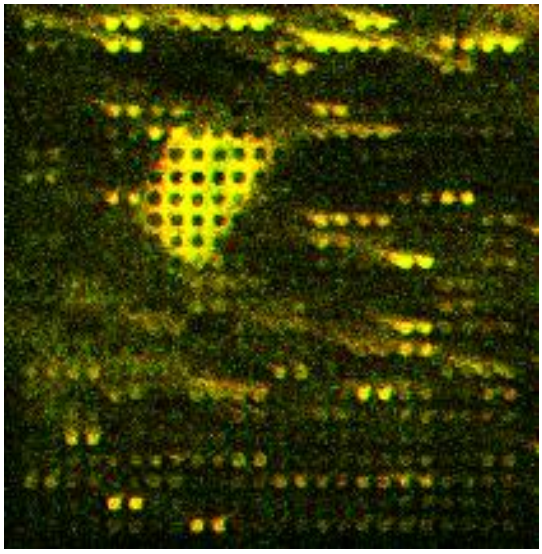
Uneven Spotting/Blotting

- Problems with print tips or with overly viscous solution
- Problems with humidity in spottiing chamber

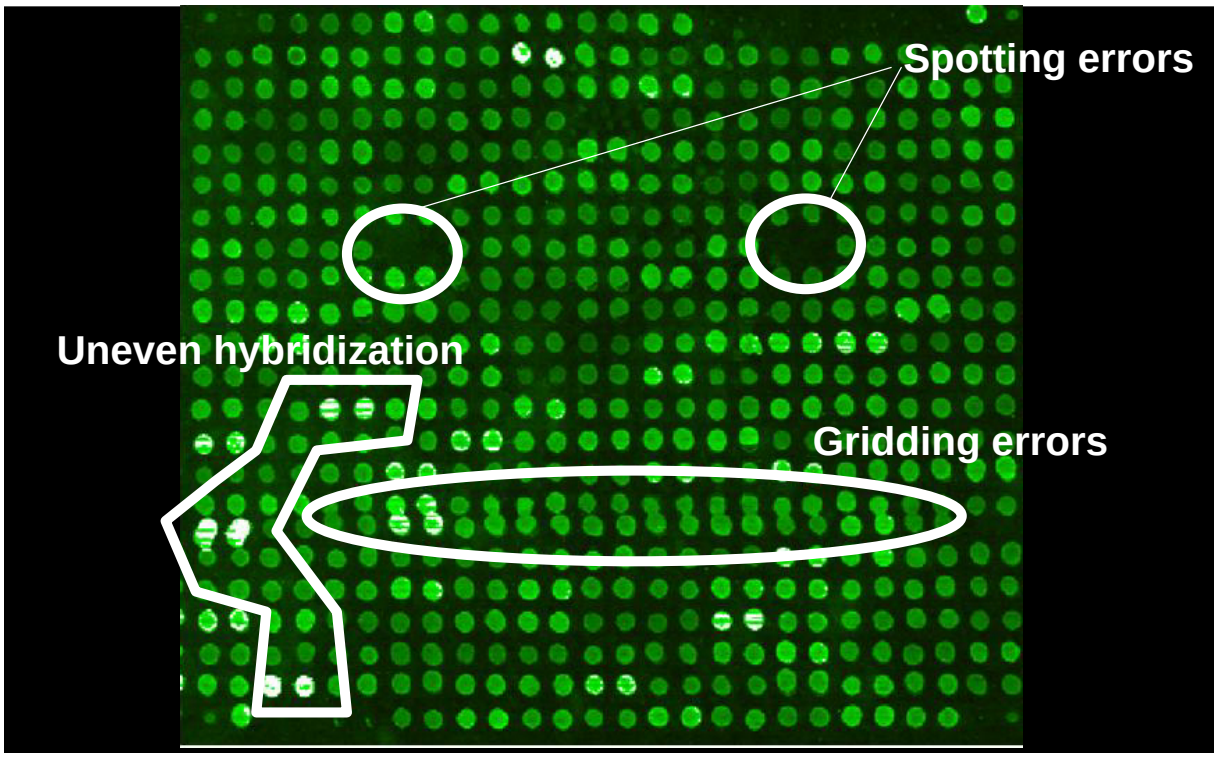


High Background

- Insufficient Blocking
- Precipitation of labelled probe



Gridding Errors

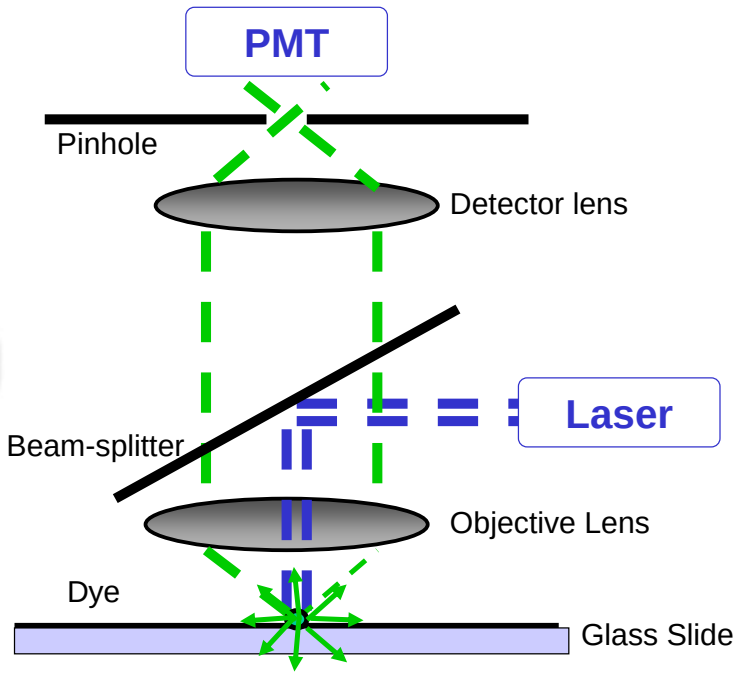


Key Steps in Microarray Analysis

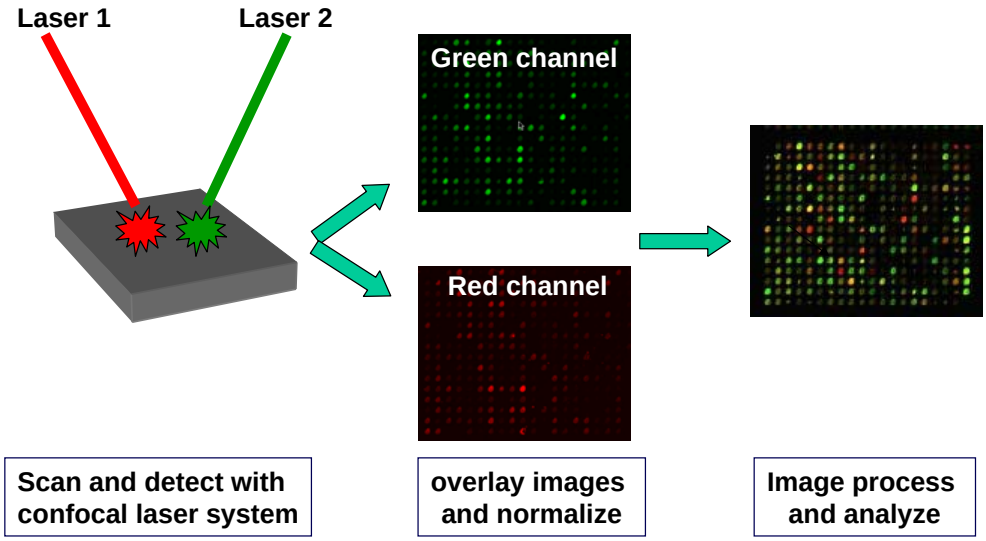
- Quality Control (checking microarrays for errors or problems)
- Image Processing
  - Gridding
  - Segmentation (spot picking)
  - Data Extraction (intensity, QC)
- Data Analysis and Data Mining

Microarray Scanning





Microarray Principles



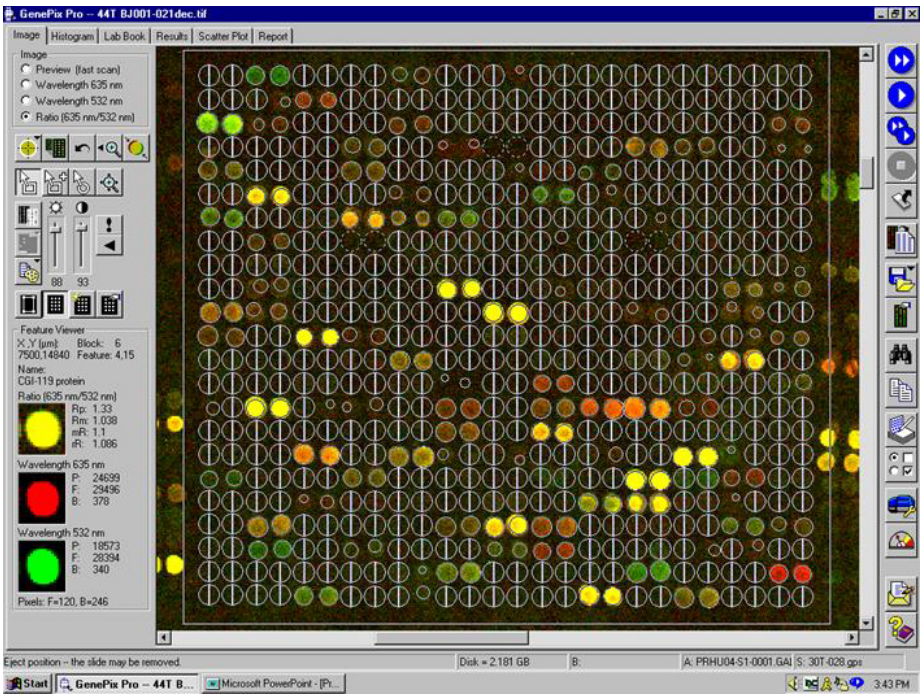
Microarray Images

- **Resolution**
  - standard 10μm [currently, max 5μm]
  - 100μm spot on chip = 10 pixels in diameter
- **Image format**
  - TIFF (tagged image file format) 16 bit (64K grey levels)
  - 1cm x 1cm image at 16 bit = 2Mb (uncompressed)
  - other formats exist i.e. SCN (Stanford University)
- **Separate image for each fluorescent sample**
  - channel 1, channel 2, etc.

Image Processing

- **Addressing or gridding**
  - **Assigning coordinates to each of the spots**
- **Segmentation or spot picking**
  - **Classifying pixels either as foreground or as background**
- **Intensity extraction (for each spot)**
  - **Foreground fluorescence intensity pairs (R, G)**
  - **Background intensities**
  - **Quality measures**

Gridding

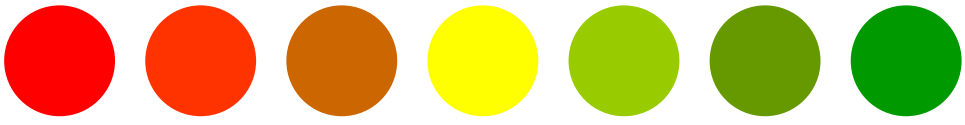


Gridding Considerations

- Separation between rows and columns of grids
- Individual translation of grids
- Separation between rows and columns of spots within each grid
- Small individual translation of spots
- Overall position of the array in the image
- *Automated & manual methods available*

Spot Picking

- Classification of pixels as foreground or background (fluorescence intensities determined for each spot are a measure of transcript abundance)
- Large selection of methods available, each has strengths & weaknesses

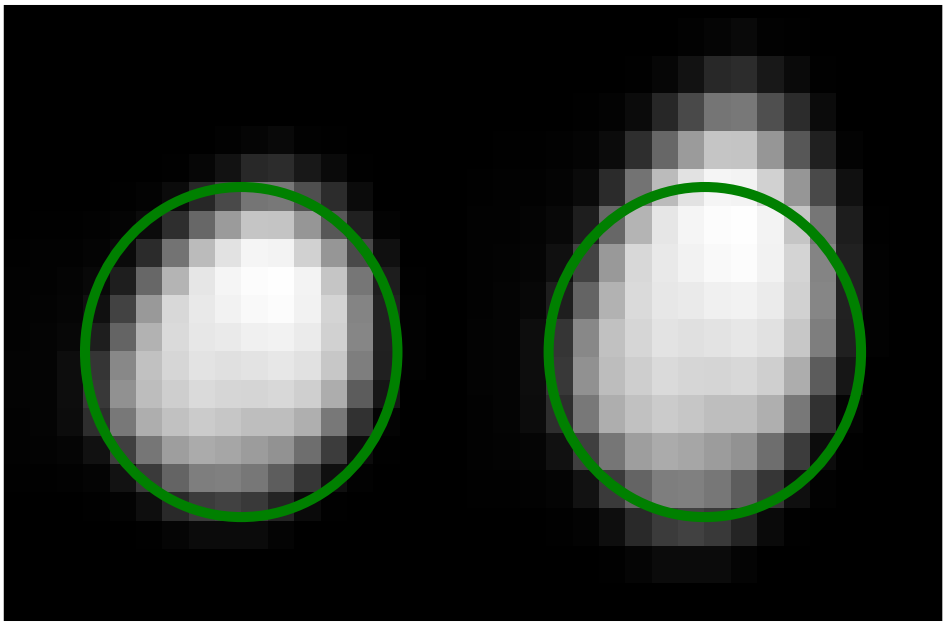


Spot Picking

- **Segmentation/spot picking methods:**
  - **Fixed circle segmentation**
  - **Adaptive circle segmentation**
  - **Adaptive shape segmentation**
  - **Histogram segmentation**

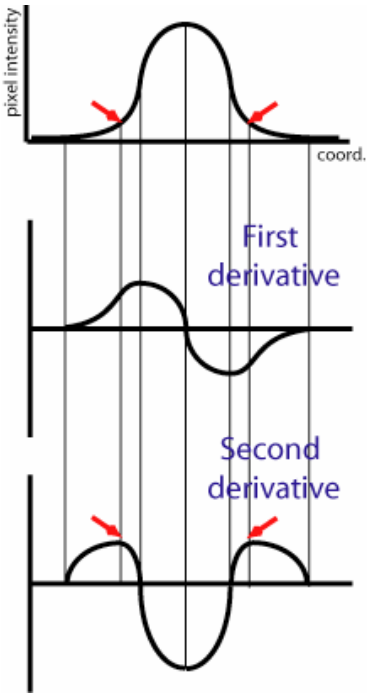
Fixed circle	**ScanAlyze, GenePix, QuantArray**
Adaptive circle	**GenePix, Dapple**
Adaptive shape	**Spot, region growing and watershed**
Histogram method	**ImaGene, QuantArraym DeArray and adaptive thresholding**

Fixed Circle Segmentation

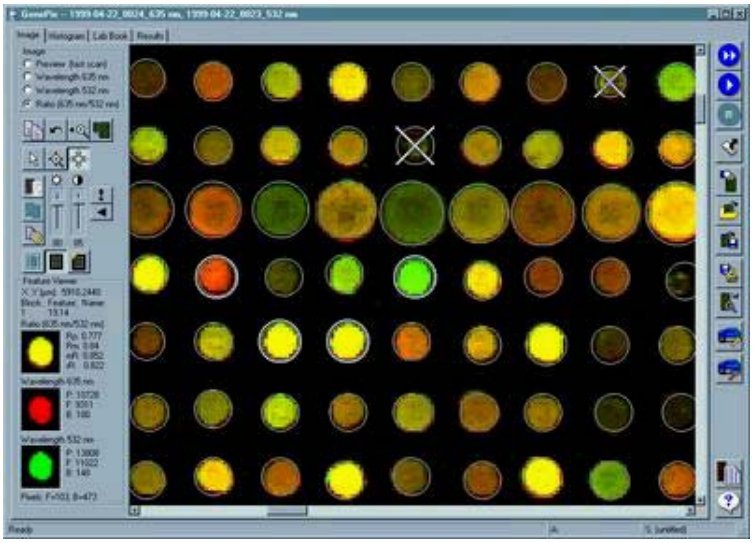


Adaptive Circle Segmentation

- **The circle diameter is estimated separately for each spot**
- ***GenePix* finds spots by detecting edges of spots (second derivative)**
- **Problematic if spot exhibits oval shapes**

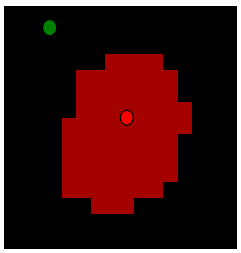


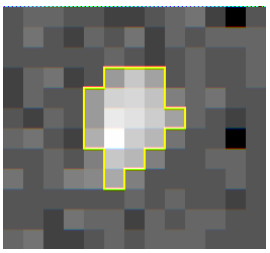
Adaptive Circle Segmentation



Adaptive Shape Segmentation

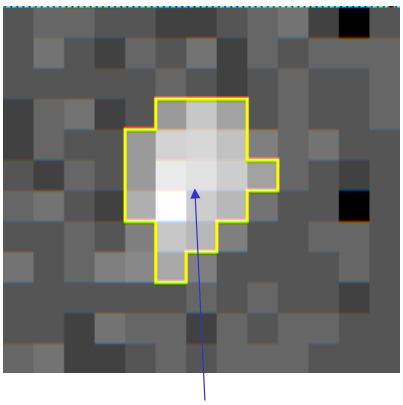
Edge detection or *Seeded Region Growing*
(R. Adams and L. Bishof (1994) - Regions grow outwards from the seed points preferentially according to the difference between a pixel's value and the running mean of values in an adjoining region





Information Extraction

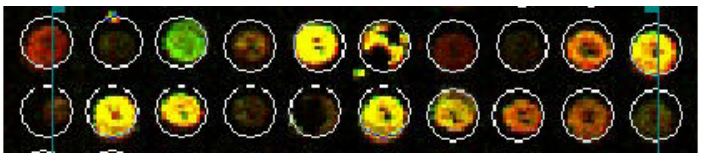
- **Spot Intensities**
  - mean (pixel intensities)
  - median (pixel intensities)
- **Background values**
  - Local Background
  - Morphological opening
  - Constant (global)
- **Quality Information**



Spot Intensity

- The total amount of hybridization for a spot is proportional to the *total fluorescence* at the spot
- Spot intensity = sum of pixel intensities within the spot mask
- Since later calculations are based on ratios between cy5 and cy3, we compute the average* pixel value over the spot mask
- Can use ratios of medians instead of means

Means vs. Medians

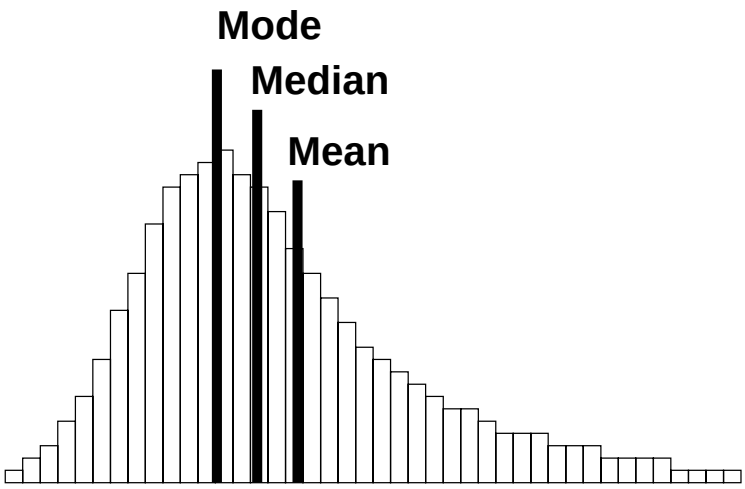


row	col	ch1_sig_mea	ch2_sig_mea	ch1_sig_med	ch2_sig_med
1	1	56000	2000	58000	1900
1	2	1000	600	600	800
1	3	2000	60000	3000	59000

↓

etc.

Mean, Median & Mode



Mean, Median, Mode

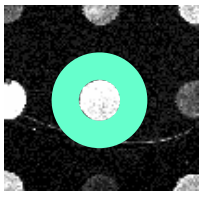
- In a Normal Distribution the mean, mode and median are all equal
- In skewed distributions they are unequal
- **Mean** - average value, affected by extreme values in the distribution
- **Median** - the “middlemost” value, usually half way between the mode and the mean
- **Mode** - most common value

Background Intensity

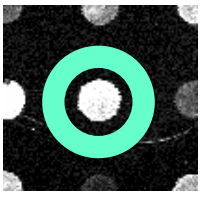
- A spot's measured intensity includes a contribution of non-specific hybridization and other chemicals on the glass
- Fluorescence intensity from regions not occupied by DNA can be different from regions occupied by DNA

Local Background Methods

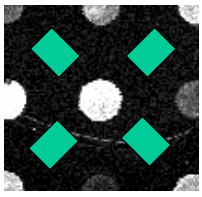
- Focuses on small regions around spot mask
- Determine median pixel values in this region
- Most common approach



ScanAlyze



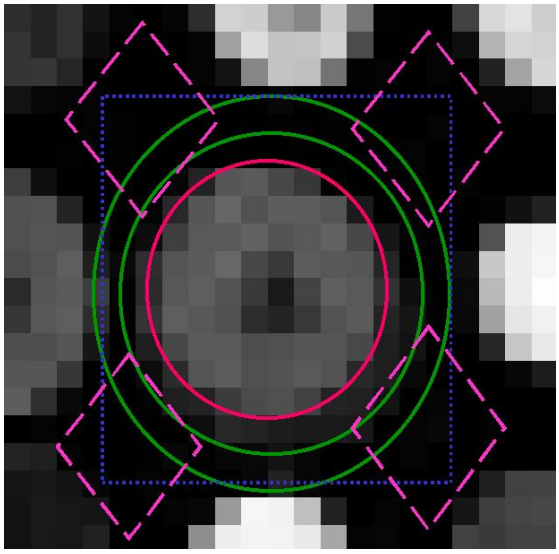
ImaGene



Spot, GenePix

- ***By not considering the pixels immediately surrounding the spots, the background estimate is less sensitive to the performance of the segmentation procedure***

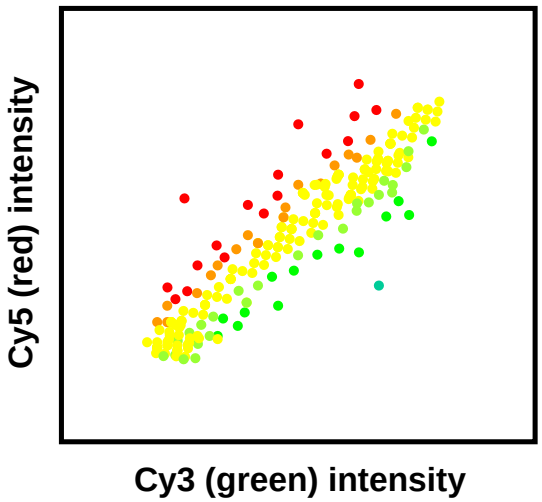
Local Background Methods



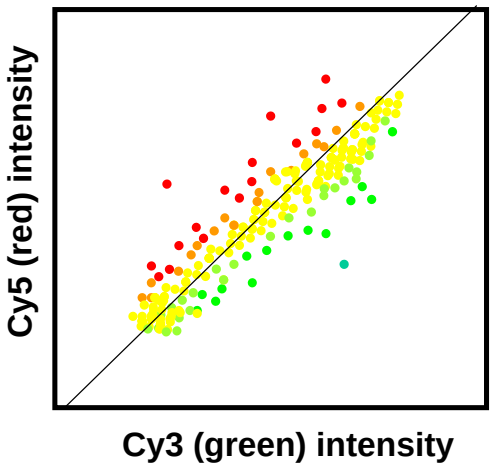
Quality Measurements

- **Array**
  - ***Correlation between spot intensities***
  - **Percentage of spots with no signals**
  - **Distribution of spot signal area**
  - **Inter-array consistency**
- **Spot**
  - **Signal / Noise ratio**
  - **Variation in pixel intensities**
  - **ID of “bad spots” (spots with no signal)**

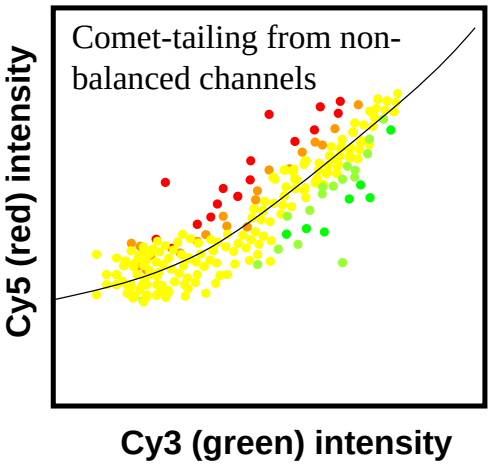
A Microarray Scatter Plot



Correlation

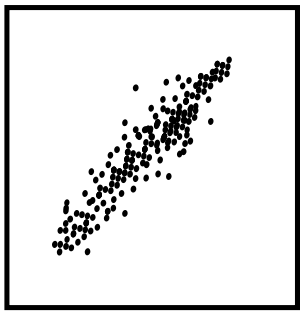


Linear

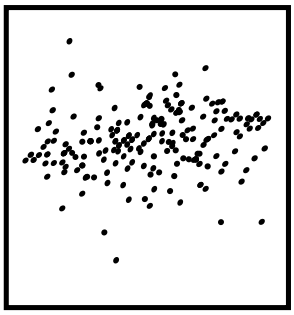


Non-linear

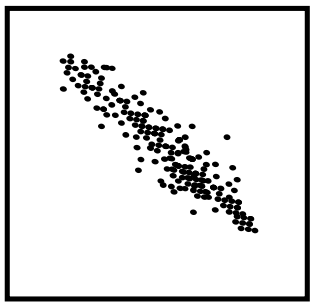
Correlation



“+” correlation

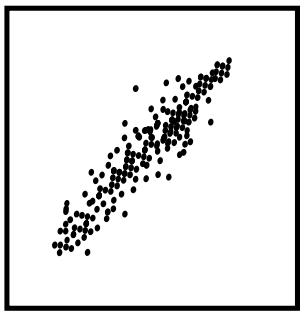


Uncorrelated



“-” correlation

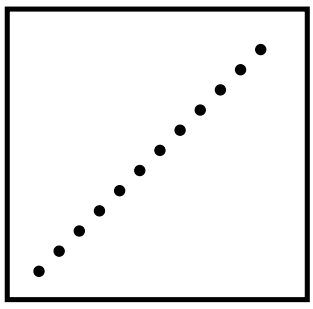
Correlation



High correlation



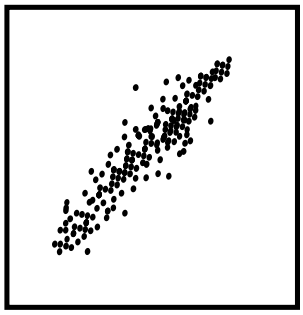
Low correlation



Perfect correlation

Correlation Coefficient

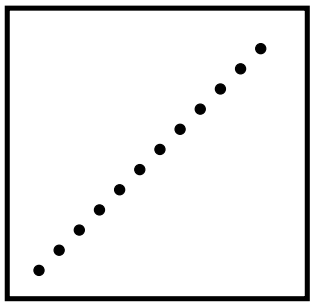
$$r = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\Sigma(x_i - \mu_x)^2(y_i - \mu_y)^2}}$$



r = 0.85



r = 0.4

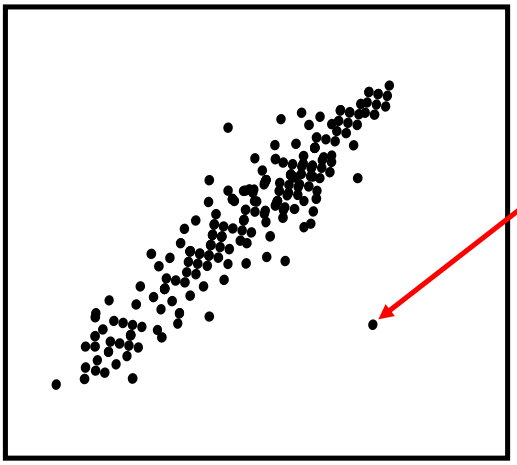


r = 1.0

Correlation Coefficient

- Sometimes called coefficient of linear correlation or Pearson product-moment correlation coefficient
- A quantitative way of determining what model (or equation or type of line) best fits a set of data
- Commonly used to assess most kinds of predictions or simulations

Correlation and Outliers



A single “bad” point can destroy a good correlation

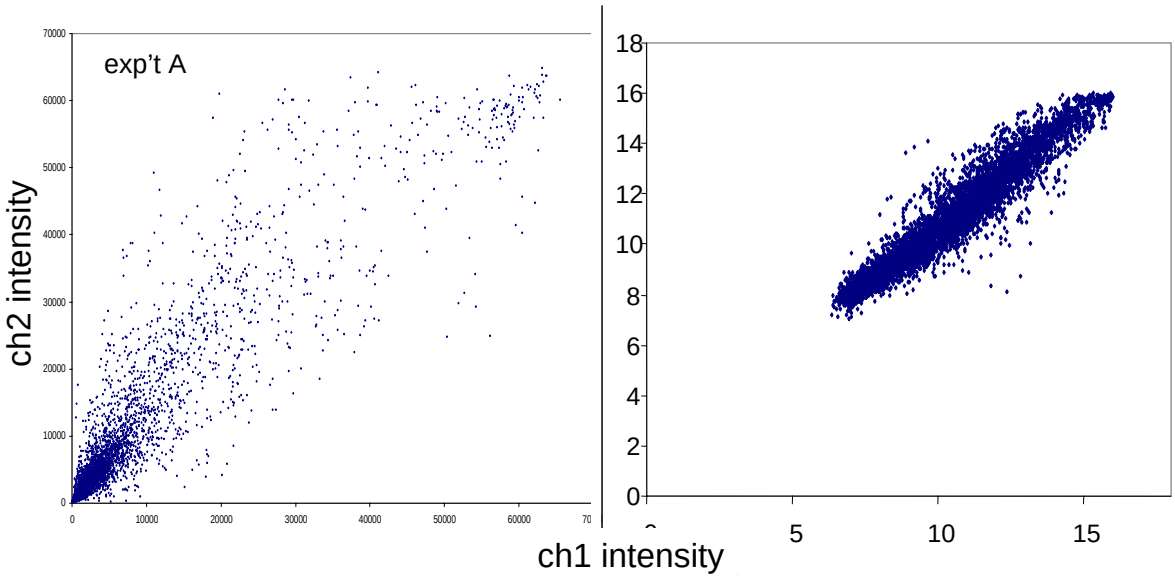
Outliers

- **Can be both “good” and “bad”**
- **When modeling data -- you don’t like to see outliers (suggests the model is bad)**
- **Often a good indicator of experimental or measurement errors -- only you can know!**
- **When plotting gel or microarray expression data you do like to see outliers**
- **A good indicator of something significant**

Log Transformation

linear scale

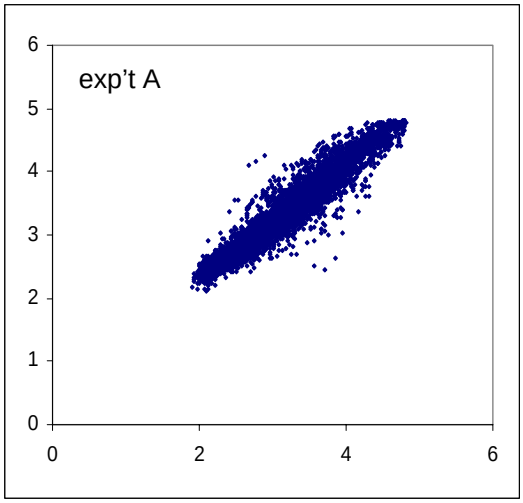
$\log_2$ scale

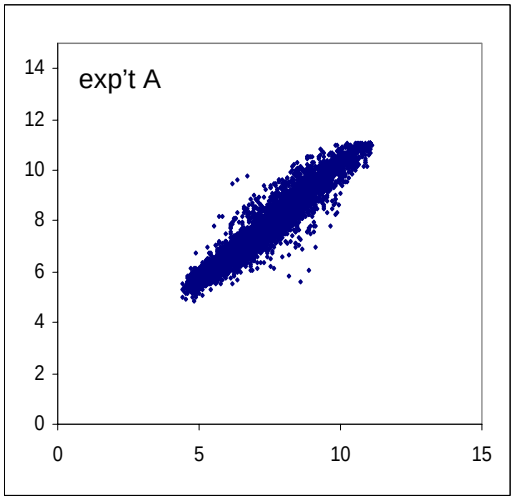


Choice of Base is Not Important

$\log_{10}$

ln





Why Log_2 Transformation?

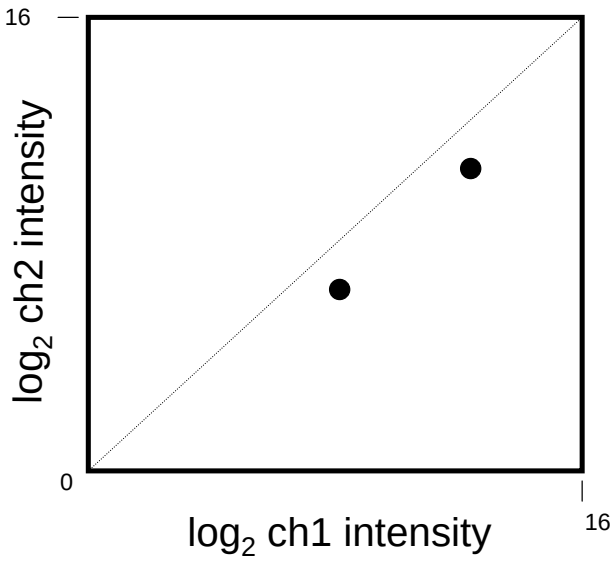
- Makes variation of intensities and ratios of intensities more independent of absolute magnitude
- Makes normalization additive
- Evens out highly skewed distributions
- Gives more realistic sense of variation
- Approximates normal distribution
- Treats up- and down- regulated genes symmetrically

Log Transformations

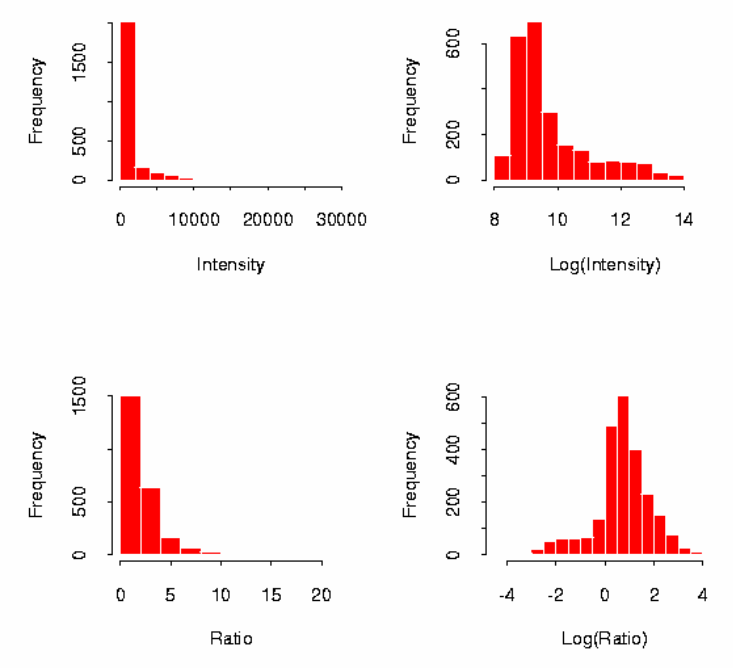
Applying a log transformation makes the variance and offset more proportionate along the entire graph

ch1	ch2	ch1/ch2
60 000	40 000	1.5
3000	2000	1.5

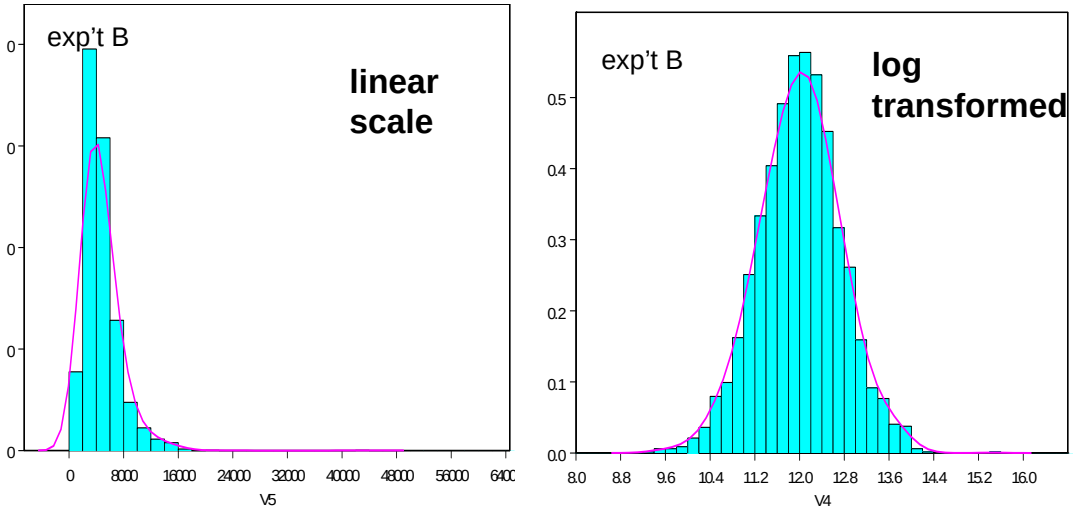
log_2 ch1	log_2 ch2	log_2 ratio
15.87	15.29	0.58
11.55	10.97	0.58



Log Transformations



Log Transformation



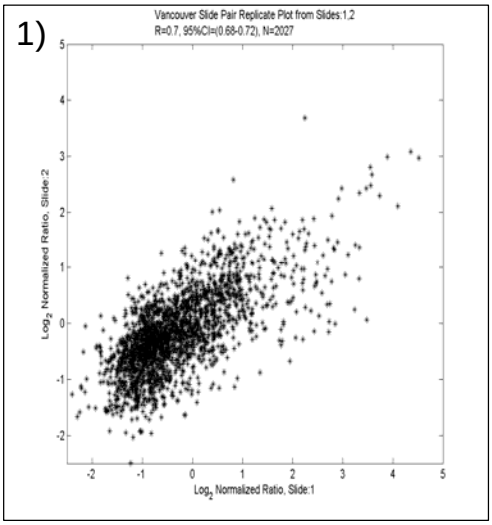
Normalization

- **Reduces systematic (multiplicative) differences between two channels of a single hybridization or differences between hybridizations**
- **Several Methods:**
  - **Global mean method**
  - **(Iterative) linear regression method**
  - **Curvilinear methods (e.g. loess)**
  - **Variance model methods**

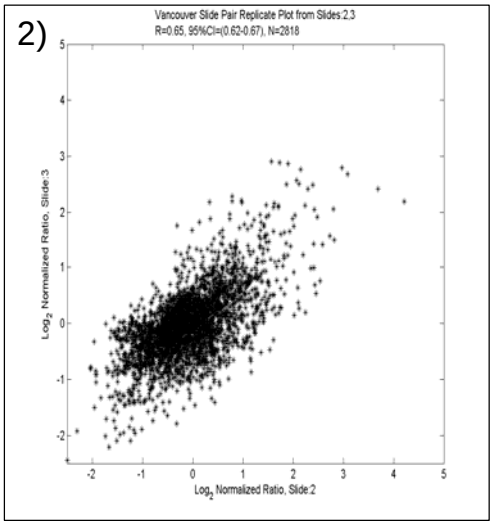
Try to get a slope ~1 and a correlation of ~1

Example Where Normalization is Needed

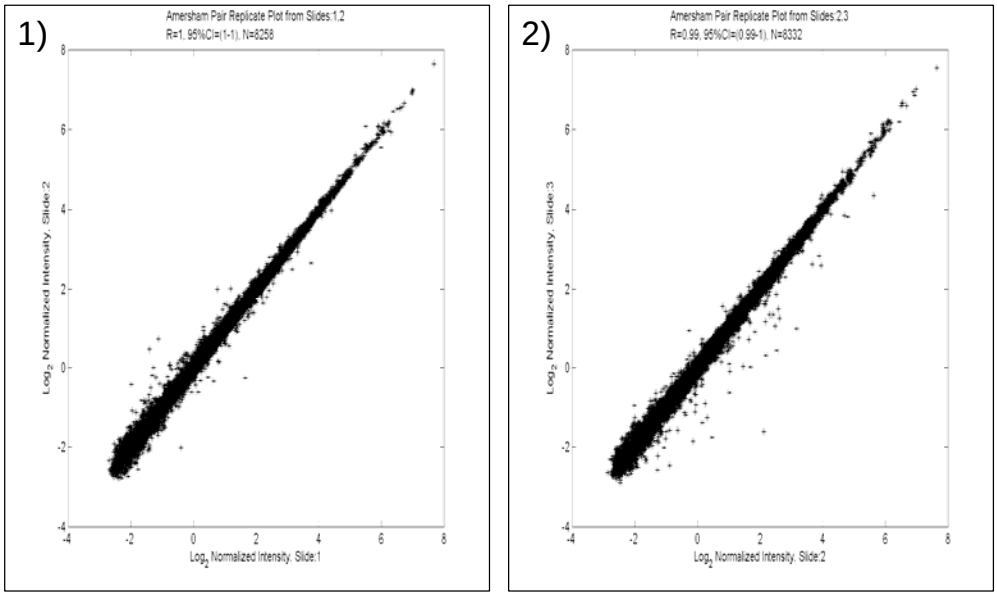
1)



2)



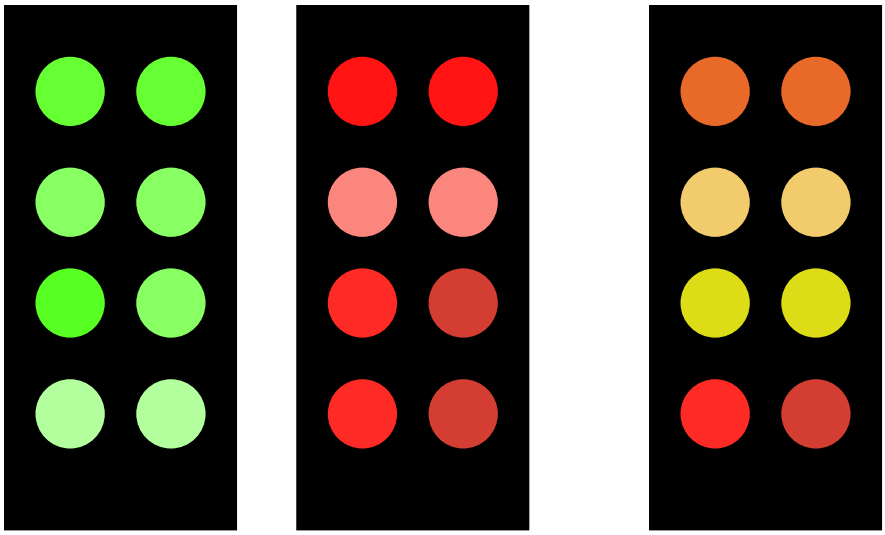
Example Where Normalization is Not Needed



Normalization to a Global Mean

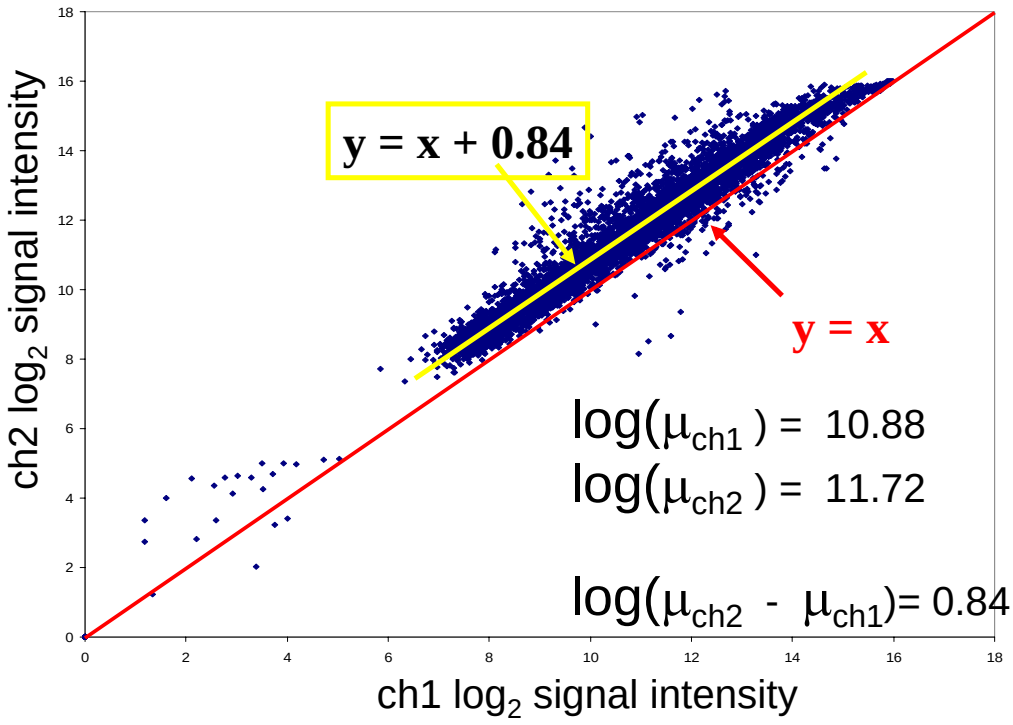
- **Calculate mean intensity of all spots in ch1 and ch2**
  - e.g. μch2 = 25 000 μch2/μch1 = 1.25
  - μch1 = 20 000
- **On average, spots in ch2 are 1.25X brighter than spots in ch1**
- **To normalize, multiply spots in ch1 by 1.25**

Hypothetical Data

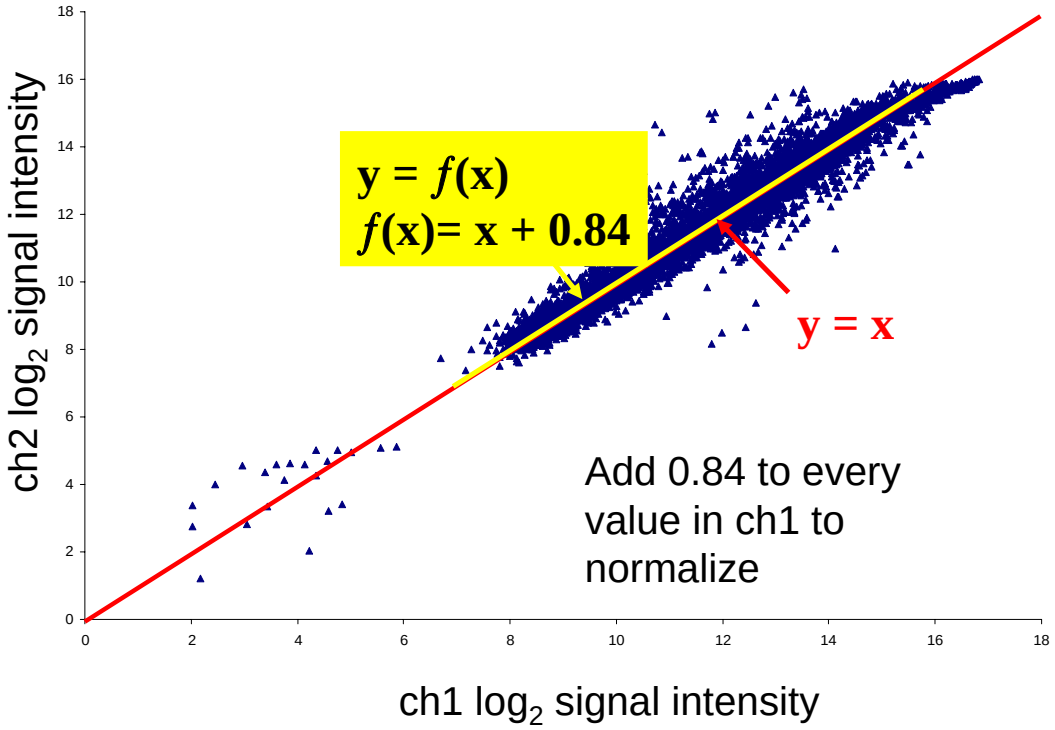


Ch 1 Ch 2 Ch1 + Ch2

Pre-normalized Data



Normalized Microarray Data



Finkelstein et al. (2001) www.camda.edu

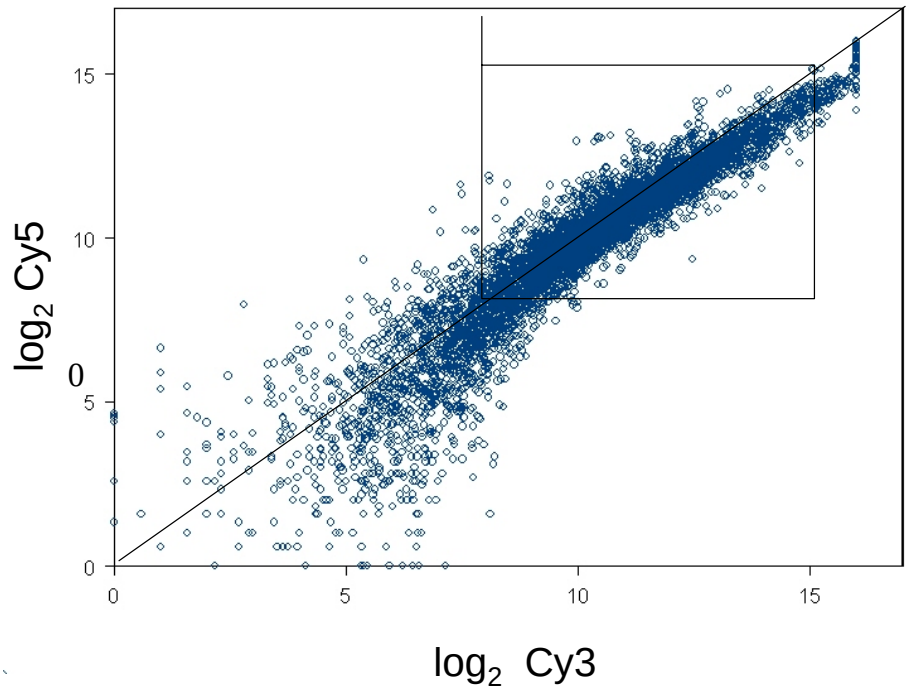
Normalization by Iterative Linear Regression

fit a line (y=mx+b) to the data set

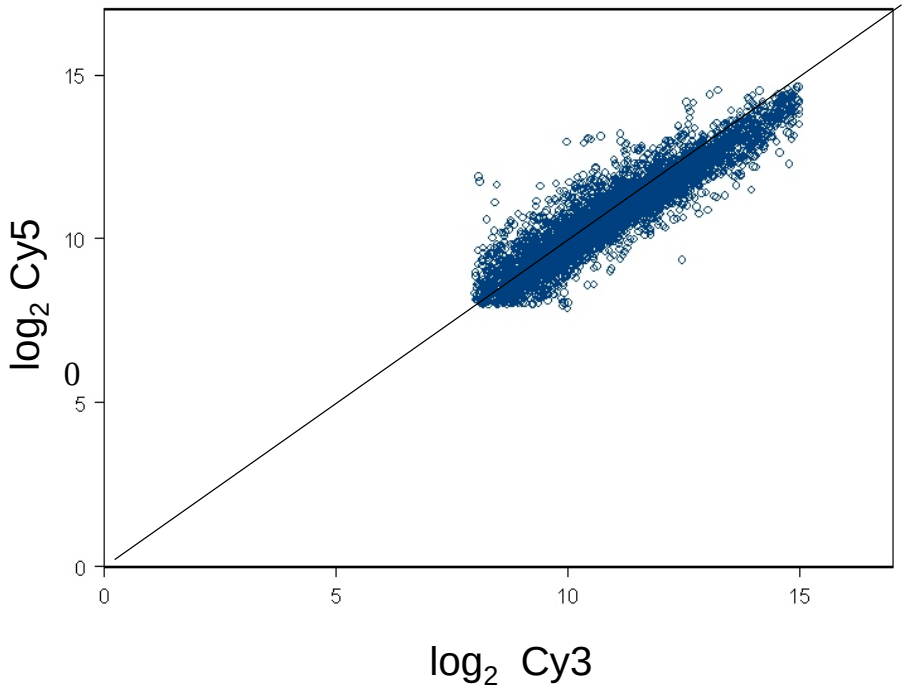
set aside outliers (residuals > 2 x s.d.)

repeat until r^2 changes by < 0.001 → then apply slope and intercept to the original dataset

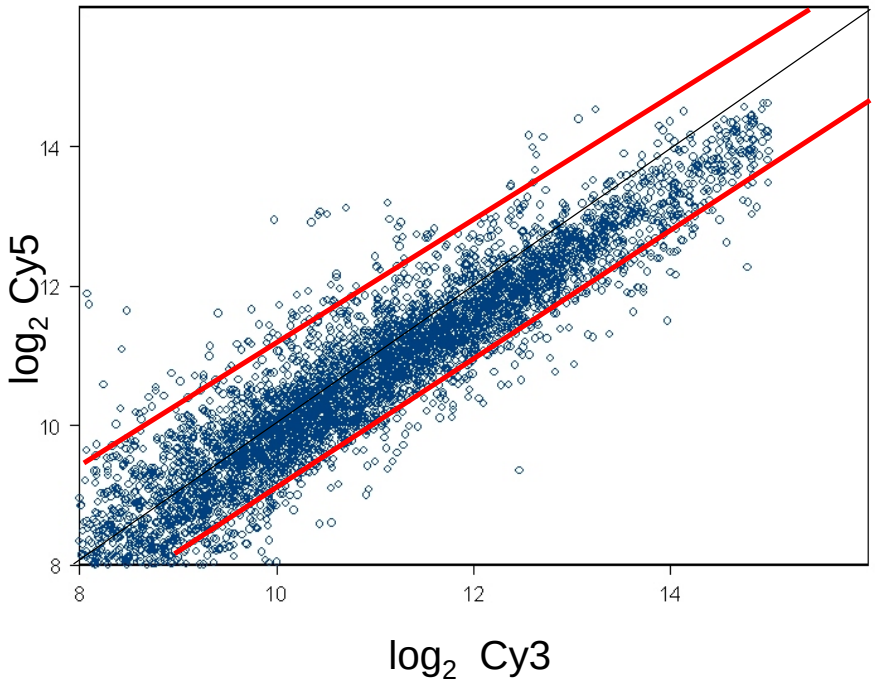
Channel 1 vs. 2 – Raw Data



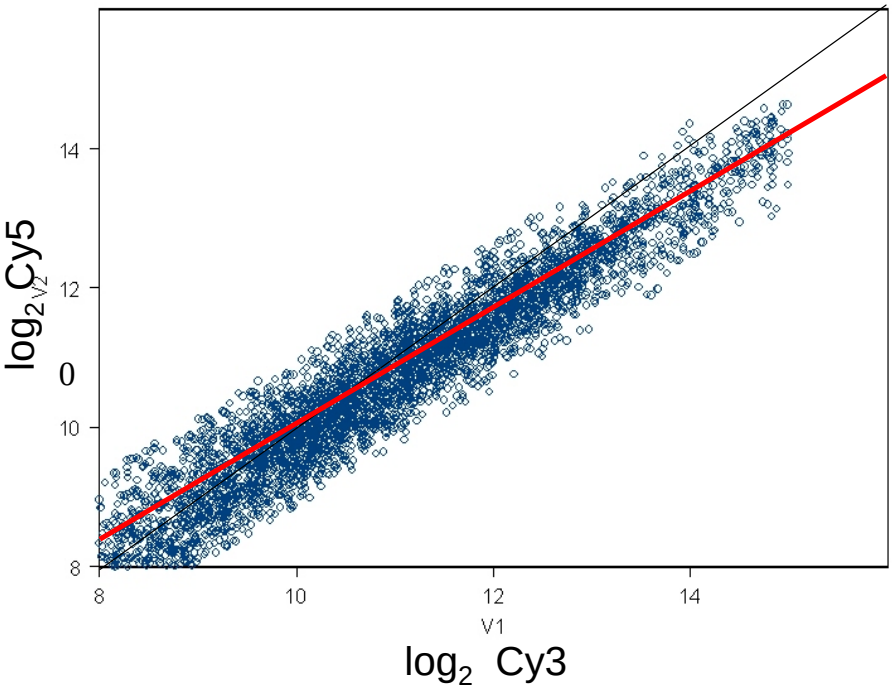
Ch. 1 vs. 2 – Outlier Removal



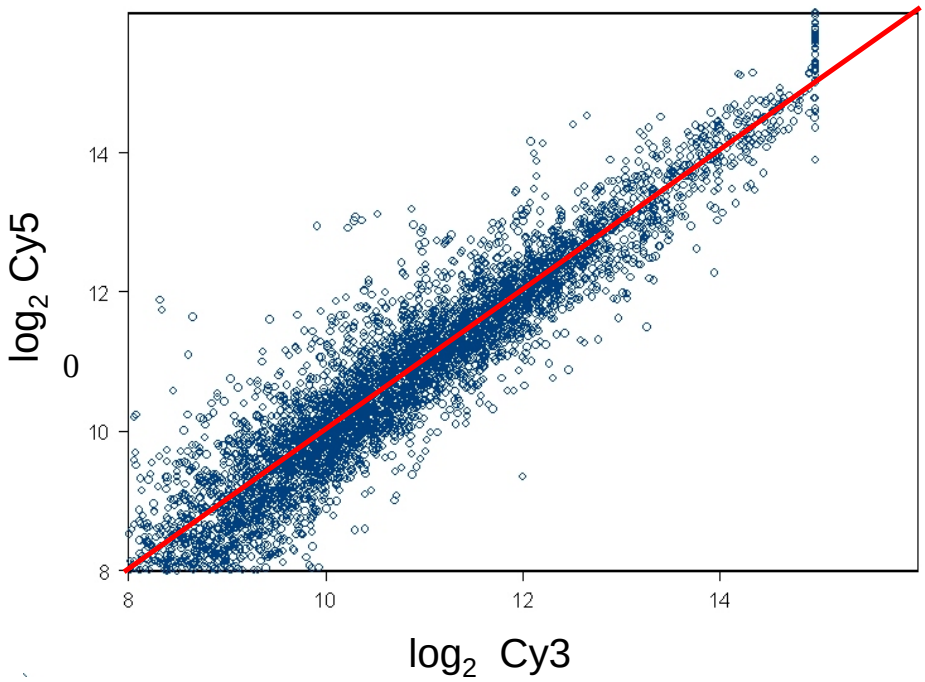
Ch. 1 vs. 2 – Outlier Removal



Ch. 1 vs. 2 – Slope Adjust



Ch. 1 vs. 2 – Normalized

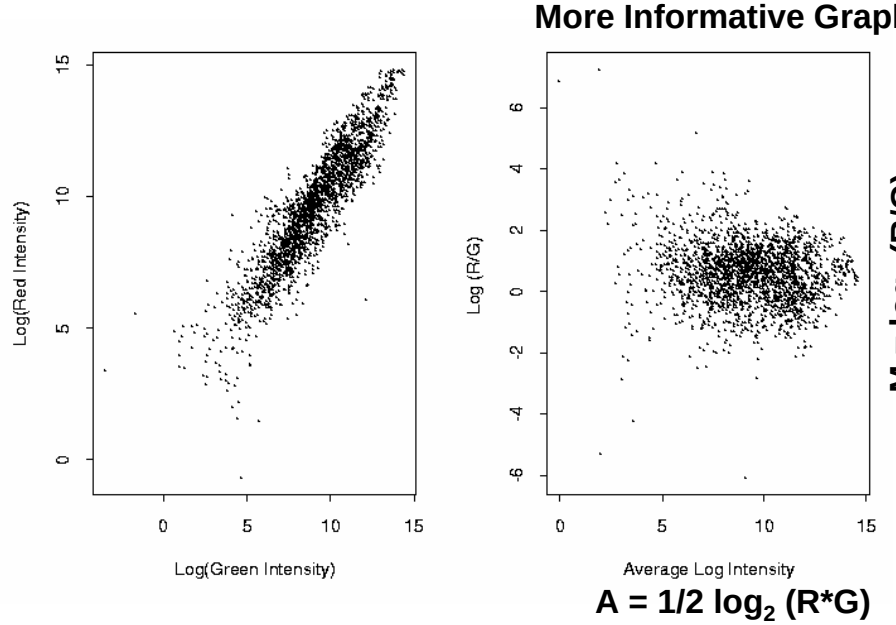


Normalization to Loess Curve

- **A curvilinear form of normalization**
- **For each spot, plot ratio vs. mean (ch1,ch2) signal in log scale (A vs. M)**
- **Use statistical programs (e.g. S-plus, SAS, or R) to fit a loess curve (local regression) through the data**
- **Offset from this curve is the normalized expression ratio**

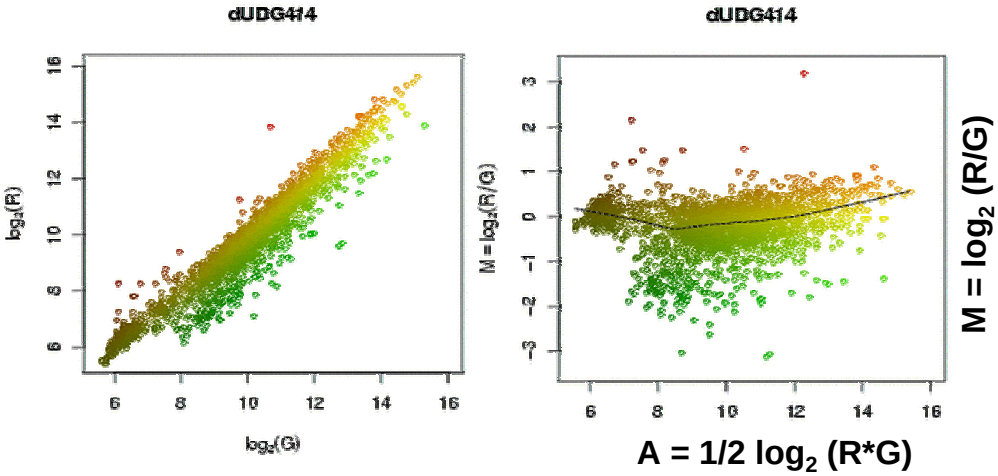
The A versus M Plot

More Informative Graph



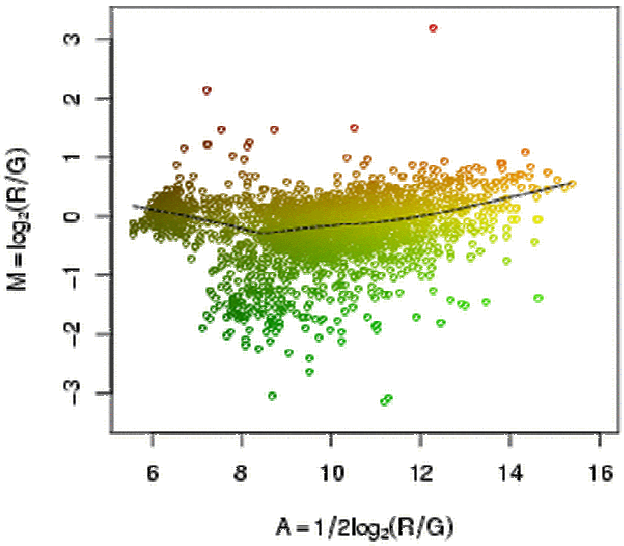
A vs. M Plot

More Informative Graph

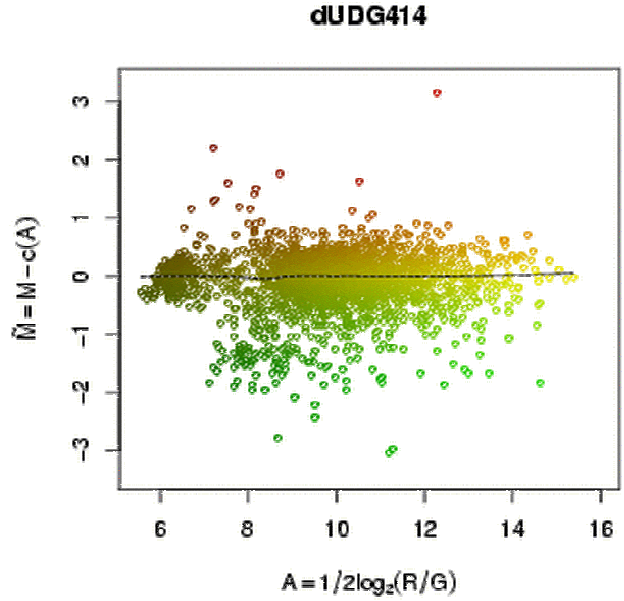


Prior To Normalization

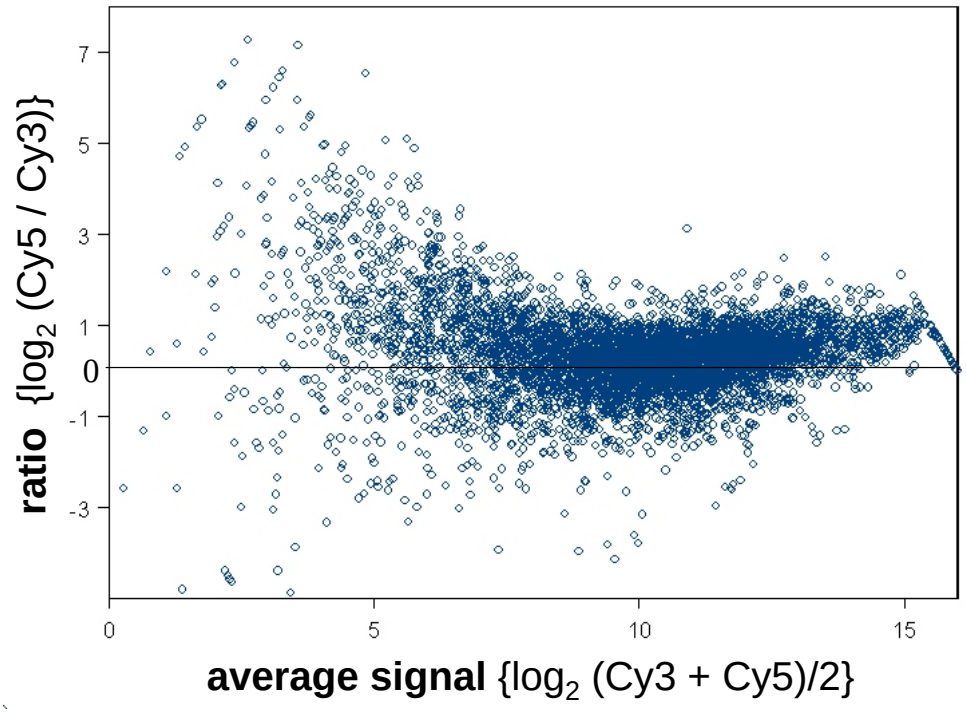
Non-normalized data $\{(M,A)\}_{n=1..5184}$:
$M = \log_2(R/G)$



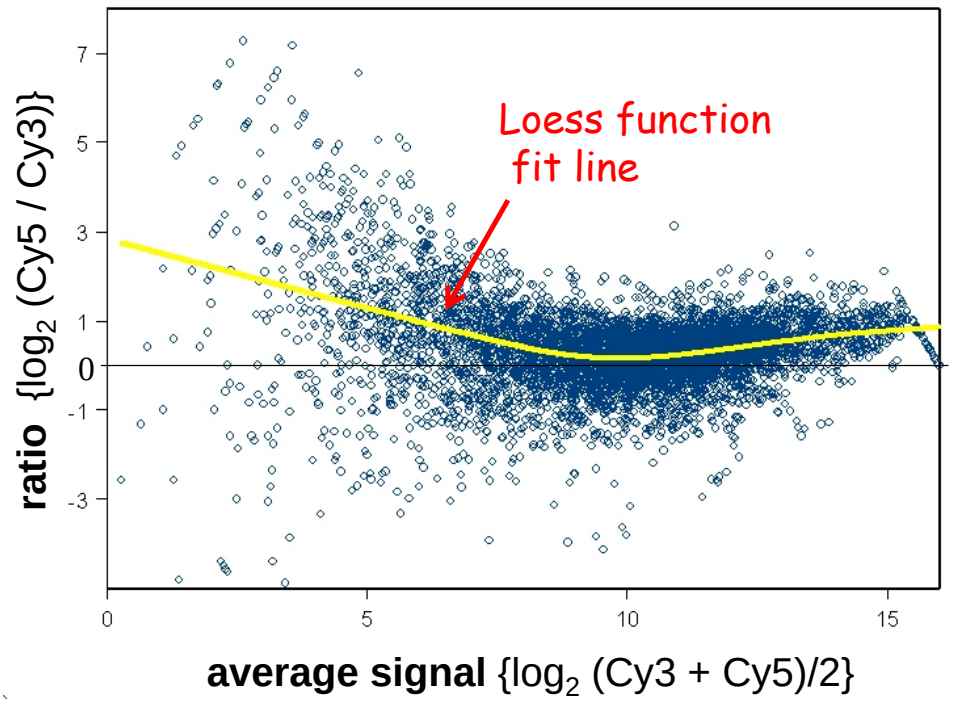
Global (Loess) Normalization



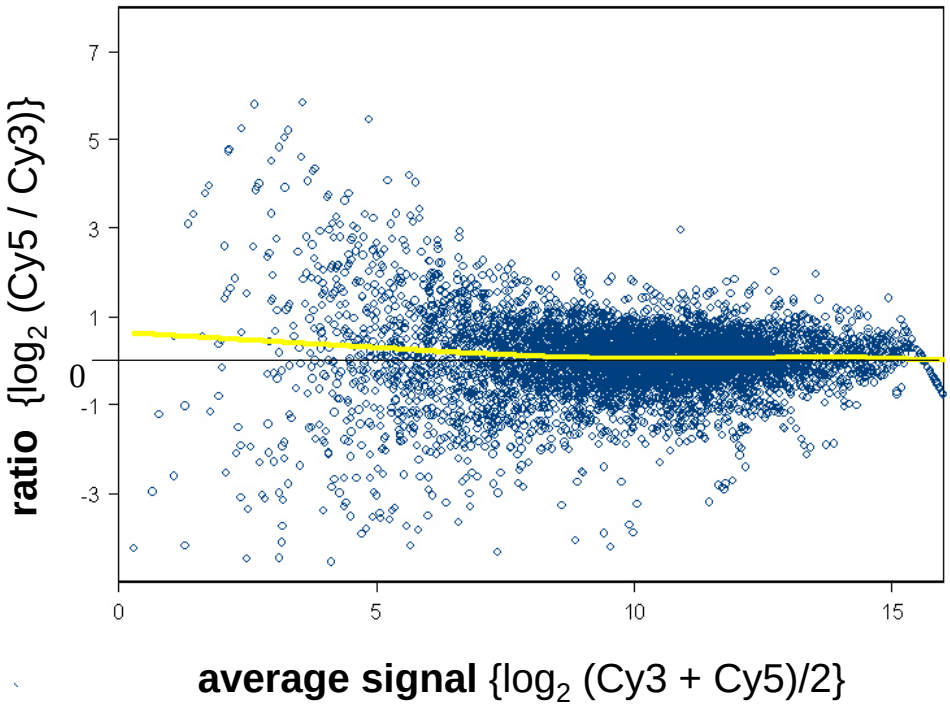
A vs. M Plot



Loess Function



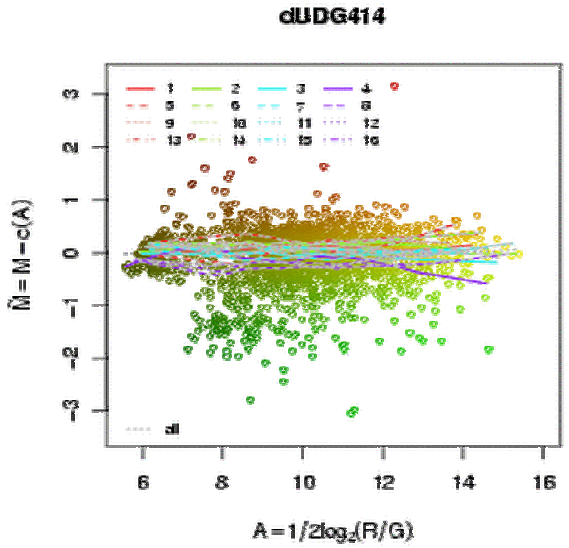
Data After Normalization

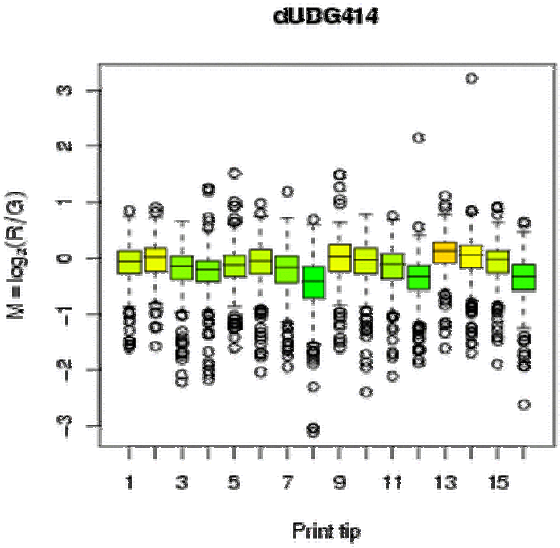


Print-tip Normalization

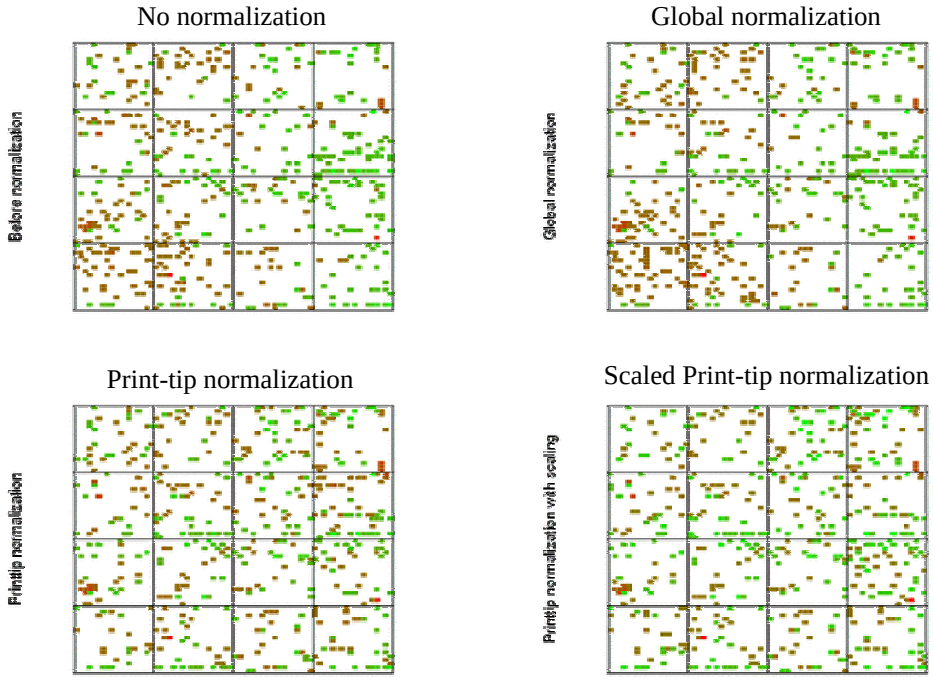
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Print-tip layout





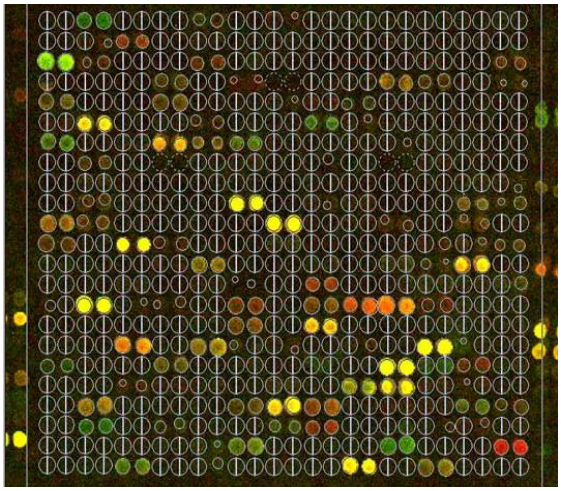
Spatial Normalization



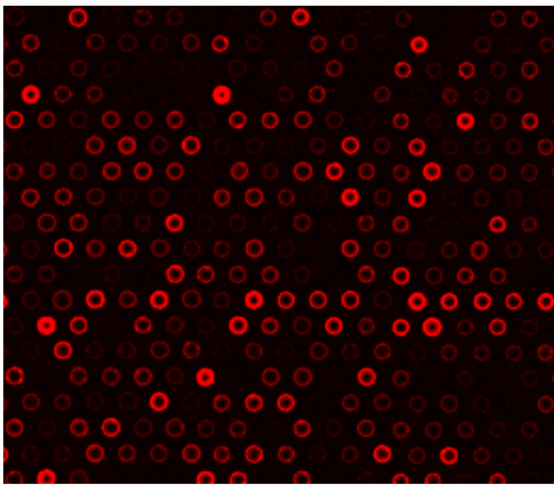
Quality Measurements

- **Array**
  - ***Correlation between spot intensities***
  - **Percentage of spots with no signals**
  - **Distribution of spot signal area**
  - **Inter-array consistency**
- **Spot**
  - **Signal / Noise ratio**
  - **Variation in pixel intensities**
  - **ID of “bad spots” (spots with no signal)**

Quality Assessment



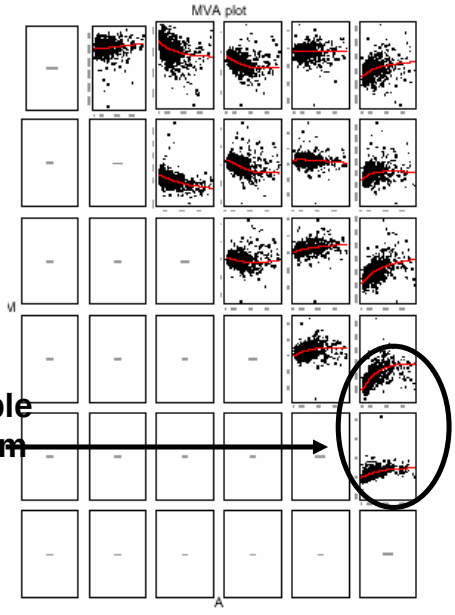
OK quality



High quality

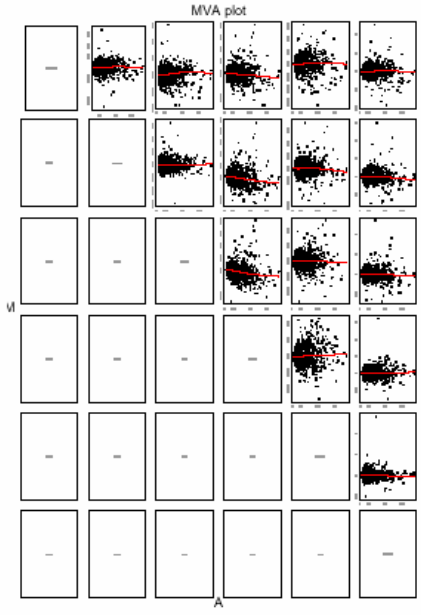
Inter-Array Consistency

Pre-normalized

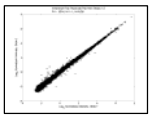
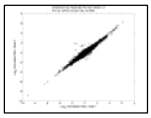
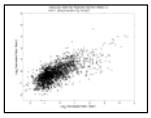
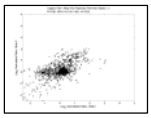
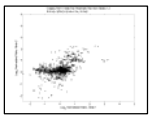


Possible problem

Normalized



Quality Assessment

High Quality Array	1) R=1	95%CI=(1-1)	N=8258	
	2) R=0.99	95%CI=(0.99-1)	N=8332	
	3) R=0.99	95%CI=(0.99-0.99)	N=8290	
High Quality Array	1) R=0.98	95%CI=(0.98-0.98)	N=7694	
	2) R=0.97	95%CI=(0.97-0.98)	N=7873	
	3) R=0.97	95%CI=(0.97-0.97)	N=7694	
Good Quality Array	1) R=0.7	95%CI=(0.68-0.72)	N=2027	
	2) R=0.65	95%CI=(0.62-0.67)	N=2818	
	3) R=0.61	95%CI=(0.59-0.64)	N=2001	
Poor Quality Array	1) R=0.66	95%CI=(0.62-0.69)	N=1028	
	2) R=0.86	95%CI=(0.85-0.87)	N=1925	
	3) R=0.64	95%CI=(0.61-0.68)	N=1040	
Poor Quality Array	1) R=0.49	95%CI=(0.44-0.54)	N=942	
	2) R=0.81	95%CI=(0.8-0.83)	N=1700	
	3) R=0.57	95%CI=(0.52-0.61)	N=973	

Microarrays in Practice

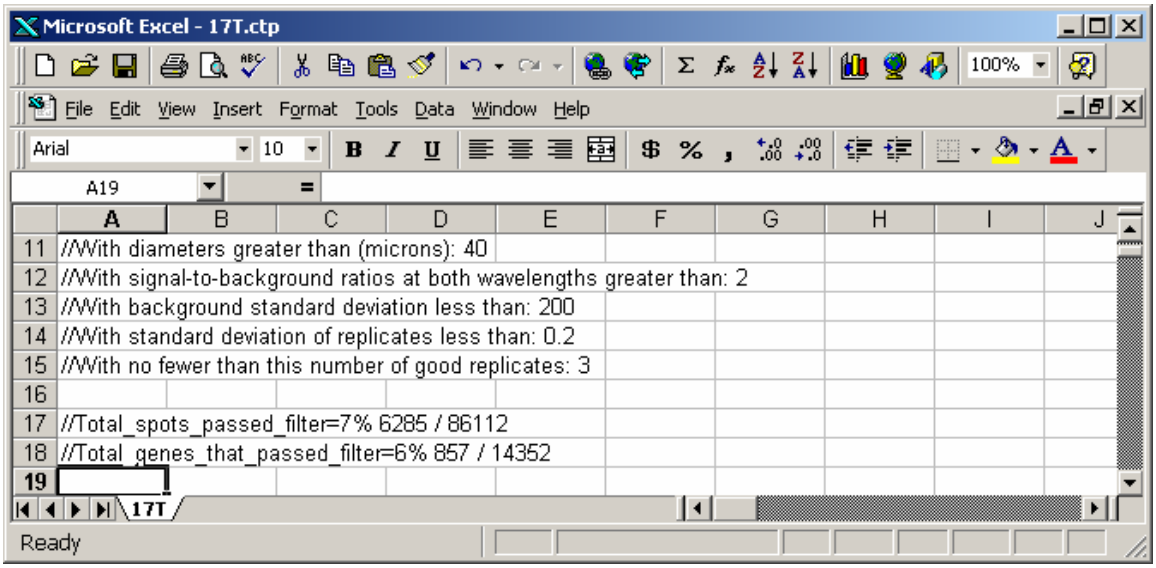
- **14,352 spots from a collection of synthetic DNA 70mers from Operon**
- **Each oligo is spotted 2x side-by-side**
- **Each slide is prepared and scanned in triplicate (6 spots per gene)**
- **Each slide image is hand inspected & QCd**
- **Scanning system is calibrated by hand**

Microarrays in Practice

- **Gene ID, Gene Name**
- **Foreground** median pixel intensity (red, green channels)
- **Background** median pixel intensity (red, green channels)
- **Ratio of medians** of background-corrected intensities
- **Normalized Ratio** of medians of background-corrected intensities
- **Percent Saturation** (red, green channels)
- **Diameter** of spot found
- **Background standard deviation** (red, green)
- **Robust average** over replicated experimental spots.
- **Robust standard deviation** over replicated experimental spots.

Microarray Data Cleansing

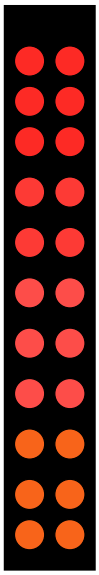
Before entering the database, microarray data is filtered:



Data Filtering/Cleansing

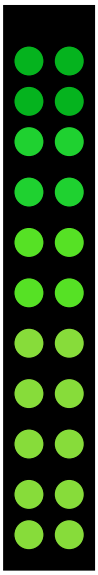
- **Purpose is to set aside spots with low, high, or negative intensity values**
  - **low intensity values are associated with high variance (a sensitivity issue)**
  - **very high signals may be saturated**
  - **negative signals can be produced by background subtraction**

Final Result



Highly Exp	
Trx	16.8
Enh1	13.2
Hin2	11.8
P53	8.4
Calm	7.3
Ned3	5.6
P21	5.5
Antp	5.4
Gad2	5.2
Gad3	5.1
Erp3	5.0

Reduced Exp	
GPD	0.11
Shn2	0.13
Alp4	0.22
OncB	0.23
Nrd1	0.25
LamR	0.26
SetH	0.30
LinK	0.32
Mrd2	0.32
Mrd3	0.33
TshR	0.34



Key Steps in Microarray Analysis

- Quality Control (checking microarrays for errors or problems)
- Image Processing
  - Gridding
  - Segmentation (peak picking)
  - Data Extraction (intensity, QC)
- Data Analysis and Data Mining

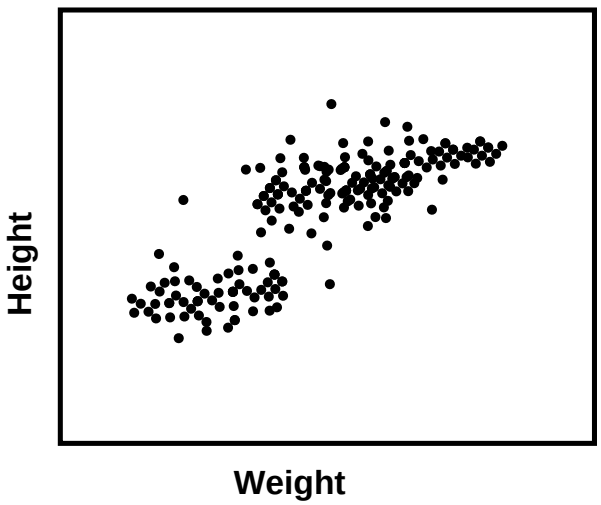
Identifying Patterns of Gene Expression

- Key Goal: identify co-regulated groups of genes
- This leads to:
  - inferences about physiological responses
  - generalizations about large data sets
  - identification of regulatory cascades
  - assignment of possible function to uncharacterized genes
  - identification of shared regulatory motifs

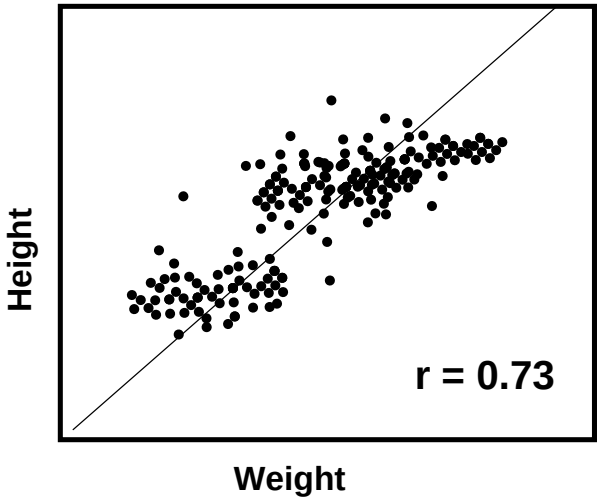
Analysis Methods

- Hierarchical clustering
- K-means clustering
- Self organizing feature maps
- Support vector machines
- Neural networks
- Principle Component Analysis
- *Data Mining & Data Enrichment*

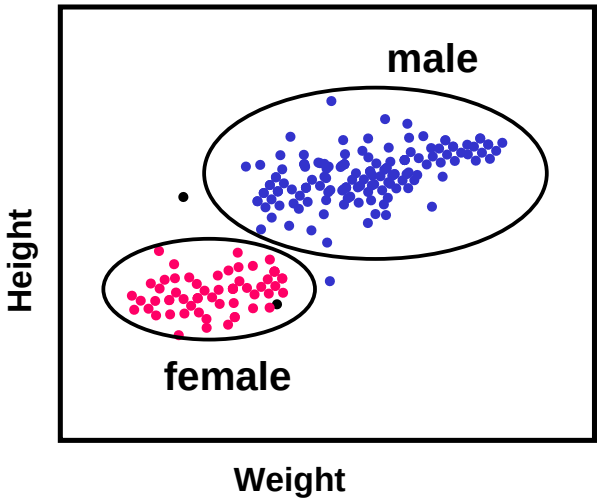
Detecting Clusters



Is it Right to Calculate a Correlation Coefficient?



Or is There More to This?



Clustering Applications in Bioinformatics

- Microarray or GeneChip Analysis
- 2D Gel or ProteinChip Analysis
- Protein Interaction Analysis
- Phylogenetic and Evolutionary Analysis
- Structural Classification of Proteins
- Protein Sequence Families

Clustering

- Definition - a process by which objects that are logically similar in characteristics are grouped together.
- Clustering is different than Classification
- In classification the objects are assigned to pre-defined classes, in clustering the classes are yet to be defined
- Clustering helps in classification

Clustering Requires...

- A method to measure similarity (a similarity matrix) or dissimilarity (a dissimilarity coefficient) between objects
- A threshold value with which to decide whether an object belongs with a cluster
- A way of measuring the “distance” between two clusters
- A cluster seed (an object to begin the clustering process)

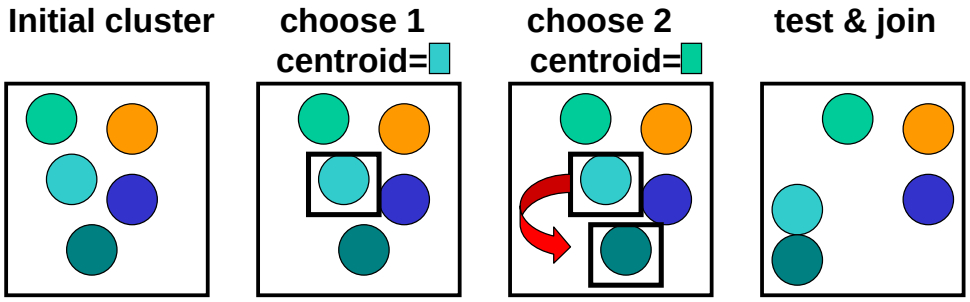
Clustering Algorithms

- K-means or Partitioning Methods - divides a set of N objects into M clusters -- with or without overlap
- Hierarchical Methods - produces a set of nested clusters in which each pair of objects is progressively nested into a larger cluster until only one cluster remains
- Self-Organizing Feature Maps - produces a cluster set through iterative “training”

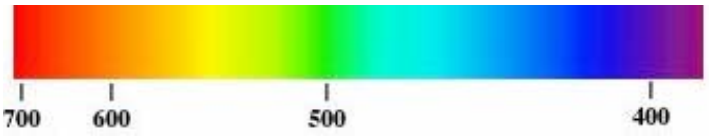
K-means or Partitioning Methods

- *Make the first object the centroid for the first cluster*
- *For the next object calculate the similarity to each existing centroid*
- *If the similarity is greater than a threshold add the object to the existing cluster and redetermine the centroid, else use the object to start new cluster*
- *Return to step 2 and repeat until done*

K-means or Partitioning Methods



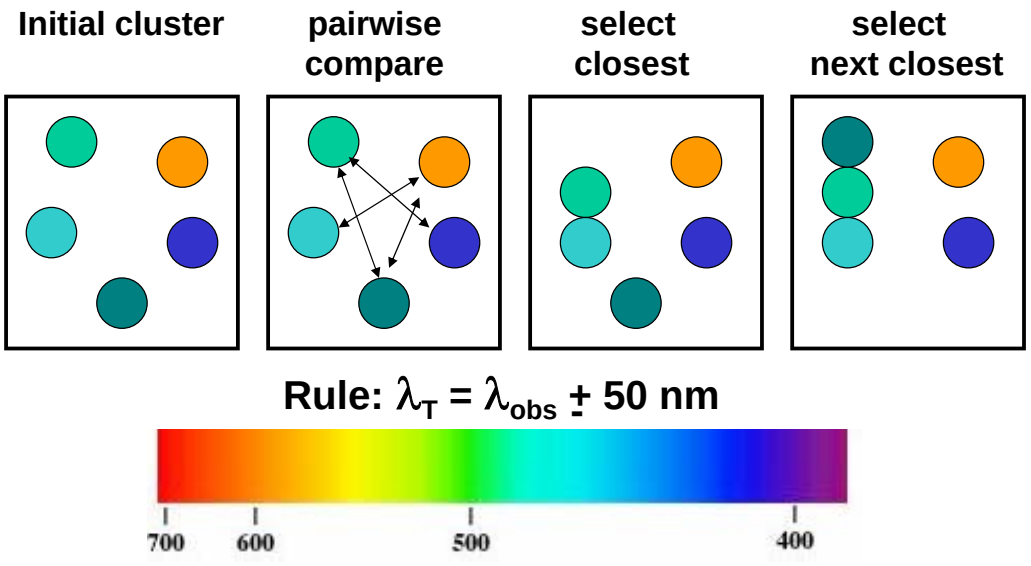
Rule: $\lambda_T = \lambda_{centroid} \pm 50$ nm



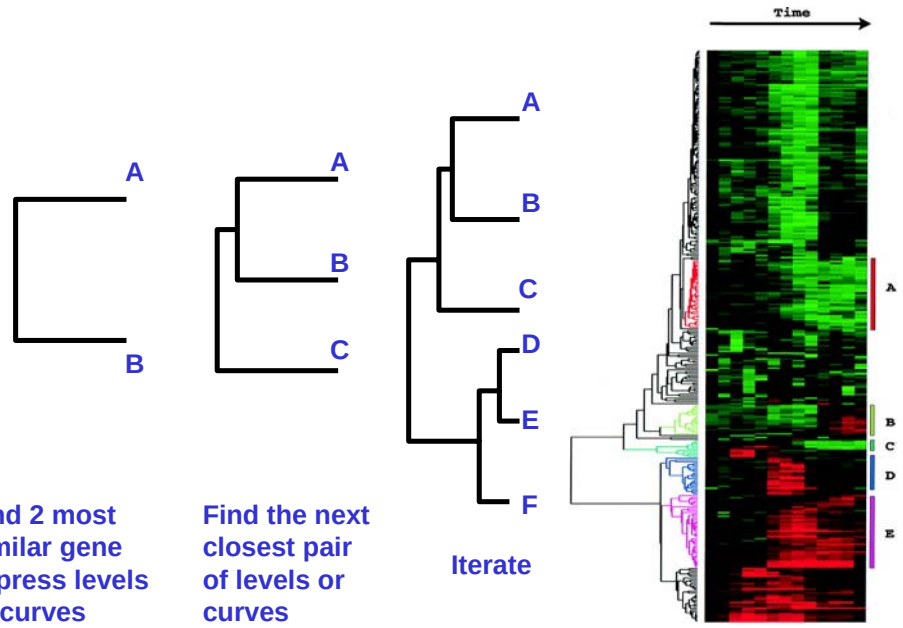
Hierarchical Clustering

- ***Find the two closest objects and merge them into a cluster***
- ***Find and merge the next two closest objects (or an object and a cluster, or two clusters) using some similarity measure and a predefined threshold***
- ***If more than one cluster remains return to step 2 until finished***

Hierarchical Clustering



Hierarchical Clustering

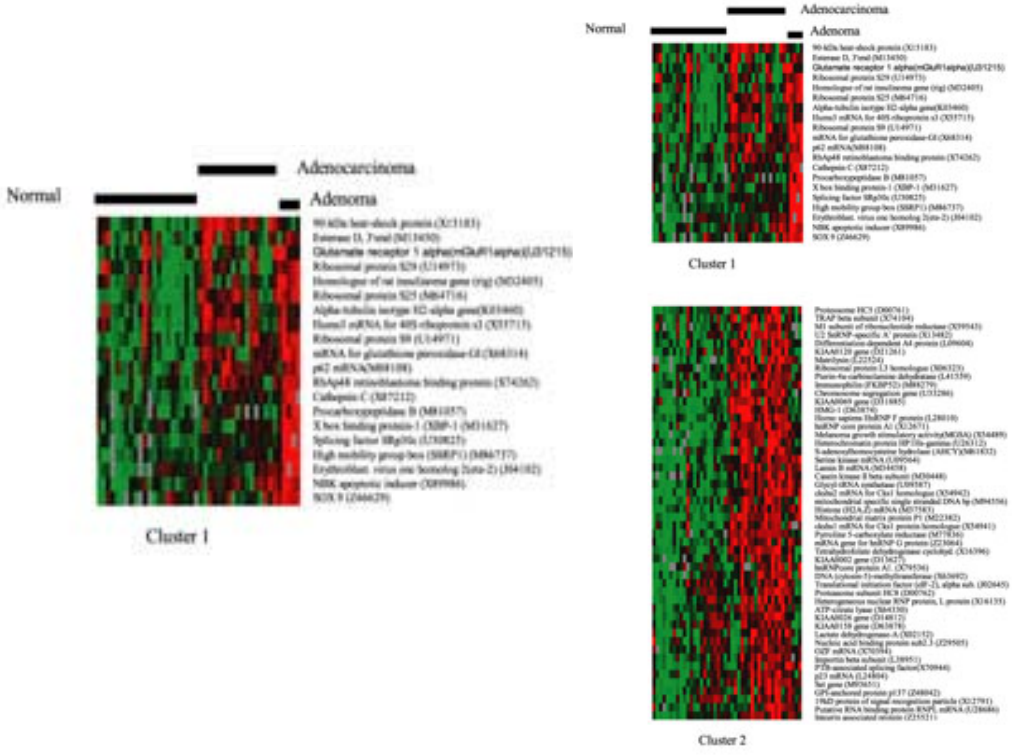


Find 2 most similar gene express levels or curves

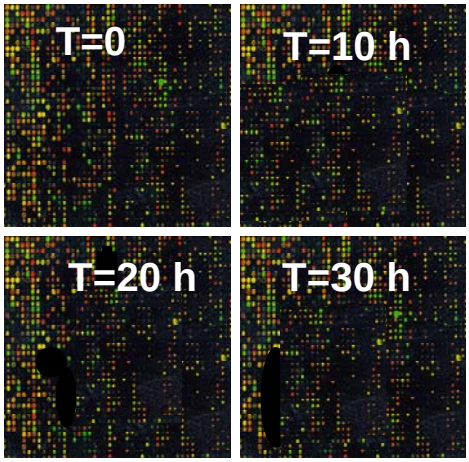
Find the next closest pair of levels or curves

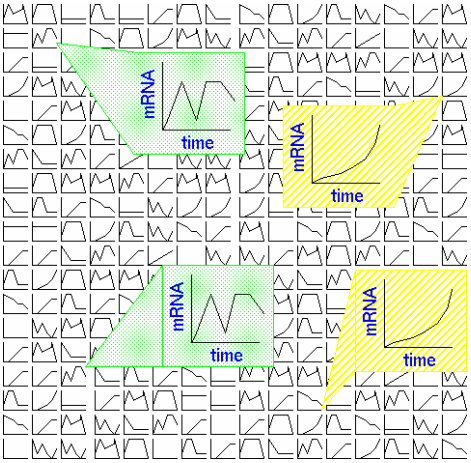
Iterate

Results

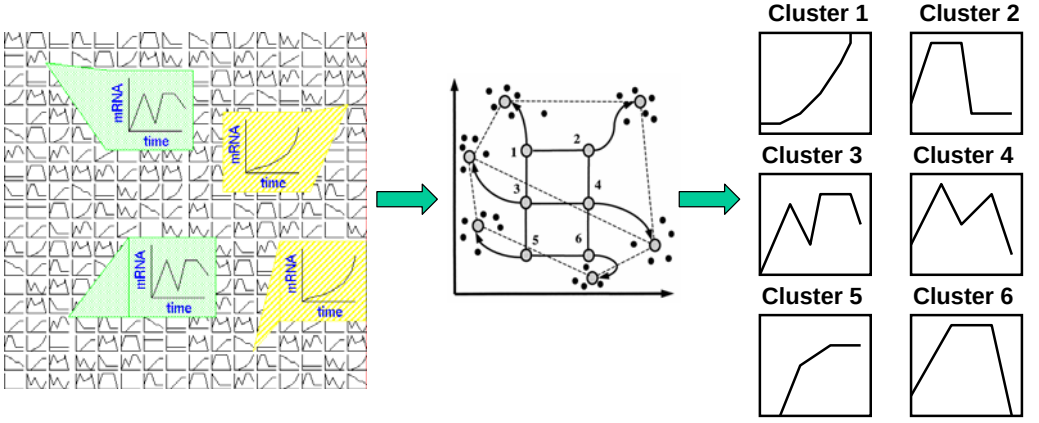


Self-Organizing Feature Maps





Self-Organizing Feature Maps



Plot Chip Data

Compute Feature Map with 6 nodes

Examine Clusters for Biological Meaning

Data Mining



The Genomic Abyss

Data Mining

- Enriches the data content of previously collected experimental data -- adds to data dimemsionality
- Adds “prior knowledge” to raw or hard-to-interpret data
- Key to finding new relations or hidden connections
- Key to enhancing quality of predictions

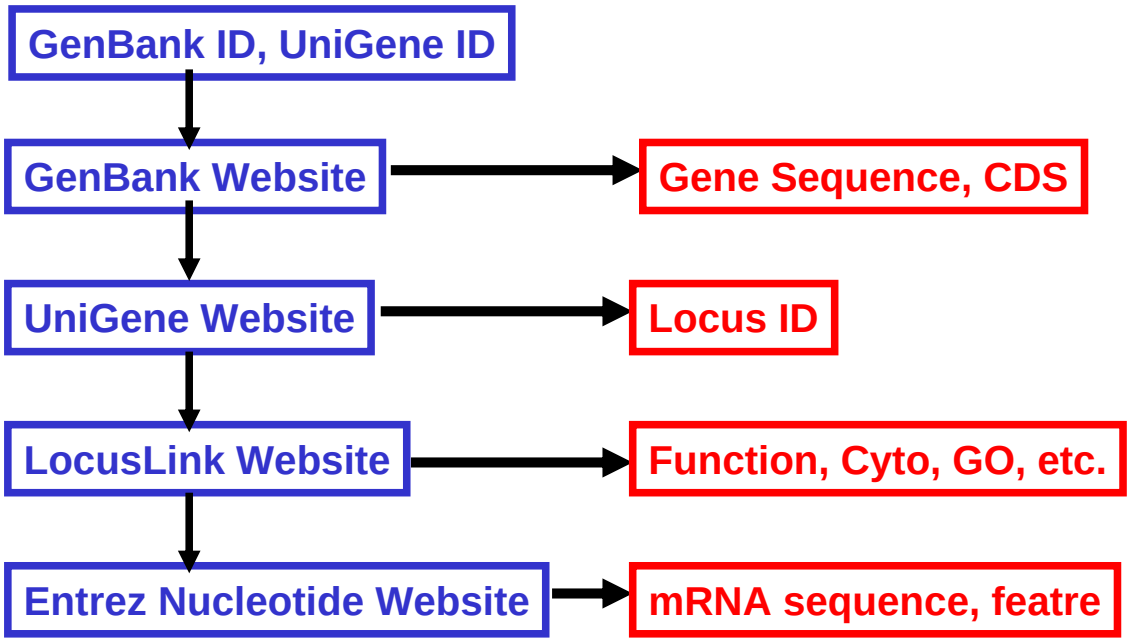
Data Mining Challenges

- Terabytes of data “out there, somewhere”
- Data is constantly being updated,corrected
- New genes being added, old ones deleted
- Data sources are highly distributed
- Data sources are of variable quality (errors)
- Data is in various forms (image, numeric, natural language text, temporal, spatial)
- *“mining a moving target”*

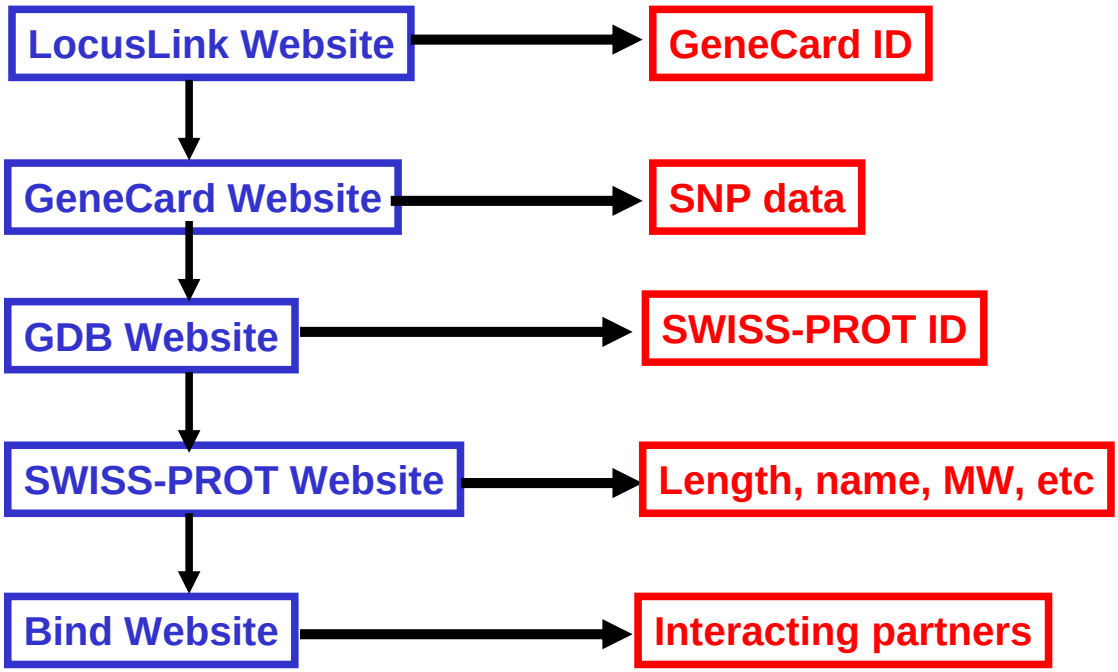
Data Mining Solutions

- Use web-based agents (web-bots) to collect, compare and update data
- Identify key (reliable) databases
- Identify key “data fields” of interest or potential importance
- Generate a web-based, web-accessible ArrayCard to link genetic, metabolic, proteomic data to geneID

Mining Microarray Data



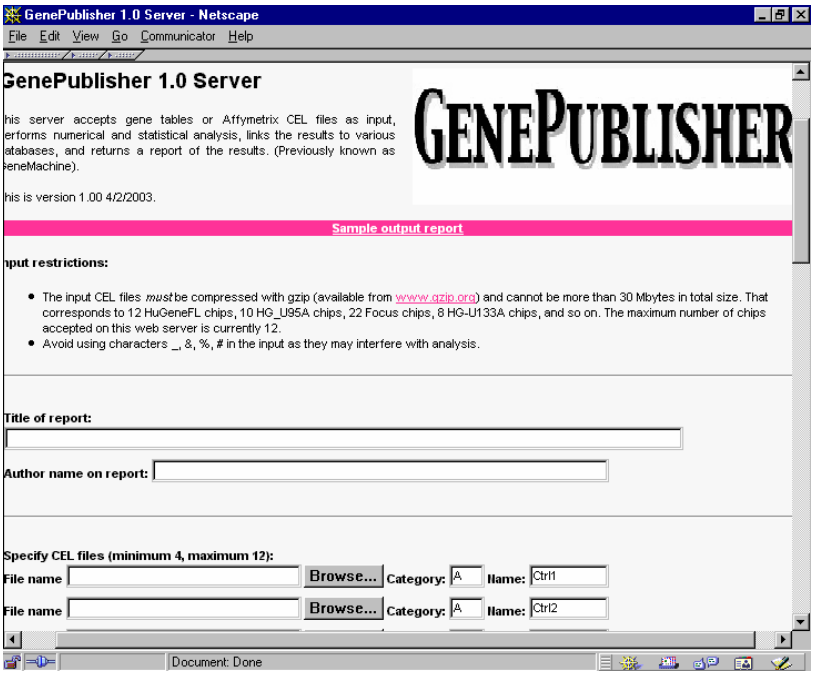
Mining Microarray Data



Putting it All Together

- Use clustering methods to identify unusual patterns or groups that associate with a disease state or conditions
- Use data mining to interpret the results in terms of existing biological or physiological knowledge
- Or selectively combine mined data with expression data to conduct further clustering

GenePublisher



http://www.cbs.dtu.dk/services/GeneMachine/

Microarray Protocols & Links

- http://www.microarrays.org/protocols.html
- http://brownlab.stanford.edu/protocols.html
- http://oz.Berkeley.EDU/users/terry/zarray/Html/
- http://ihome.cuhk.edu.hk/~b400559/arraysoft.html

  http://genome-www5.Stanford.EDU/MicroArray//SMD/resinfo.html
- http://www.bio.davidson.edu/courses/genomics/chip/chip.html