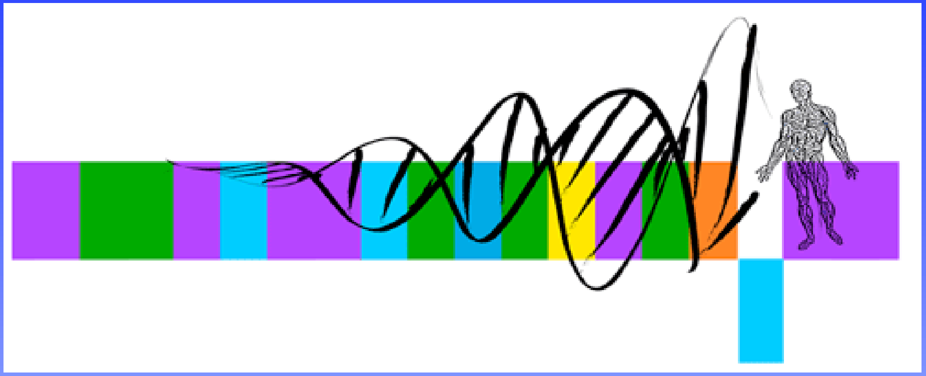


Bioinformatics for Biologists

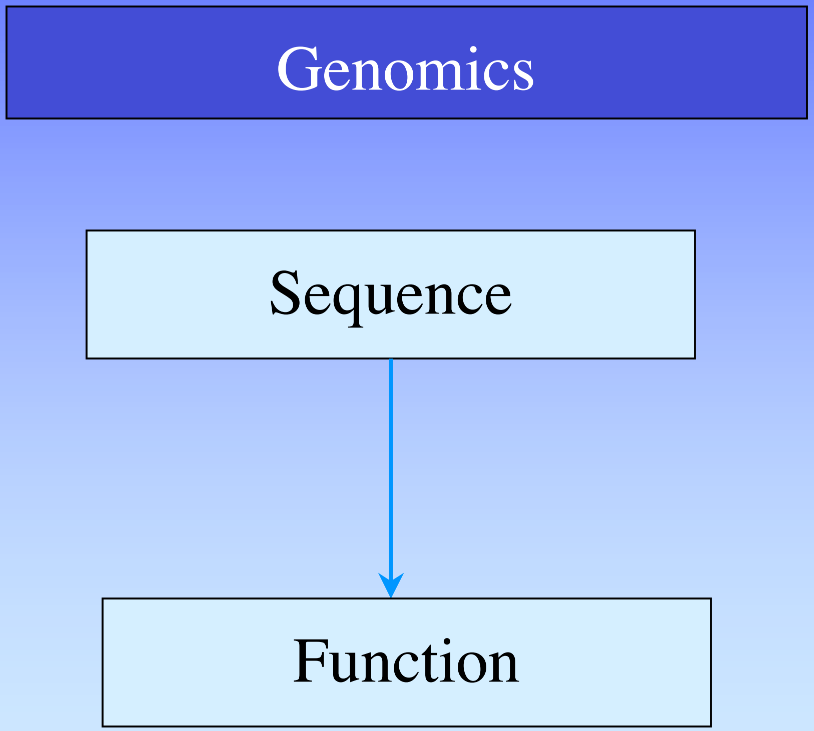
Functional Genomics: Microarray Data Analysis

Fran Lewitter, Ph.D.
Head, Biocomputing
Whitehead Institute

Outline

- Introduction
- Working with microarray data
  - Normalization
  - Analysis
    - Distance metrics
    - Clustering methods

Research Trends



Genomics

Sequence

Function

- How are genes regulated?
- How do genes interact?
- What are the functional roles of different genes?
- How does expression level of a gene differ in different tissues?

Transcriptional Profiling

(Adapted from Quackenbush 2001)

- Study of patterns of gene expression across many experiments that survey a wide array of cellular responses, phenotypes and conditions

- Simple analysis - what's up/down regulated?

- More interesting - identify patterns of expression for insight into function, etc.

Microarray Data

Collect data on *n* DNA samples (e.g. rows, genes, promotors, exons, etc.) for *p* mRNA samples of tissues or experimental conditions (eg. columns, time course, pathogen exposure, mating type, etc)

$$\text{Matrix } (n\ x\ p) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Multivariate Analysis

Concerned with datasets with more than one response variable for each observational or experimental unit (e.g. matrix X with *n* rows (genes) and *p* columns (tissue types))

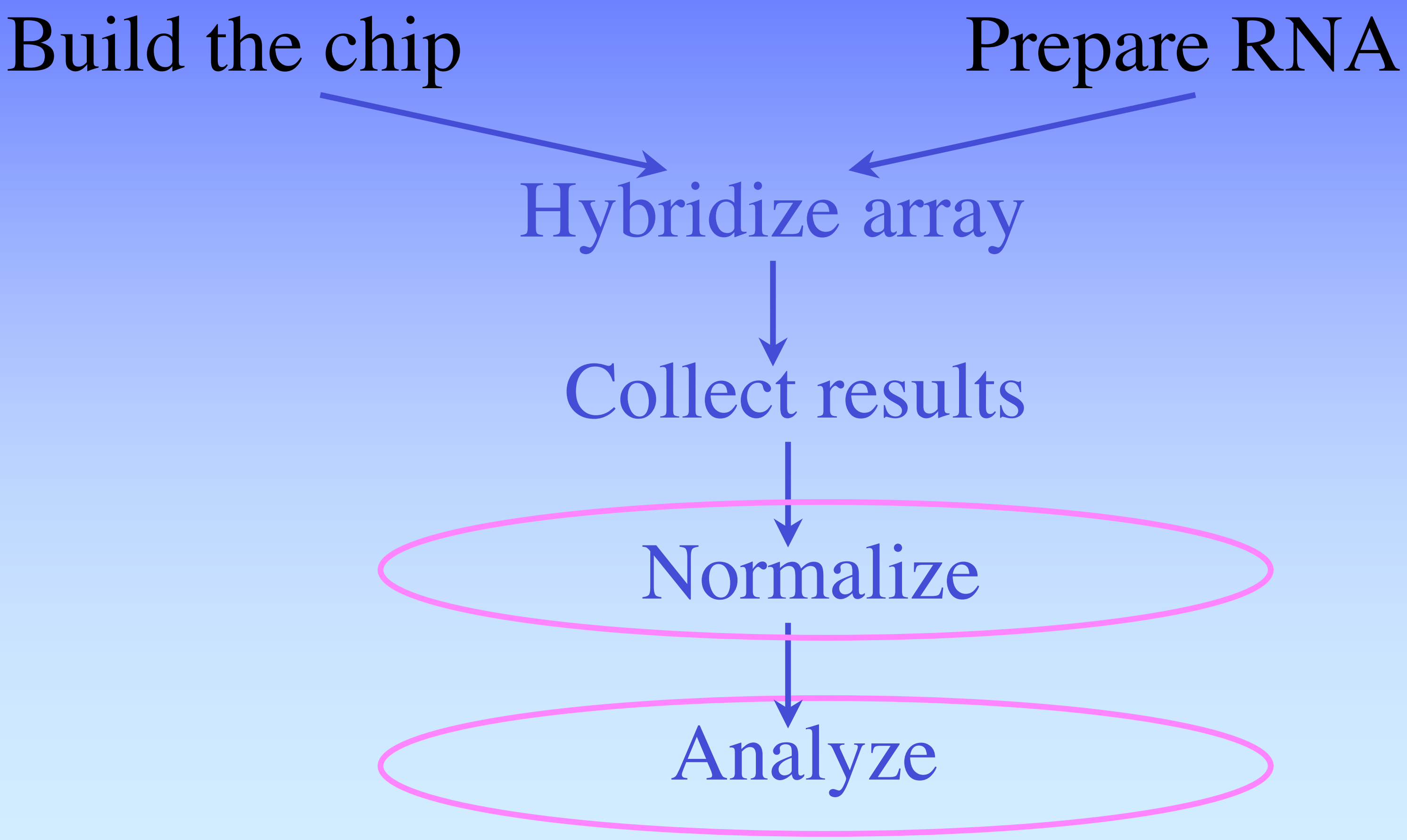
- Hierarchical (phylogenetic trees) vs non-hierarchical (k-means)
- Divisive vs agglomerative
- Supervised vs unsupervised
  - Divide cases into groups vs discover structure of data

Multivariate Methods

- Cluster analysis - discover groupings among cases of X
  - Hierarchical produces dendograms
  - K-means - choose a prespecified number of clusters
  - Self Organizing Maps

- Principal component analysis (PCA)
  - Linear method, unsupervised, seeks linear combinations of the columns of X with maximal (or minimal) variance (graphical)

DNA Microarrays



Build the chip

Prepare RNA

Hybridize array

Collect results

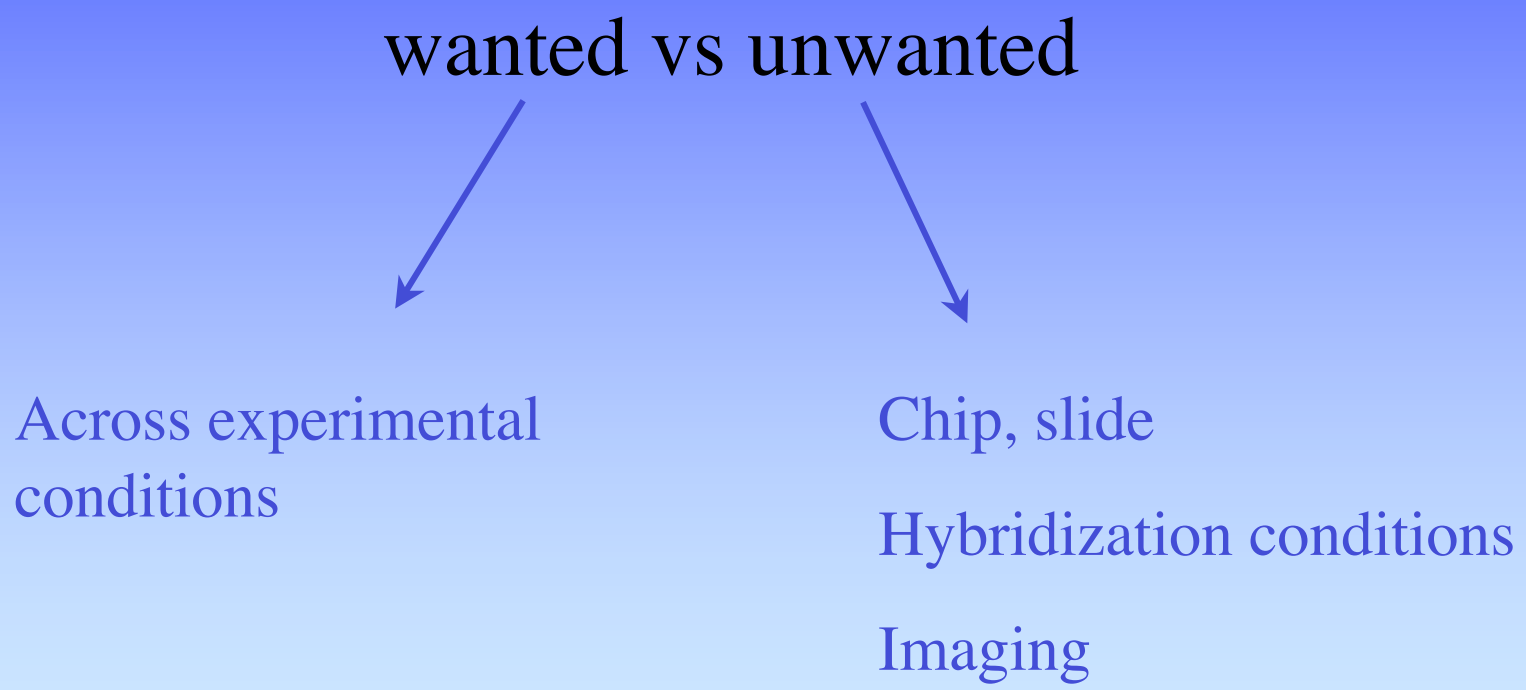
Normalize

Analyze

Data Normalization

- Correct for systematic bias in data
  - Avoid it, recognize it, correct it, discard outliers
- First step for comparing data from one array to another

Sources of variation



wanted vs unwanted

Across experimental conditions

Chip, slide

Hybridization conditions

Imaging

Normalization Approaches

Compensate for experimental variability

- Housekeeping genes
- Spiked in controls
- Total intensity normalization
- LOWESS correction

Expression Ratios

- Let R = a query sample
- Let G = a reference sample
- Then the ratio, $T_i = R_i/G_i$
- Need to transform these to $\log_2$
- Examples: T = 2/1 = 2; T=1/2 = .5
- Examples: $\log_2(2) = 1$; $\log_2(.5) = -1$

Total Intensity Normalization

(Adapted from Quackenbush 2002)

Assumptions: (1) start with equal amounts of RNA for the two samples; (2) arrayed elements represent random sample of genes in the organism

a.

$$N_{total} = \frac{\sum_{i=1}^{Narray} R_i}{\sum_{i=1}^{Narray} G_i}$$

b. Rescale intensities:

$$G'_i = N_{total} G_i \text{ and } R'_i = R_i$$

c.

$$T'_i = \frac{R'_i}{G'_i} = \frac{1}{N_{total}} \frac{R_i}{G_i}$$

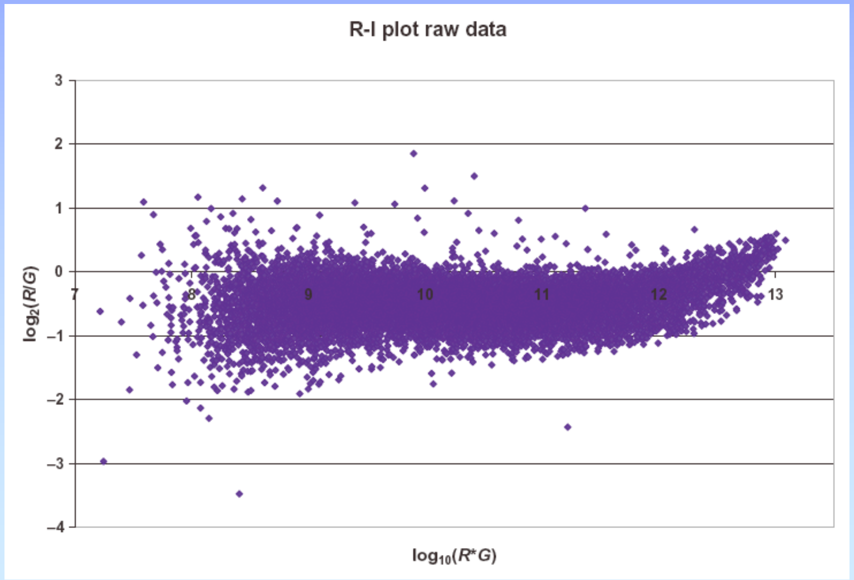
d.

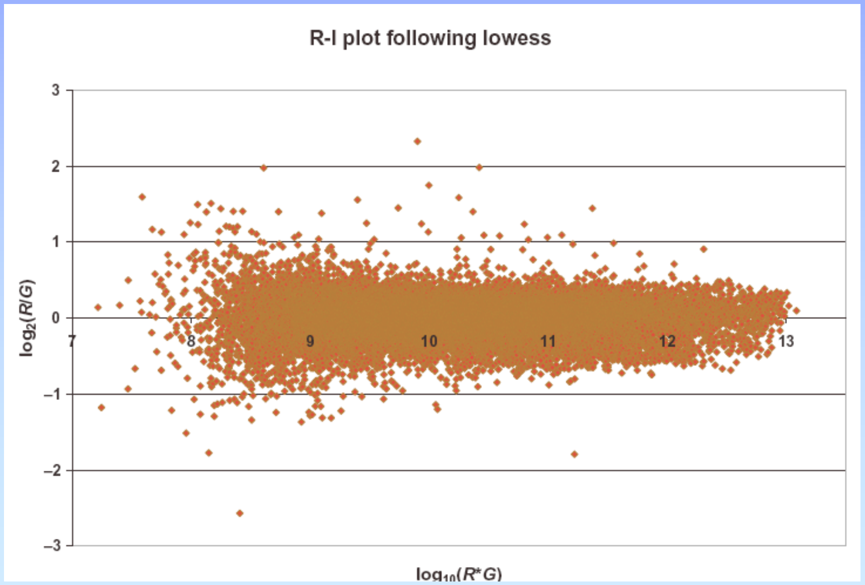
$$\log_2(T'_i) = \log_2(T_i) - \log_2(N_{total})$$

LOWESS - The R-I Plot

(Adapted from Quackenbush 2002)

- Data exhibit an intensity-dependent structure
- Uncertainty in intensity and ratio measurements is greater at lower intensities





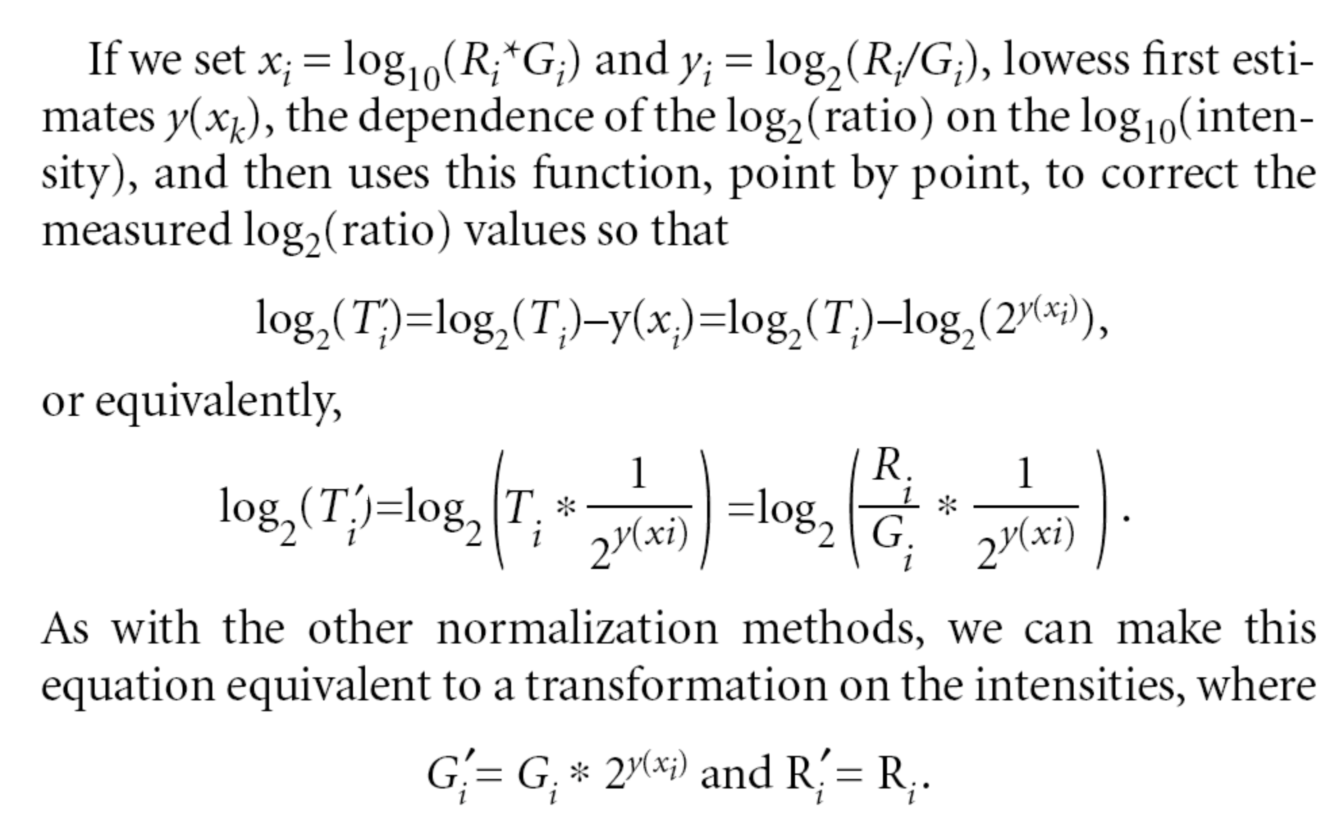
LOWESS - The R-I Plot

(Adapted from Quackenbush 2002)

- Plot $\log_2$(R/G) ratio as a function of $\log_{10}$(R*G) product intensity
- Shows intensity specific artifacts in the measurements of ratios
- Correct using a local weighted linear regression

LOWESS Normalization

(From Quackenbush 2002)



If we set $x_i = \log_{10}(R_i{}^*G_i)$ and $y_i = \log_2(R_i/G_i)$, lowess first estimates $y(x_k)$, the dependence of the $\log_2$(ratio) on the $\log_{10}$(intensity), and then uses this function, point by point, to correct the measured $\log_2$(ratio) values so that

$$\log_2(T'_i)=\log_2(T_i)-y(x_i)=\log_2(T_i)-\log_2(2^{y(x_i)}),$$

or equivalently,

$$\log_2(T'_i)=\log_2\left(T_i * \frac{1}{2^{y(xi)}}\right)=\log_2\left(\frac{R_i}{G_i} * \frac{1}{2^{y(xi)}}\right).$$

As with the other normalization methods, we can make this equation equivalent to a transformation on the intensities, where

$$G'_i = G_i * 2^{y(x_i)} \text{ and } \mathrm{R}'_i = \mathrm{R}_i.$$

After normalization

(Adapted from Quackenbush 2001)

- Data reported as an "expression ratio" or as a logarithm of the expression ratio

- Expression ratio is the normalized value of the expression level for a particular gene in the query sample divided by its normalized value for the control

- Use log of expression ratio for easier comparisons

Distance Metrics

(Adapted from Quackenbush 2001)

- Metric distances - d_{ij} between two vectors, *i* and *j*, must obey several rules:
  - Distance must be positive definite, $dij \geq 0$
  - Distance must be symmetric, $d_{ij} = d_{ji}$, so that the distance from *i* to *j* is the same as the distance from *j* to *i*.
  - An object is zero distance from itself, $d_{ii} = 0$.
  - When considering three objects, *i*, *j* and *k*, $d_{ik} \leq d_{ij} + d_{jk}$. This is sometimes called the 'triangle' rule.

Distance Metrics

(Adapted from Quackenbush 2001)

- The most common metric distance is Euclidean distance,which is a generalization of the familiar Pythagorean theorem. In a three-dimensional space, the Euclidean distance, d_{12}, between two points, (x_1,x_2,x_3) and (y_1,y_2,y_3) is given by:

$$d_{12} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2},$$

- where (x_1,x_2,x_3) are the usual Cartesian coordinates (x,y,z).

More on distance

(Adapted from Quackenbush 2001)

The generalization of this to higher-dimensional expression spaces is straightforward.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \, ,$$

where x_i and y_i are the measured expression values, respectively, for genes X and Y in experiment i, and the summation runs over the n experiments under analysis.

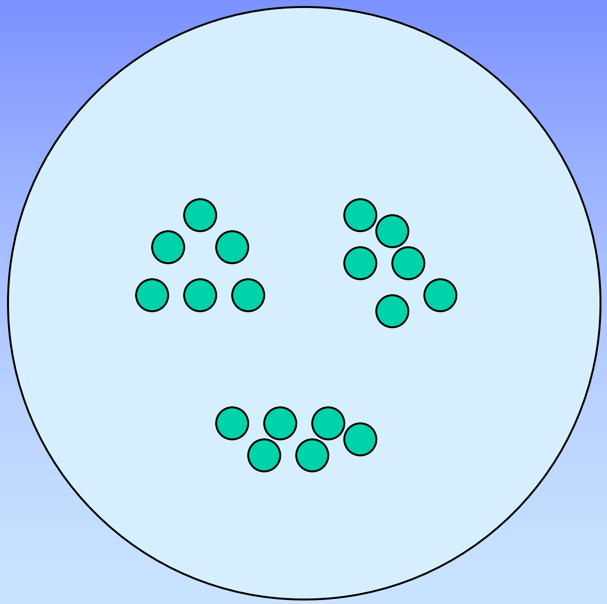
Semi-metric distances

(Adapted from Quackenbush 2001)

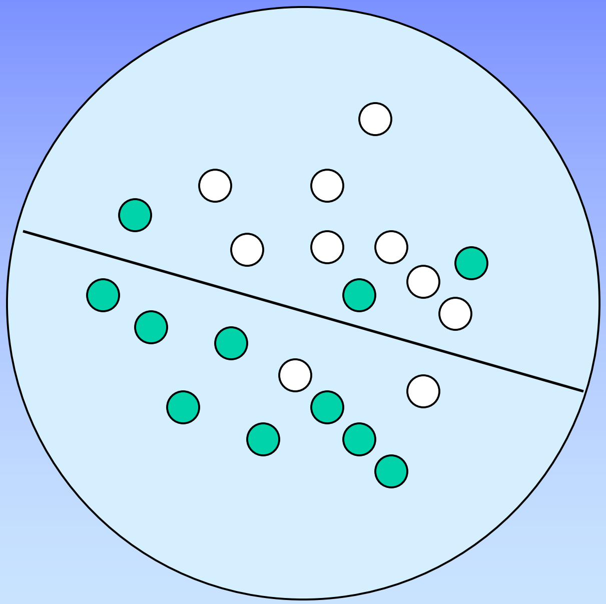
- Distance measures that obey the first three consistency rules, but fail to maintain the triangle rule are referred to as semi-metric.

- Pearson correlation coefficient is most commonly used semi-metric distance measure

Clustering vs Classification



Unsupervised



Supervised

Hierarchical methods

(Adapted from Dudoit and Gentleman, 2002)

- Produces a tree or dendogram
- Don't need to specify how many clusters
- The tree can be built in two distinct ways
  - bottom-up: agglomerative clustering
  - top-down: divisive clustering

Agglomerative methods

(Adapted from Dudoit and Gentleman, 2002)

- Start with *n* mRNA sample clusters
- At each step, merge two closest clusters using a measure of between-cluster dissimilarity reflecting shape of the clusters
- Between-cluster dissimilarity measures
  - Unweighted Pair Group Method with Arithmetic mean (UPGMA): average of pairwise dissimilarities
  - Single-link: minimum of pairwise dissimilarities
  - Complete-link: maximum of pairwise dissimilarities

Divisive methods

(Adapted from Dudoit and Gentleman, 2002)

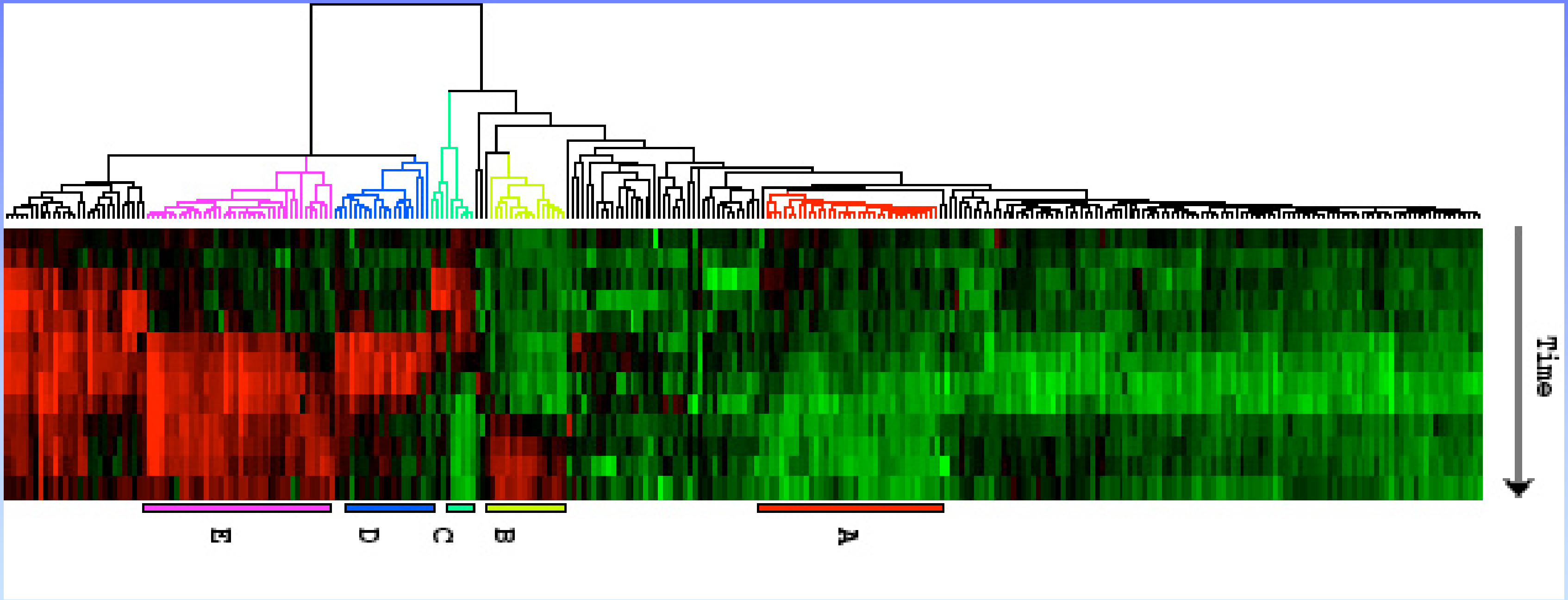
- Start with only one cluster
- At each step, split clusters into parts
- Advantages: obtain main structure of the data, i.e., focus on upper levels of dendogram
- Disadvantages: computational difficulties when considering all possible divisions into two groups

Hierarchical Clustering

(Adapted from Quackenbush 2001)

- ***Agglomerative*** - single expression profiles are joined to form groups….forming a single tree
  - Pairwise distance matrix is calculated for all genes to be clustered
  - Distance matrix is searched for the 2 most similar genes or clusters
  - Two selected clusters are merged to produce new cluster
  - Distances calculated between this new cluster and all other clusters

Dendogram

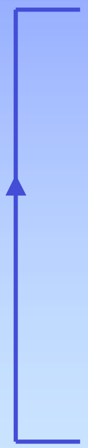


Eisen et al 1998

K-means Clustering

(Adapted from Quackenbush 2001)

- ***Divisive*** - good if you know the number (*k*) of clusters to be represented in the data
  - Initial objects randomly assigned to one of *k* clusters
  - Average expression vector calculated for each cluster & compute distance between clusters
  - Objects moved between clusters and intra- and inter-cluster distances are measured with each move
  - Expression vectors for each cluster are recalculated
  - Shuffling proceeds until moving any more objects would make clusters more variable (> intra-cluster distances and decreasing inter-cluster dissimilarity



Self Organizing Maps (SOM)

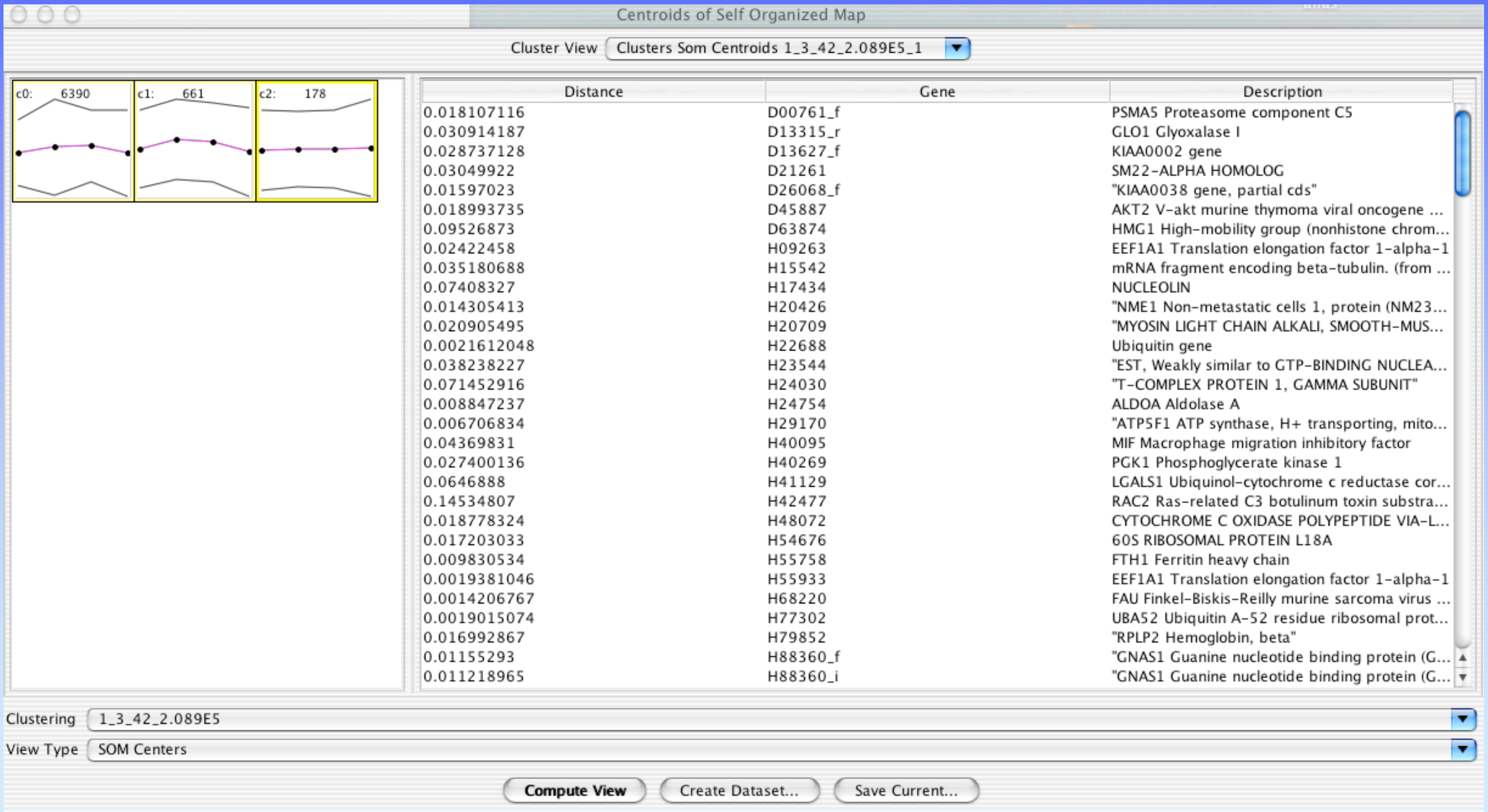
(Adapted from Quackenbush 2001)

- Neural-network based divisive clustering approach
  - Assigns genes to a series of partitions
  - User defines a geometric configuration for the partitions
  - Random vectors are generated for each partition
  - Vectors are first ‘trained’ using an iterative process until data most effectively separated

SOMs Continued

- Random vectors are constructed and assigned to each partition
- A gene is picked at random and, using a selected distance metric, the reference vector that is closest to the gene is identified
- The reference vector is then adjusted so that it is more similar to the vector of the assigned gene
- Genes are mapped to relevant partitions depending on the reference vector to which they are most similar

SOMs from GeneCluster



Cluster View: Clusters Som Centroids 1_3_42_2.089E5_1

Distance	Gene	Description
0.018107116	D00761_f	PSMA5 Proteasome component C5
0.030914187	D13315_r	GLO1 Glyoxalase I
0.028737128	D13627_f	KIAA0002 gene
0.03049922	D21261	SM22-ALPHA HOMOLOG
0.01597023	D26068_f	"KIAA0038 gene, partial cds"
0.018993735	D45887	AKT2 V-akt murine thymoma viral oncogene ...
0.09526873	D63874	HMG1 High-mobility group (nonhistone chrom...
0.02422458	H09263	EEF1A1 Translation elongation factor 1-alpha-1
0.035180688	H15542	mRNA fragment encoding beta-tubulin. (from ...
0.07408327	H17434	NUCLEOLIN
0.014305413	H20426	"NME1 Non-metastatic cells 1, protein (NM23...
0.020905495	H20709	"MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUS...
0.0021612048	H22688	Ubiquitin gene
0.038238227	H23544	"EST, Weakly similar to GTP-BINDING NUCLEA...
0.071452916	H24030	"T-COMPLEX PROTEIN 1, GAMMA SUBUNIT"
0.008847237	H24754	ALDOA Aldolase A
0.006706834	H29170	"ATP5F1 ATP synthase, H+ transporting, mito...
0.04369831	H40095	MIF Macrophage migration inhibitory factor
0.027400136	H40269	PGK1 Phosphoglycerate kinase 1
0.0646888	H41129	LGALS1 Ubiquinol-cytochrome c reductase cor...
0.14534807	H42477	RAC2 Ras-related C3 botulinum toxin substra...
0.018778324	H48072	CYTOCHROME C OXIDASE POLYPEPTIDE VIA-L...
0.017203033	H54676	60S RIBOSOMAL PROTEIN L18A
0.009830534	H55758	FTH1 Ferritin heavy chain
0.0019381046	H55933	EEF1A1 Translation elongation factor 1-alpha-1
0.0014206767	H68220	FAU Finkel-Biskis-Reilly murine sarcoma virus ...
0.0019015074	H77302	UBA52 Ubiquitin A-52 residue ribosomal prot...
0.016992867	H79852	"RPLP2 Hemoglobin, beta"
0.01155293	H88360_f	"GNAS1 Guanine nucleotide binding protein (G...
0.011218965	H88360_i	"GNAS1 Guanine nucleotide binding protein (G...

Clustering: 1_3_42_2.089E5

View Type: SOM Centers

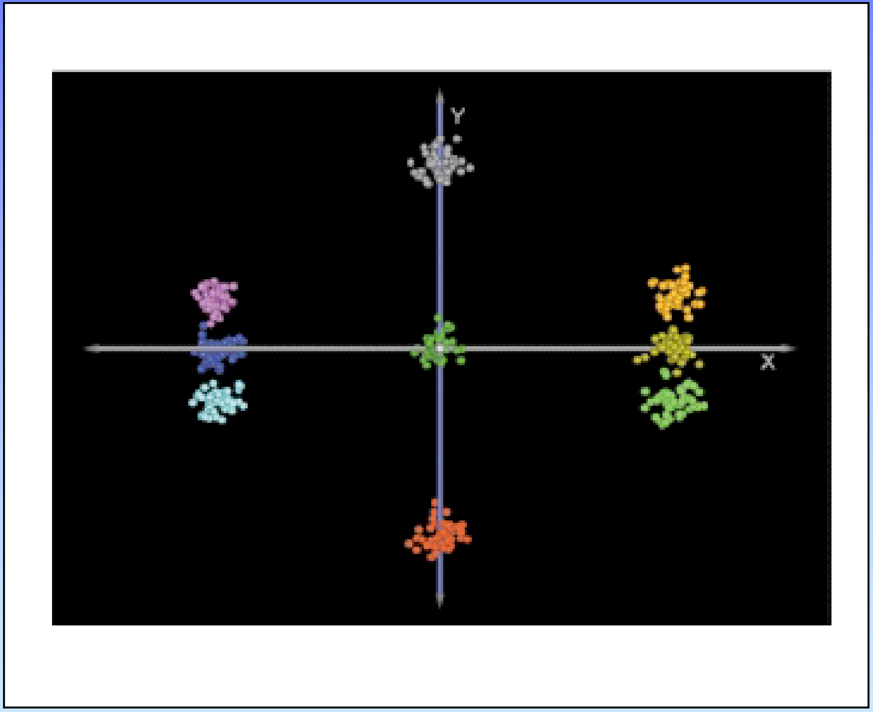
Principal Component Analysis

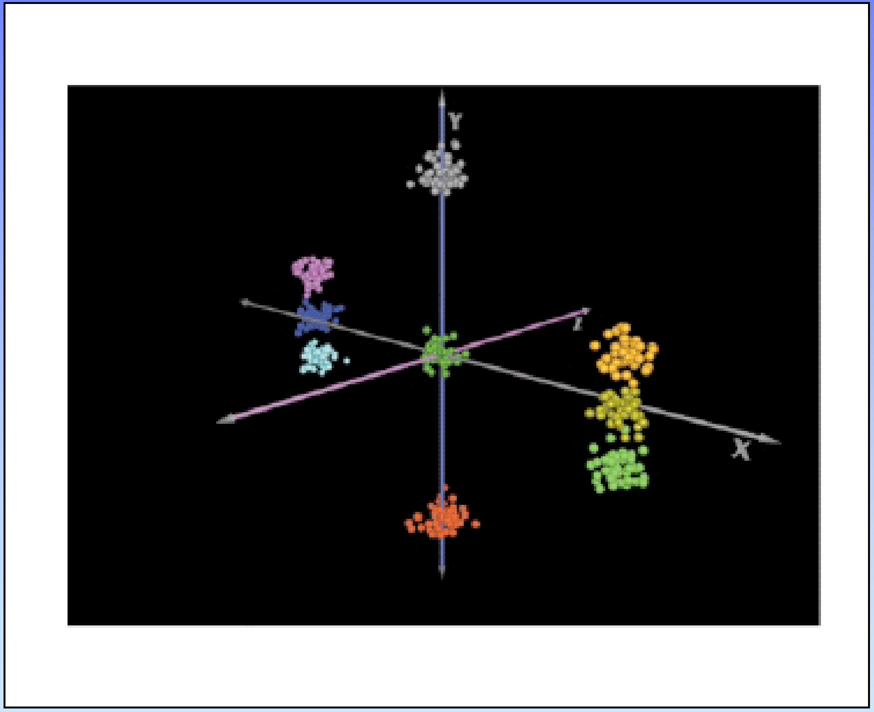
(Adapted from Quackenbush 2001)

- Data reduction method
- AKA singular value decomposition
- Used to pick out patterns in data
- Provide projection of complex data sets onto reduced, easily visualized space
- Difficult to define precise clusters but can give you an idea of # of clusters for SOMs or k-means

Principal Component Analysis

(Quackenbush 2001)





Quackenbush 2001

“One must remember that the results of any analysis have to be evaluated in the context of other biological knowledge.”

Supervised Learning

(Adapted from Quackenbush 2001)

- Useful if you have some previous information about which genes are expected to cluster together

- Support Vector Machine (SVM)

- Start with training set (eg. positive and negative examples)

- SVM learns to distinguish between members and non-members of a class

Warnings

(Adapted from Quackenbush 2001)

- Classification is dependent on
  - clustering method used
  - normalization of data
  - measure of similarity

Citations

- Brazma A and Vilo J. Minireview: Gene expression data analysis. *FEBS Letters* 480:17-24, 2000.
- Quackenbush J. Computational Analysis of Microarray Data. *Nature Review | Genetics* 2:418-427, 2001.
- Quackenbush J. Microarray data normalization and transformation. *Nature Genetics Supp*. 32:496-501, 2002.
- Dudoit S and Gentleman R. Classification in microarray experiments. Statistics and Genomics Short Course - Lecture 5, January 2002 (http://www.bioconductor.org/workshop.html)

Available Tools

- GeneCluster (WI/MIT Genome Center)
- Cluster & TreeView (Eisen)
- GeneSpring (Silicon Genetics)
- Spotfire (Spotfire)
- R Statistics Package/Bioconductor
- Matlab (modules from Churchill, JAX)

Lists of Tools

- Rockefeller University (formerly)
  - http://www.nslij-genetics.org/microarray/
- R Statistics Package Microarry Tools
  - http://www.stat.uni-muenchen.de/~strimmer/rexpress.html
- Bioconductor Project
  - http://www.bioconductor.org/
- EBI
  - http://ep.ebi.ac.uk/Links.html
  - http://ep.ebi.ac.uk/EP/