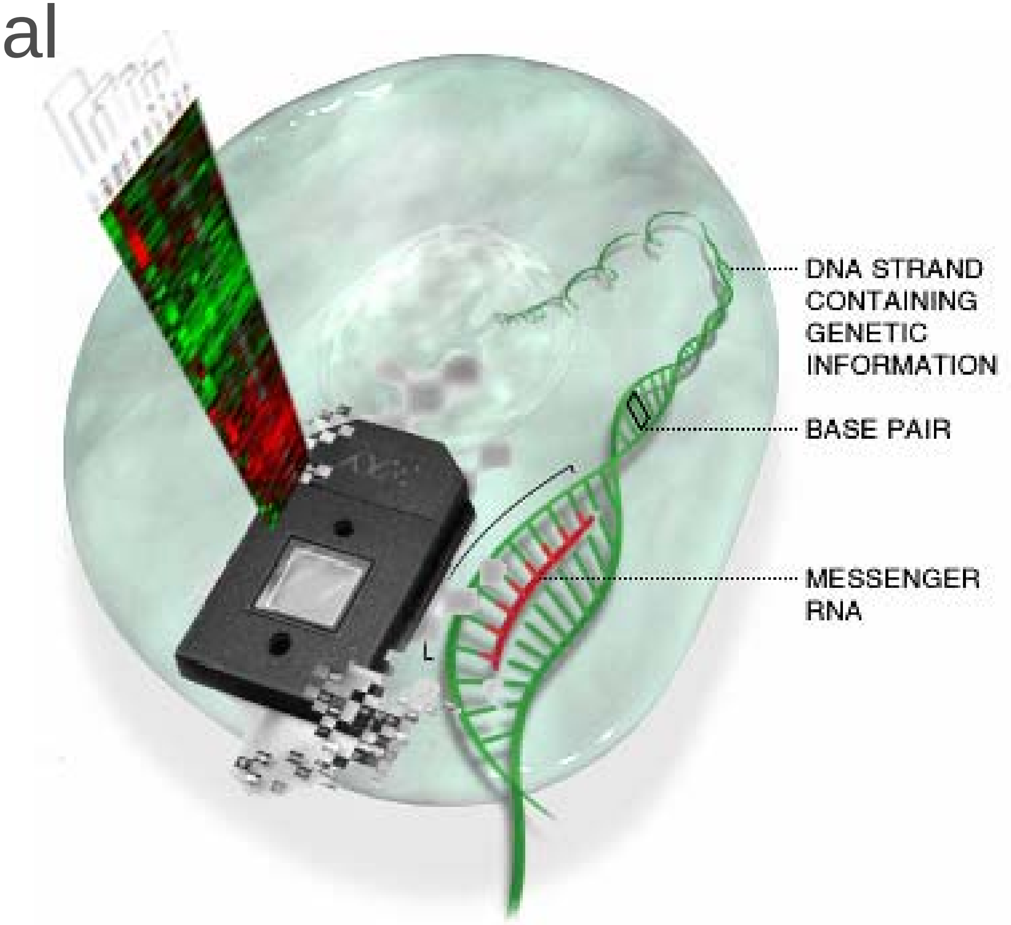
A Survey of Free Microarray Data Analysis Tools

Piali Mukherjee

Institute for Computational Biomedicine (ICB)

http://icb.med.cornell.edu

pim2001@med.cornell.edu



http://www.trii.org

Experimental Design

Biological Question → Microarray Experiment

Data Pre-Processing

Image quantification and analysis

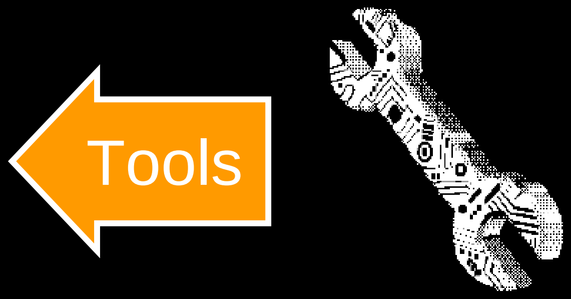
Quality Control: filtering and normalizing each chip for noise/backgound

Data Analysis

Normalizing multiple experiments

Statistical Estimation and Testing

Clustering and Prediction



Biological Verification

RT-PCR, RPA, northern blots etc.

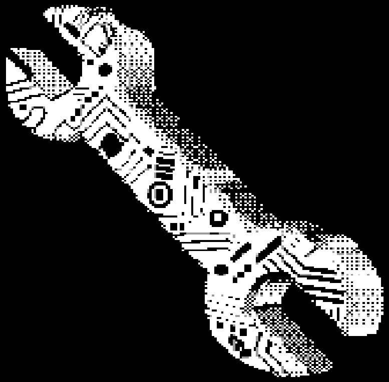
Functional Analysis

Functionally Clustering based on databases, Pathway analysis

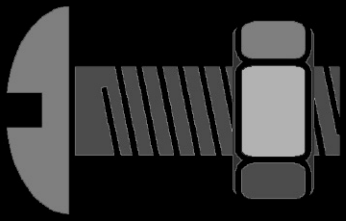
Data Analysis

- **Quality Control (Background correction and Filtering)**
  - Example: filtering the dataset to include only positive values above background
- **Normalization (or Scaling)**
  - Per chip and multi-chip
  - Example: Global Averaging or Loess (locally weighted regression) smoothing for a custom two color experiment, MBEI (Model Based Expression Index) or RMA (Robust Multi-chip Average) for Affymetrix experiment
- **Statistical Analysis (or Calculating Differential Expression)**
  - Ranking genes using a statistical test for significance (example: ANOVA, T-test or Z-score)
  - Multiple testing Correction (example: Bonferroni correction)
  - Selecting a significance cut off (example: p-value < 0.05)
- **Clustering and Classification (Studying Co-regulation)**
  - Hierarchical or K-means
  - supervised or unsupervised
  - SOM (Self Organizing Maps), LDA (Linear Discriminant Analysis), PCA (Principal Components Analysis)

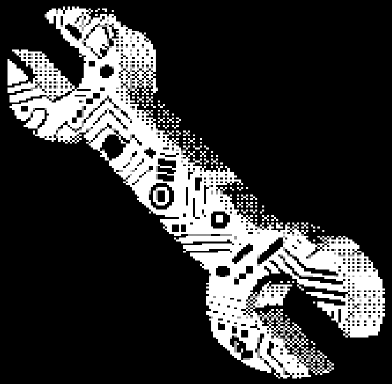
Free Data Analysis Tools



- **Clustering Tools:**
    - **Cluster / Tree-View** (Hierarchical Clustering)
    - **CAGED** (Bayesian/Supervised Clustering)
- **Analysis Suites:**
    - **D-Chip** (Model-based Analysis of Oligonucleotide Arrays)
    - **TIGR M4 Suite** (Analysis Suite for Spotted Two-Color Arrays)
    - **BioConductor** (R based Statistical Analysis)
- **Web based analysis tools:**
    - **Cyber-T**
    - **SNOMAD**



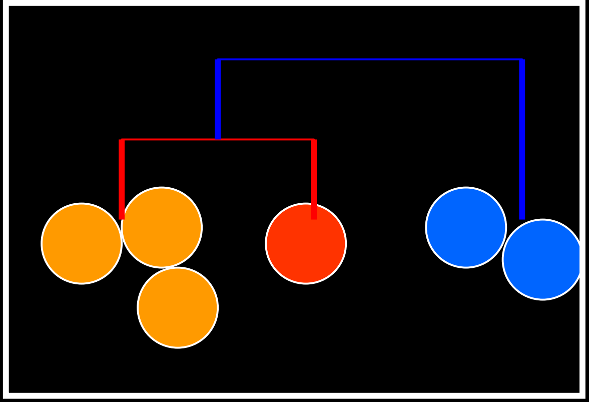
Clustering Tools



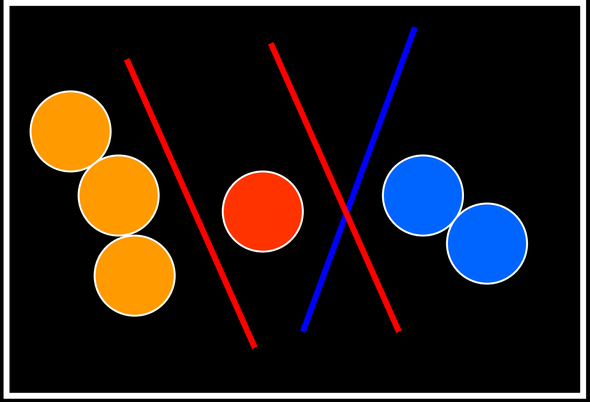
Cluster Analysis

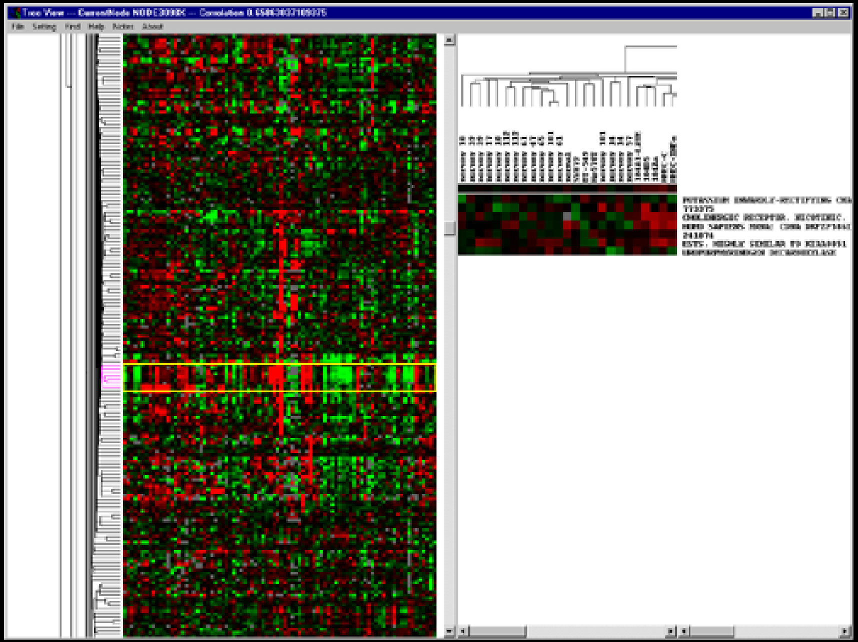
Standard statistical algorithms to arrange genes according to similarity in pattern of gene expression.

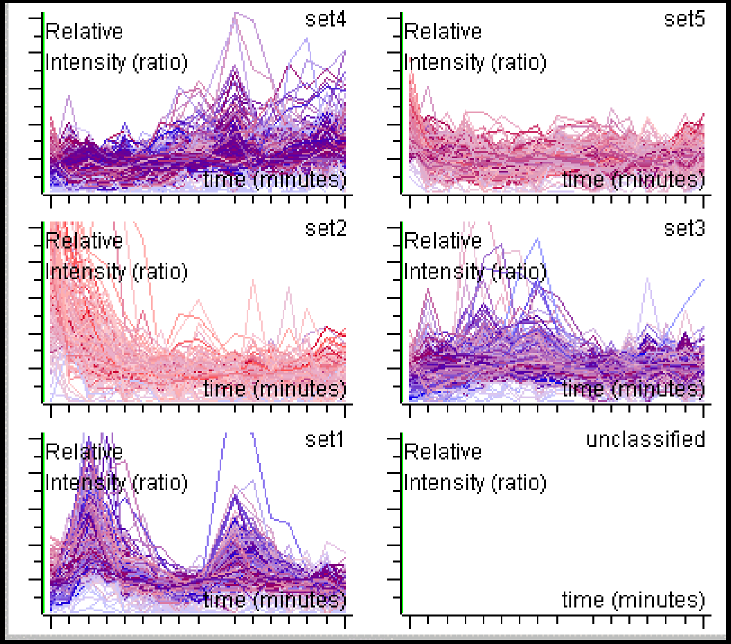
Hierarchical Clustering

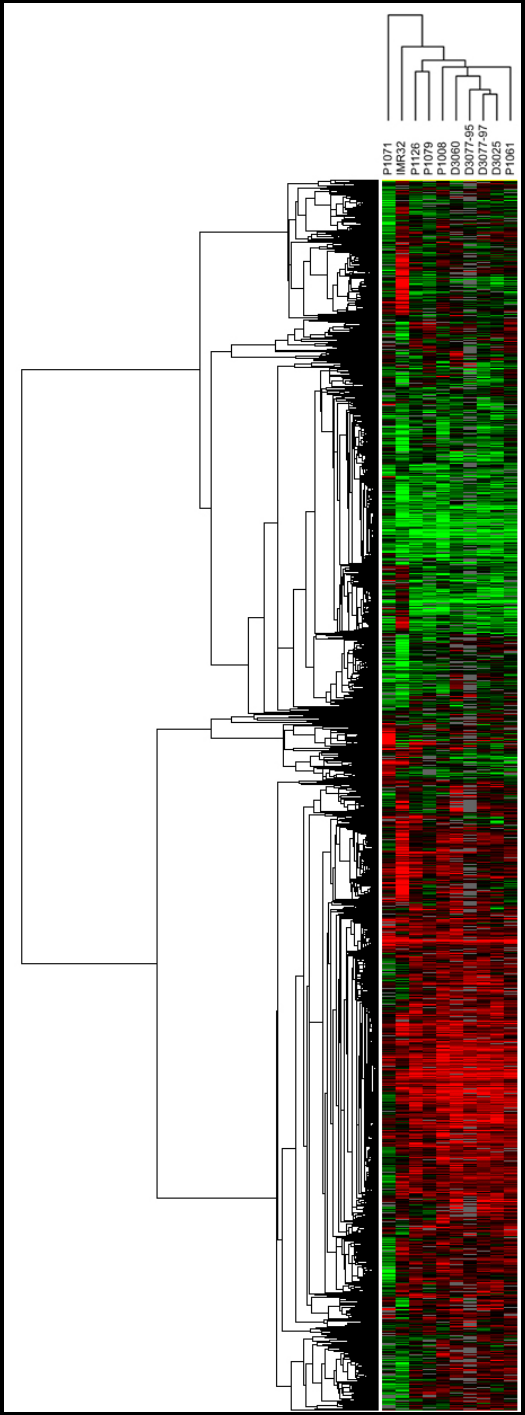


Partition Clustering




Tree View — CurrentNode NODE3098X — Correlation 0.658630037109375
File
Setting
Find
Help
Notes
About
POTASSIUM INWARDLY-RECTIFYING CHA
CHOLINERGIC RECEPTOR, NICOTINIC,
HOMO SAPIENS MRNA; CDNA DKFZP
ESTS, HIGHLY SIMILAR TO KIAA





Cluster / Tree View

Similarity metric = distance metric

Clustering genes:

Co-expression and Co-regulation go together – easier to visualize possible functional groups

Clustering arrays:

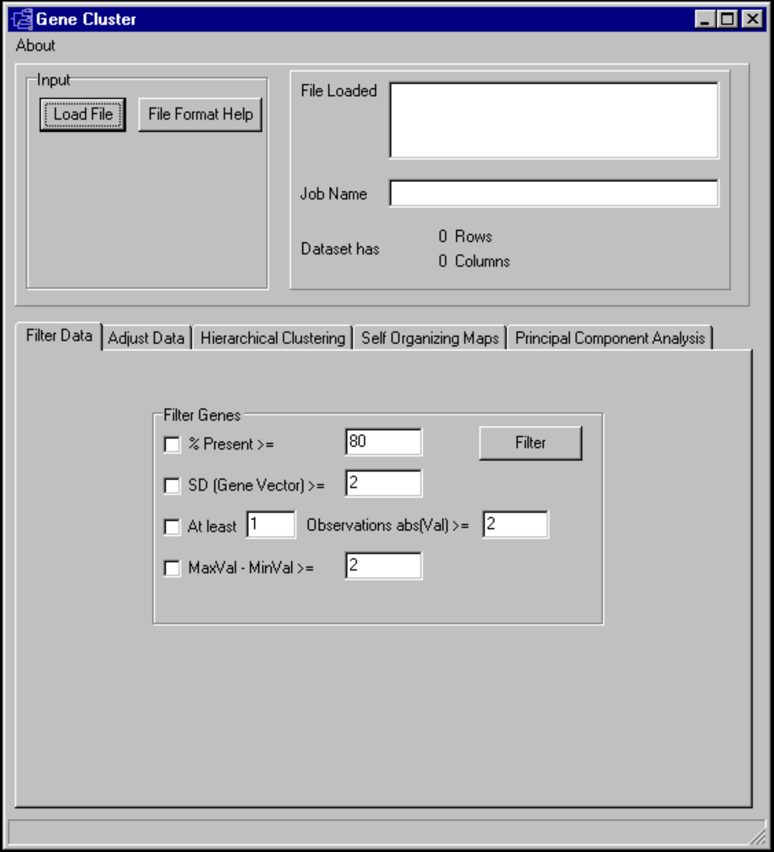
Finding new sub-classes in sample space

Two-way clustering

Available at: http://rana.lbl.gov/EisenSoftware.htm
(Eisen Lab, Stanford)

Publication: Eisen et al. (1998) PNAS 95:14863

Cluster



File Loaded

Job Name

Dataset has 0 Rows 0 Columns

Filter Genes

□ % Present >= 80

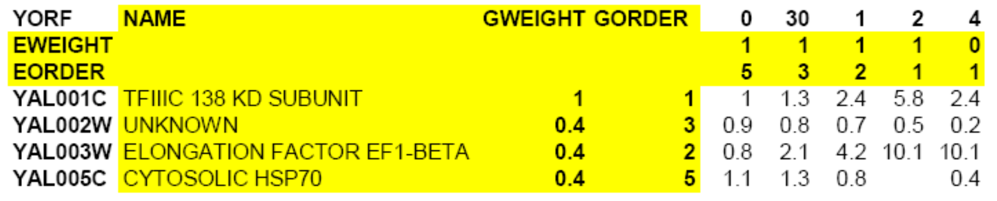
□ SD (Gene Vector) >= 2

□ At least 1 Observations abs(Val) >= 2

□ MaxVal - MinVal >= 2

http://rana.lbl.gov/EisenSoftware.htm

Load formatted data (tab-delimited text)



YORF	NAME	GWEIGHT	GORDER	0	30	1	2	4
EWEIGHT				1	1	1	1	0
EORDER				5	3	2	1	1
YAL001C	TFIIIC 138 KD SUBUNIT	1	1	1	1.3	2.4	5.8	2.4
YAL002W	UNKNOWN	0.4	3	0.9	0.8	0.7	0.5	0.2
YAL003W	ELONGATION FACTOR EF1-BETA	0.4	2	0.8	2.1	4.2	10.1	10.1
YAL005C	CYTOSOLIC HSP70	0.4	5	1.1	1.3	0.8		0.4

Filter data

SD ≥ 2, absolute expression value ≥ 2, % present ≥ 80 etc.

Adjust data

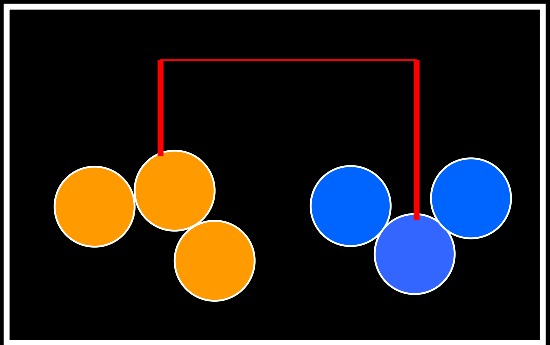
Log transform, mean/median center, row/column normalize etc.

Hierarchical clustering

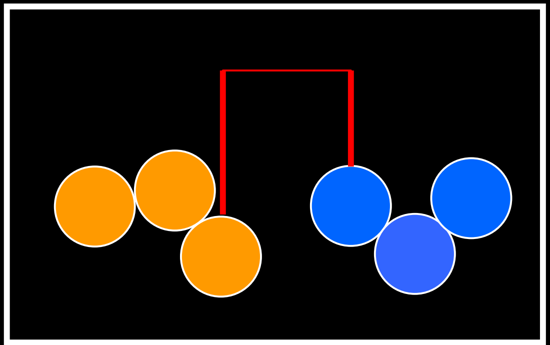
Similarity Metrics: 4 flavours of the Pearson’s correlation [r]

- **Centered** (textbook formula – linear regression in a 2 dimensional scatter plot)
- **Uncentered** (assumes mean = 0)
- **Spearman’s** (Non-parametric version)
- **Kendal’s Tau** (Non-parametric version)

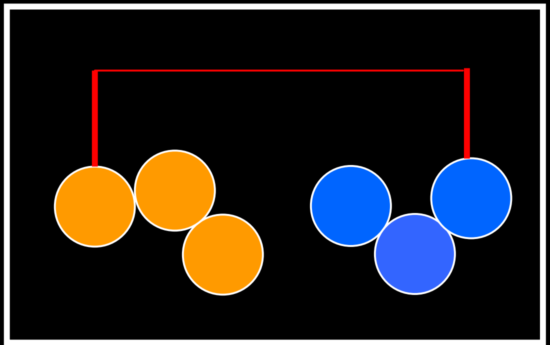
Cluster: Hierarchical clustering



Average linkage: the average distance between objects from two clusters



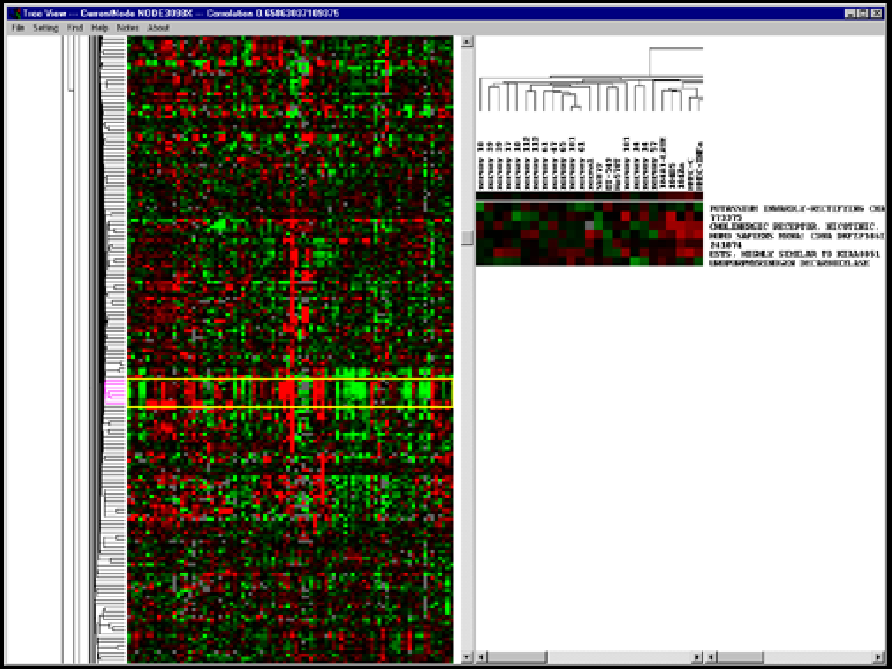
Single linkage: the distance between the closest objects from two clusters



Complete linkage: the distance between the most distant objects from two clusters

Manual: http://rana.lbl.gov/EisenSoftware.htm

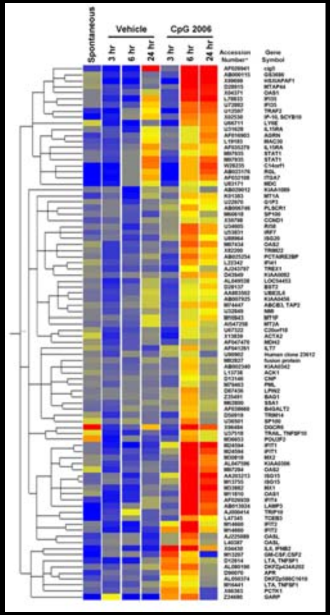
TreeView


Tree View — CurrentNode NODE3008X — Correlation 0.65863037109375
File Setting Find Help Notes About

- Visualization for text output from cluster
- Customize colors
- Various formats for import into publications

Other clustering

- K-Means (partition clustering)
- Self Organizing Maps (SOM)
- Principal Components Analysis (PCA)



Other software cluster analysis and from the Eisen Lab:

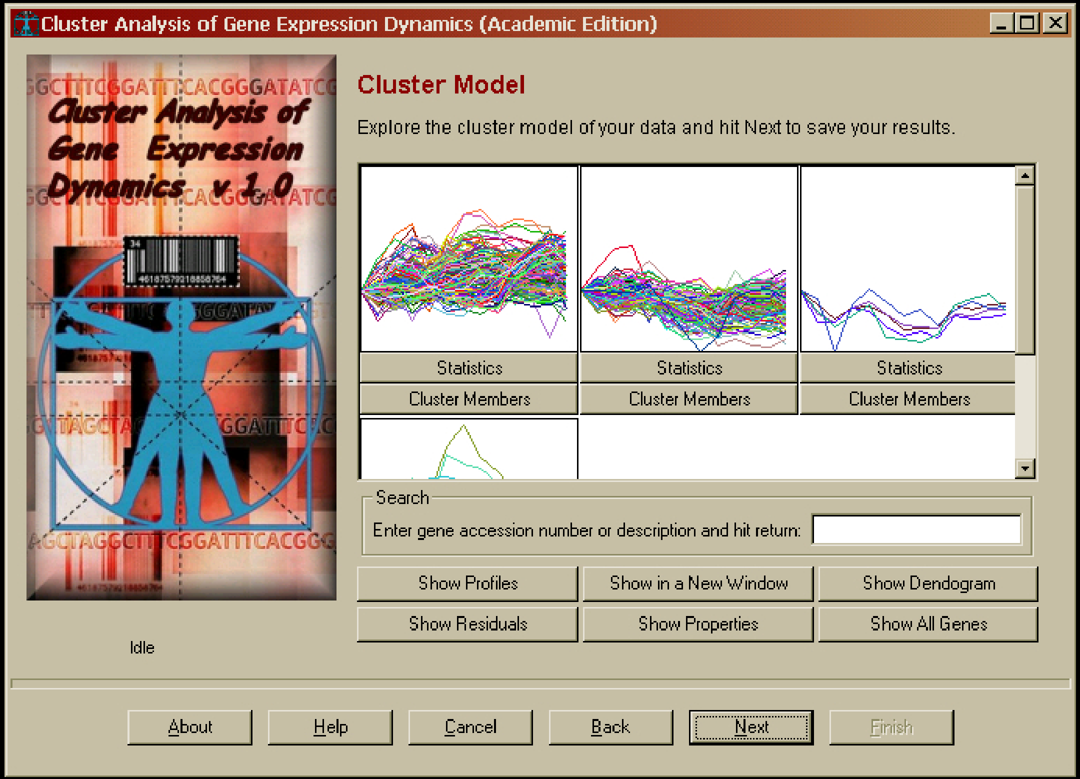
- Fuzzy K (K-means clustering software)
- Maple (java based alternative to TreeView – also allows visualizations for K-means clustering output)

http://rana.lbl.gov/EisenSoftware.htm

CAGED **C**luster **A**nalysis of **G**ene **E**xpression **D**ynamics.

Ramoni et al., 2002 (Harvard)

- **Bayesian clustering algorithm** – supervised clustering
- **Designed for temporal (time series) data -** but can be used as a Bayesian clustering program on a-temporal expression data.
- **Machine learning:** does not assume that each gene has independent observation – remembers old observations as it processes new ones



Cluster Analysis of Gene Expression Dynamics (Academic Edition)

Cluster Model

Explore the cluster model of your data and hit Next to save your results.

Search

Enter gene accession number or description and hit return:

Idle

Seeks hypothesis that has the maximum probability given the observed data by exploring all ways of combining the observed data points, computing its posterior probability (given the observed data), and selecting the most probable one.

http://www.genomethods.org/caged/

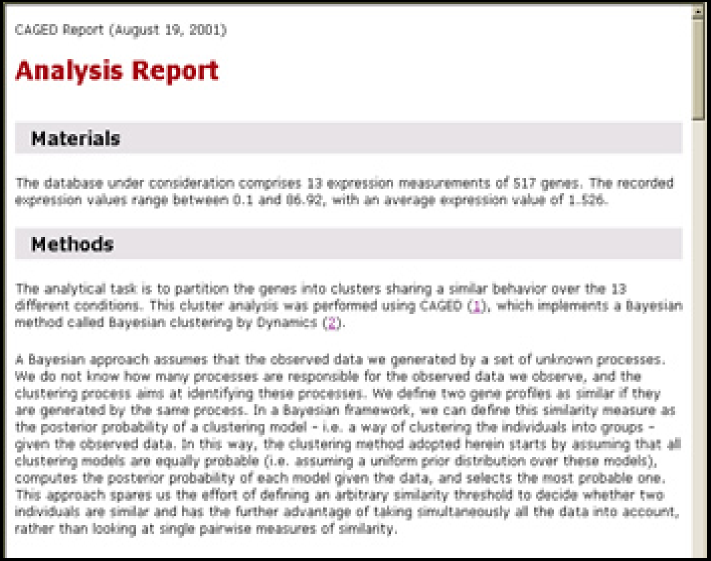
CAGED

http://www.genomethods.org/caged/

- **Model based clustering**
  - – More sensitive than hierarchical clustering but no arbitrary threshold for number of clusters like K-means clustering

- **Modeling Parameters**
  - Robustness (Markov order, prior precision, gamma value, Bayes factor)
  - Similarity measure/metric used in the heuristic/learning process (Euclidian, correlation, none etc.)
  - Transformation (log, square, square root etc.)
  - Generates a separate most probable statistical model for each cluster

- **Analysis report**
  - HTML with links to external databases (UniGene, GenBank etc.)
  - Generates methods section
  - Importable file formats for images
  - Allows popular visualizations: (histograms, dendograms/heatmaps etc.)



CAGED Report (August 19, 2001)

Analysis Report

Materials

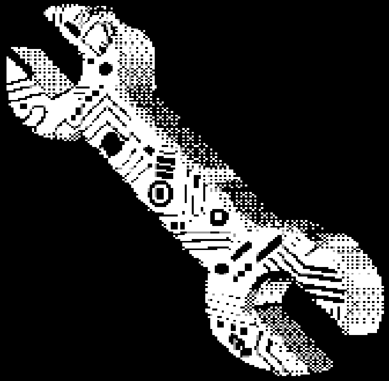
The database under consideration comprises 13 expression measurements of 517 genes. The recorded expression values range between 0.1 and 06.92, with an average expression value of 1.526.

Methods

The analytical task is to partition the genes into clusters sharing a similar behavior over the 13 different conditions. This cluster analysis was performed using CAGED (1), which implements a Bayesian method called Bayesian clustering by Dynamics (2).

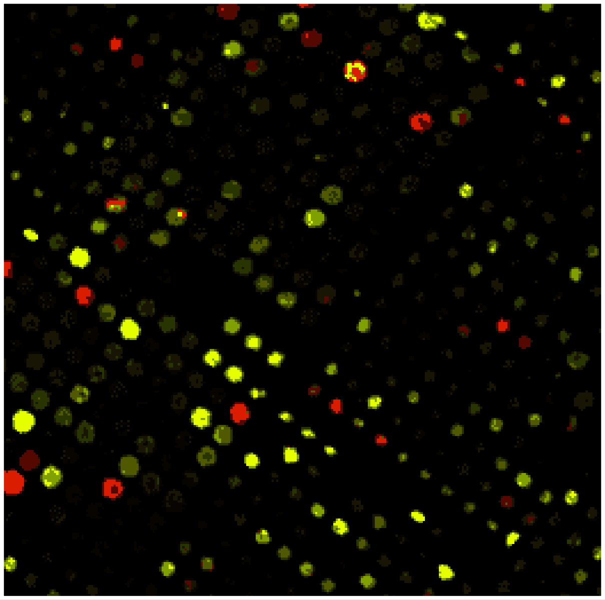
A Bayesian approach assumes that the observed data we generated by a set of unknown processes. We do not know how many processes are responsible for the observed data we observe, and the clustering process aims at identifying these processes. We define two gene profiles as similar if they are generated by the same process. In a Bayesian framework, we can define this similarity measure as the posterior probability of a clustering model - i.e. a way of clustering the individuals into groups - given the observed data. In this way, the clustering method adopted herein starts by assuming that all clustering models are equally probable (i.e. assuming a uniform prior distribution over these models), computes the posterior probability of each model given the data, and selects the most probable one. This approach spares us the effort of defining an arbitrary similarity threshold to decide whether two individuals are similar and has the further advantage of taking simultaneously all the data into account, rather than looking at single pairwise measures of similarity.

Analysis Suites

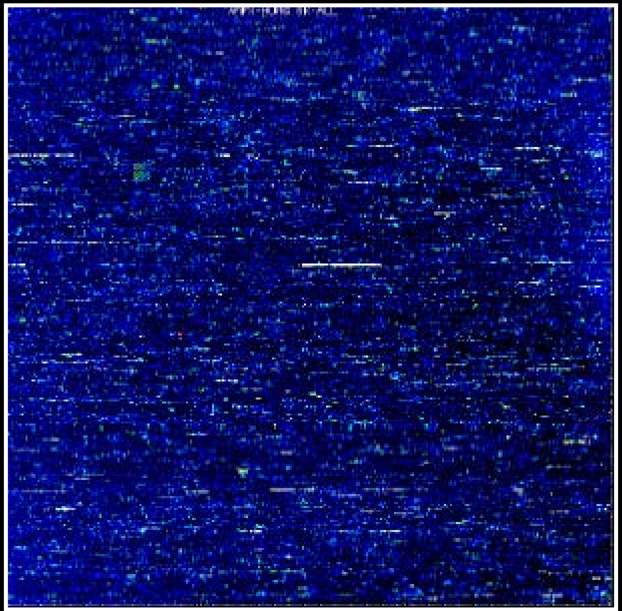


D-Chip	Oligonucleotide arrays
TIGR M4 Suite	2 dye spotted arrays
BioConductor	Both oligonucleotide and 2 dye arrays

Spotted arrays vs. Affymetrix arrays



- One probe (clone, usually cDNA) per gene
- **Two targets per array**

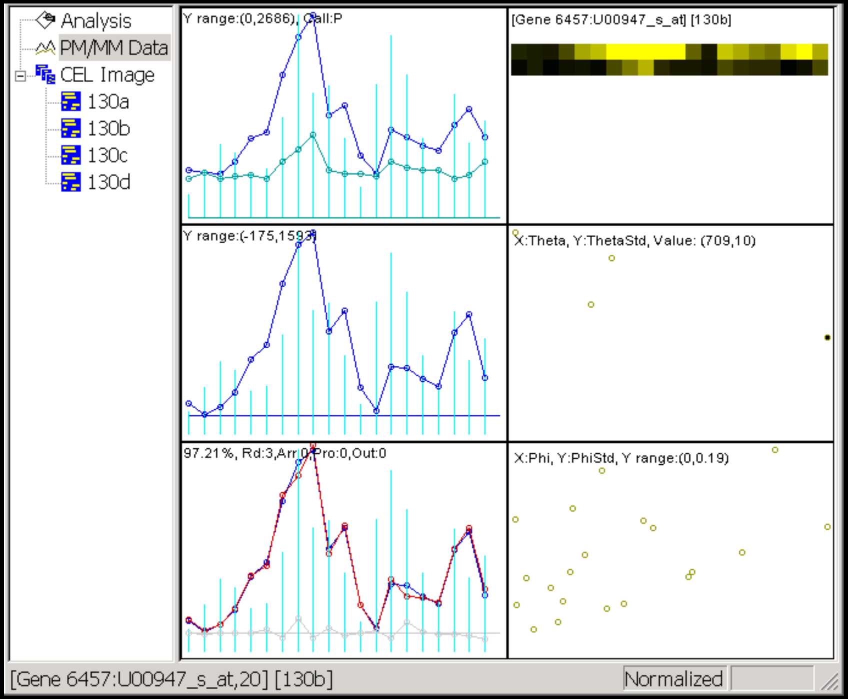


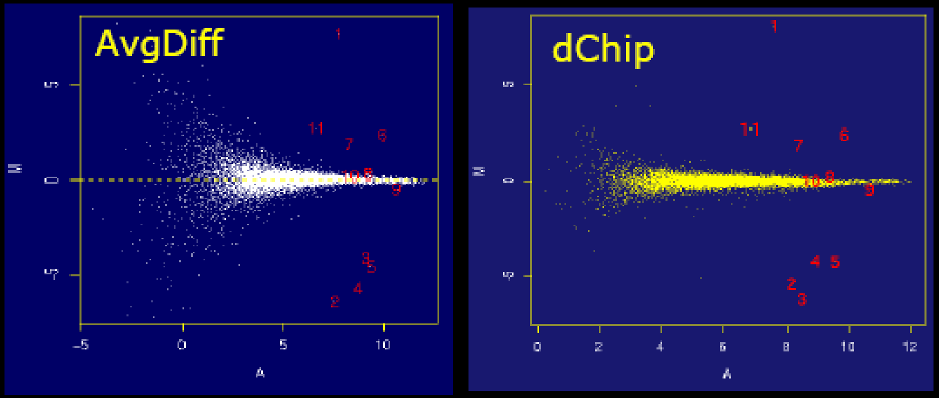
- 16-20 probe pairs (oligonucleotides) per gene
- **One target sample per array**

dChip

http://www.dchip.org/ (Wong lab, Harvard)

- Analysis of **oligonucleotide arrays** (can be used for 2 Dye arrays but mostly useful for Affymetrix type arrays)
- Reads Affy .CEL and .DAT (image files) as well as text files
- **Model based Expression Index (MBEI)** - Creates models from Probe data to calculate expression of the gene. (Not dependant on mismatch values)


Analysis
PM/MM Data
CEL Image
130a
130b
130c
130d
Y range:(0,2686), Call:P
[Gene 6457:U00947_s_at] [130b]
Y range:(-175,1593)
X:Theta, Y:ThetaStd, Value: (709,10)
97.21%, Rd:3,Arr:0,Pro:0,Out:0
X:Phi, Y:PhiStd, Y range:(0,0.19)
[Gene 6457:U00947_s_at,20] [130b]
Normalized



MA plot: M = log (Ratio); A = log (Av. Intensity)

- Instead of the average (PM-MM) analysis used by the Affy software (Av. Diff.) – dChip calculates model based errors and eliminates outliers and false positives http://www.dchip.org/ (Wong lab, Harvard)

dChip

- Allows for within chip normalization and normalization for several chips.
- Allows filtering, comparison analysis (T-test / P-value), mapping genes to chromosomes, hierarchical clustering, Linear Discriminant Analysis (LDA), PCA etc.
- Recently added features for SNP array analysis and to connect to GO databases for functional annotation
- You can combine comparisons: for example look at overlapping gene lists from two different sets of analysis etc.
- Also allows for comparison analysis of different species chips (Mouse and Human) or different chips from the same species (Human: HG_U95A and Hu6800) etc.
- Interface with R software for advanced statistical analysis

TIGR M4 Suite

http://www.tigr.org/software/tm4/
(Quackenbush lab, TIGR)

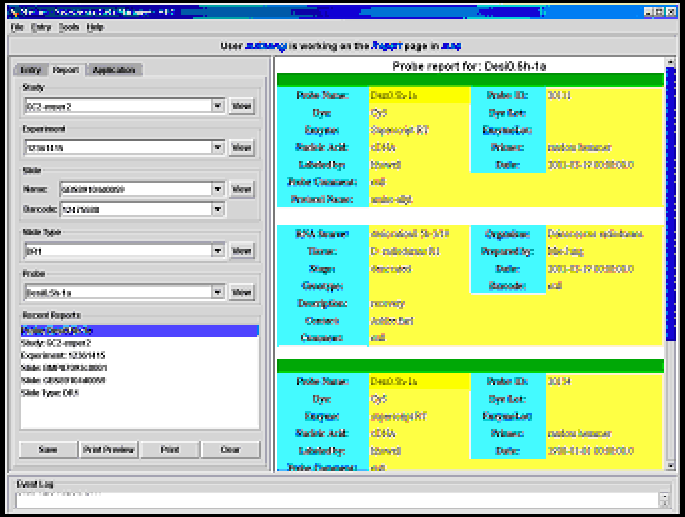
- Open source software developed mostly for spotted two-color arrays, but many of the components can be easily adapted to work with single-color formats such as GeneChips™(Affymetrix)

- The TM4 suite of tools consist of four major applications:
  - Microarray Data Manager (**MADAM**)
    - Minimal Information About a Microarray Experiment (MIAME) -compliant MySQL database
  - **Spotfinder** (image quantification tool)
  - Microarray Data Analysis System(**MIDAS**)
  - Multiexperiment Viewer(**MeV**)

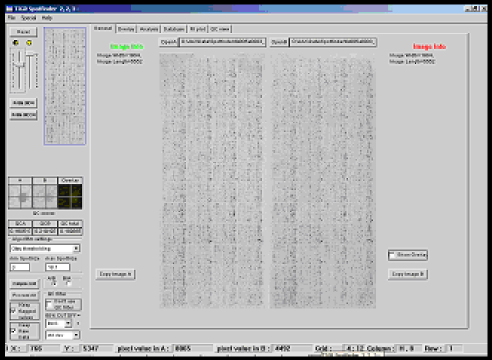
TIGR M4 Suite

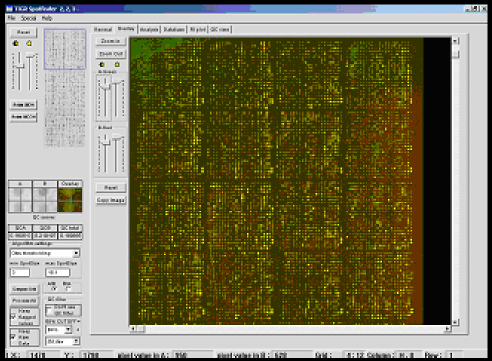
http://www.tigr.org/software/tm4/
(Quackenbush lab, TIGR)

- MADAM - designed to load and retrieve microarray data to and from a database
  - MySQL Database supplied with the software but works with any JDBC compliant database
  - Java based - Provides data entry forms, data report forms – MIAME compliant



- Spotfinder – basic image analysis for 2 color spotted arrays
  - Able to calculate and subtract background
  - Outputs in formats for other TIGR software as well as tab delimited and excel format

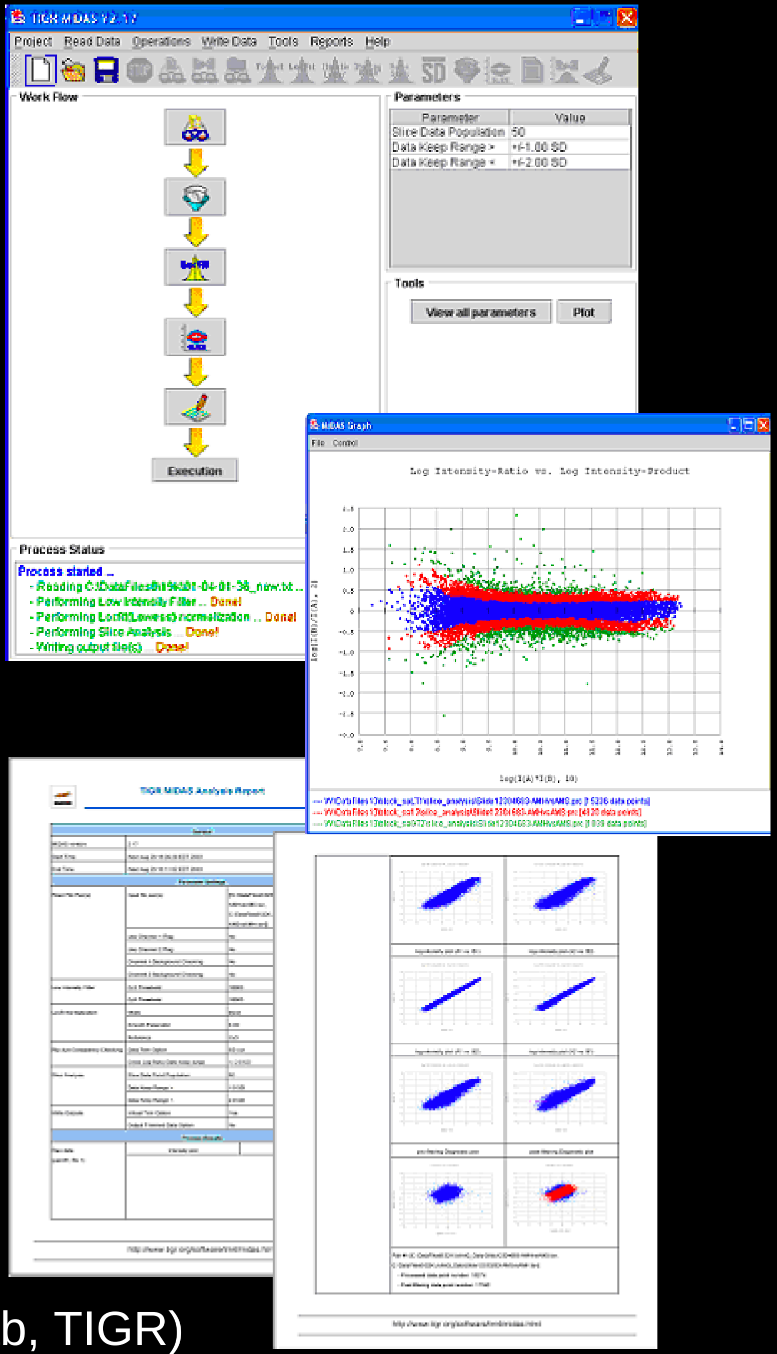




- ExpressConverter - file format transformation tool that reads GenePix file as input and generates output for TIGR microarray analysis software (MIDAS, MeV etc.)

MIDAS

- Normalization, Standardization and Filtering tool
- Global and Local normalization
  - Loess locally weighted linear regression
- Iterative linear regression and iterative log-mean centering
- Ratio statistics, Flip-Dye consistency
  - Also allows low-intensity cutoff, replicate consistency trimming
- Standard Deviation (SD) regularization
  - Adjusts Cy3-Cy5 scales for each block to have similar SD
- Z-score filtering (Slice Analysis)
- Automated report and graphs


TIGR MIDAS V2.17
Project Read Data Operations Write Data Tools Reports Help
Work Flow
Execution
Parameters
Tools
View all parameters
Plot
Process Status
MIDAS Graph
File Control
Log Intensity-Ratio vs. Log Intensity-Product
TIGR MIDAS Analysis Report

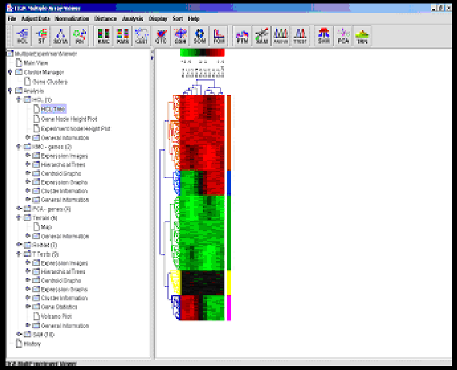
Parameter	Value
Slice Data Population	50
Data Keep Range >	+/-1.00 SD
Data Keep Range <	+/-2.00 SD

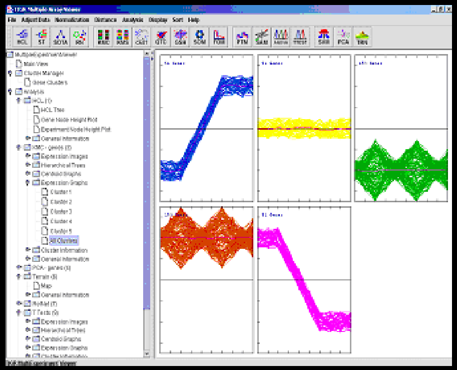
```
Process started ...
  - Reading C:\DataFiles\hi96\01-04-01-38_new.txt ...
  - Performing Low Intensity Filter ... Done!
  - Performing Lo(c)(Lowess) normalization ... Done!
  - Performing Slice Analysis ... Done!
  - Writing output file(s) ... Done!
```

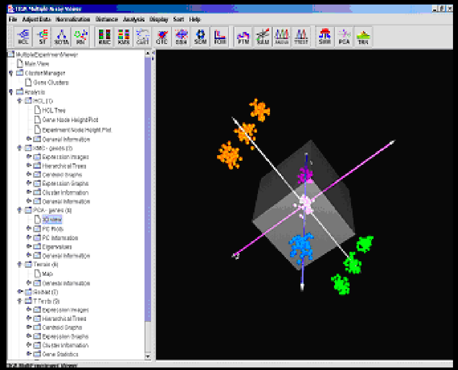
http://www.tigr.org/software/tm4/ (Quackenbush lab, TIGR)

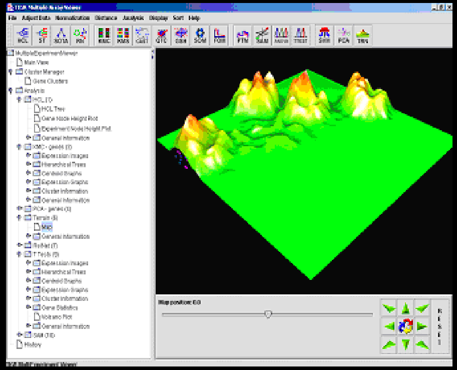
MeV – Multi-experiment Viewer

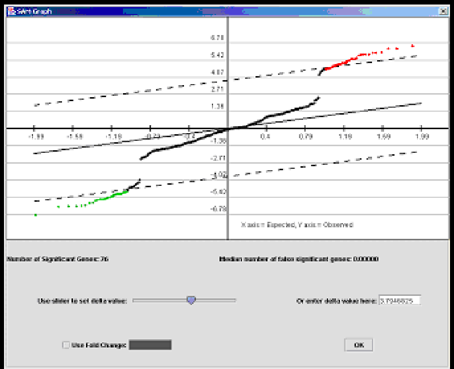
- Hierarchical and K-means clustering
- SOM, PCA, SOTA (self organizing trees – SOM type divisive approach), Figures of Merit (FOM)
- SAM
- T-test (permutations and Bonferroni correction), ANOVA
- Support Vector Machines (supervised learning)
- Gene Shaving (nested clusters)
- Randomization/Resampling
    - Bootstrapping (resampling with replacement)
    - Jackknifing (resampling without replacement)
- Relevance Networks (genes whose expression profiles are predictive of one another based on functional relationships)











X axis = Expected, Y axis = Observed

Number of Significant Genes: 76

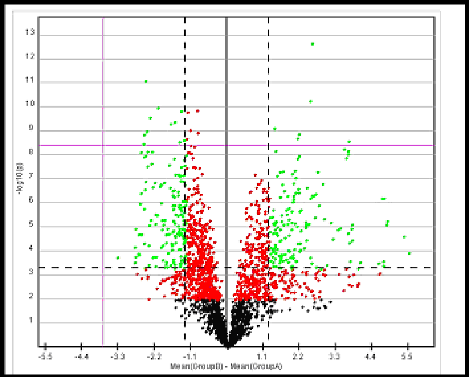
Median number of false significant genes: 0.00000

Use slider to set delta value:

Or enter delta value here: [illegible]

□ Use Fold Change:

OK



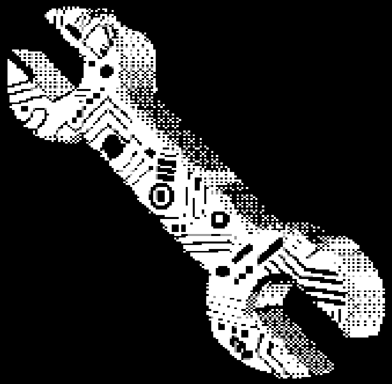
http://www.tigr.org/software/tm4/ (TIGR)

BioConductor

http://www.bioconductor.org/

- **R programming language environment** (open source version of S – S-Plus is the commercial software)
  - http://www.r-project.org
- **Requires some programming knowledge** (object oriented programming based).
- **Widgets:** graphical user interfaces have been created for some analyses
- Many applications for **both 2-dye and Affymetrix** type data.
- Allows all **popular normalization (RMA), filtering, plotting and statistical analysis (new algorithms constantly available)** and also allows you to create your own analysis packages and pipelines.
- "**annotate**" package allows annotation and literature WWW resources in real time and HTML report

Web-based Tools



- **CYBER-T – Baldi and Long, 2002 (UCI)**
  - http://visitor.ics.uci.edu/genex/cybert/ (can also be downloaded on a Unix/Linux computer as an R package)
  - Separate interfaces for 2-dye data and for data with separate control and experimental data sets (e.g. Affymetrix data)
  - General statistics (mean, median, SD, variance, T-test, fold change, p-value), Posterior Probability of Differential Expression (PPDE – calculates global false positives and negatives), Bayesian SD estimation for T-test (corrects for local variance)

- **SNOMAD – Colantuoni et al, 2002 (Pevsner Lab, Johns Hopkins)**
  - http://pevsnerlab.kennedykrieger.org/snomadinput.html
  - Allows in depth statistical evaluation of two experiments (or av. of replicates from 2 conditions etc. – does not look at variance/SD between samples)
  - Background subtraction, global and local normalizations, local variance correction (loess fit), Z scores (function of fold change, local variance and standard deviation)

Excel and Microarray Analysis

- MicroSoft Excel is a popular tool of choice for researchers
- Open Source Excel Plugins
  - SAM (Significance Analysis of Microarrays): http://www-stat.stanford.edu/~tibs/SAM/
  - PAM (Prediction Analysis of Microarrays): http://www-stat.stanford.edu/~tibs/PAM/
  - BRB Array tools:

    http://linus.nci.nih.gov/BRB-ArrayTools.html
- Hands-on Workshop: Microarray Analysis in Excel
  - http://www.trii.org

Functional Analysis Tools

- **Open Source**
  - Onto-Express (http://vortex.cs.wayne.edu/projects.htm)
  - EASE (**E**xpression **A**nalysis **S**ystematic **E**xplorer): http://david.niaid.nih.gov/david/ease.htm
  - GeneMAPP (Gene Microarray Pathway Profiler): http://www.genmapp.org/
- **Commercial**
  - Ingenuity (Pathway Analysis): http://www.ingenuity.com/

Upcoming Workshop:

Functional Interpretation of
High Throughput Data

http://www.trii.org

Links / Resources

- ICB Microarray Section:
  - http://icb.med.cornell.edu/microarray/
- Y.F.Leung's Functional Genomics site (Harvard University):
  - http://www.nslij-genetics.org/microarray/
- Wentian Li's Microarray site:
  - http://ihome.cuhk.edu.hk/%7Eb400559/
- Stanford microarray Database:
  - http://genome-www5.stanford.edu/index.shtml
- Genome Gateway at nature.com (Nature Magazine):
  - http://www.nature.com/genomics/post-genomics/
- Microarray Analysis Tutorial (Jonathan Pevsner)
  - http://pevsnerlab.kennedykrieger.org/hinxton.html