Course #412
Analyzing Microarray Data using the mAdb System

July 15-16, 2008 1:00 pm - 4:00pm
madb-support@bimas.cit.nih.gov

- Intended for users of the mAdb system who are familiar with mAdb basics
- Focus on analysis of multiple array experiments

Esther Asaki, Yiwen He

Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
   - Class Discovery - clustering, PCA, MDS
   - Class Comparison - statistical analysis
     - t-test
     - ANOVA
     - Significance Analysis of Microarrays - SAM
   - Class Prediction - PAM

Various Hands-on exercises

1. mAdb system overview

mAdb Data Workflow

Upload Data	Quality Control	Prepare Dataset	Analysis/Model	Review Annotation
File Format • GenePix • MAS5 • GCOS 1.1 • ArraySuite • Agilent FE	**Project Summary** • Summary Statistics • Array images • Graphical Report	**Dataset Extraction** • Normalization • Spot Filtering	**Analysis Tools** • Class Discovery • Class Comparison • Class Prediction	**Annotation Tools** • Feature Report • Gene Ontology • BioCarta Pathway • KEGG Pathway

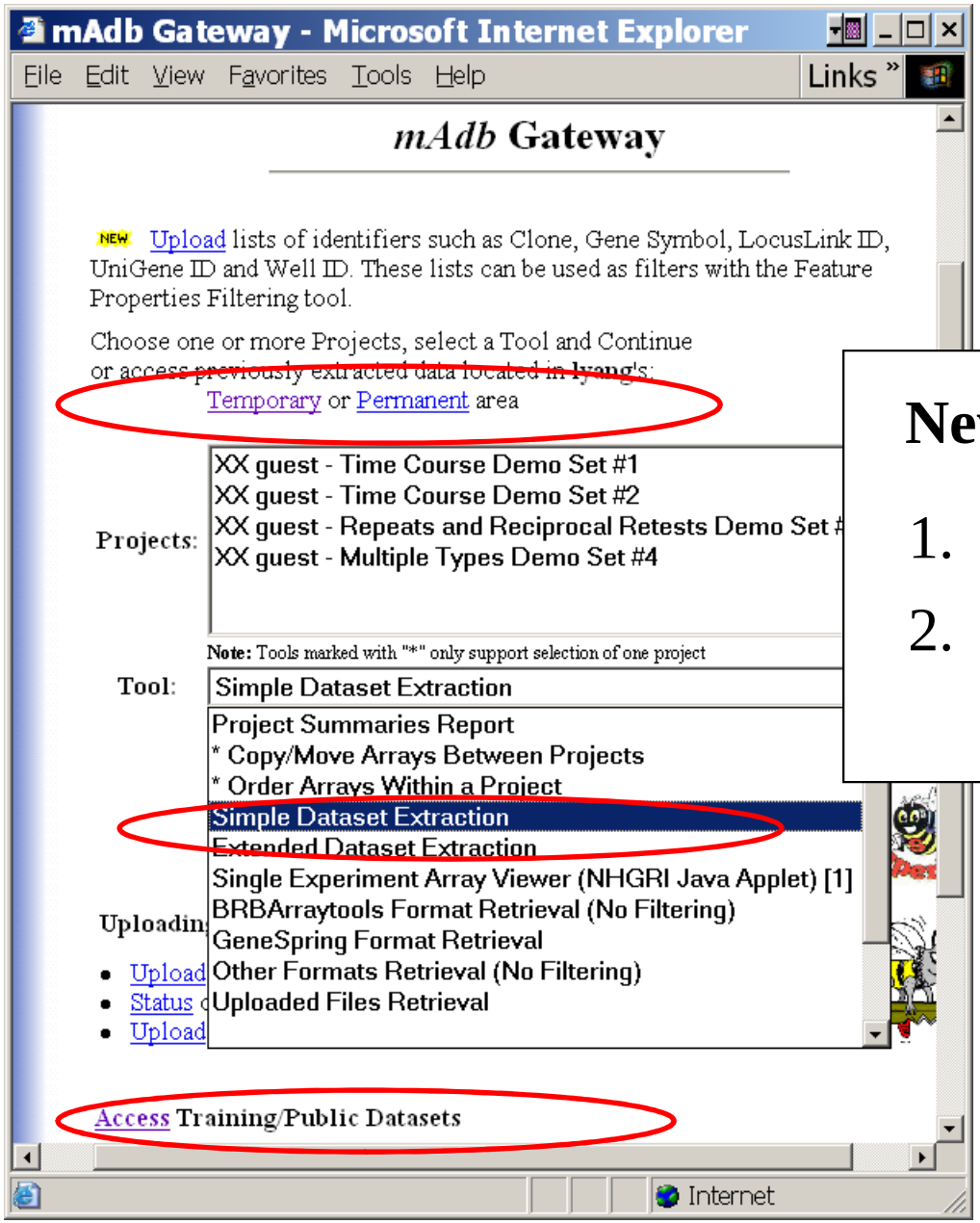
2. mAdb dataset overview

What is a dataset?

- mAdb Dataset
  - Collection of data from multiple experiments
  - Genes as rows and experiments as columns

		sample1	sample2	sample3	sample4	sample5	...
	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
Genes	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

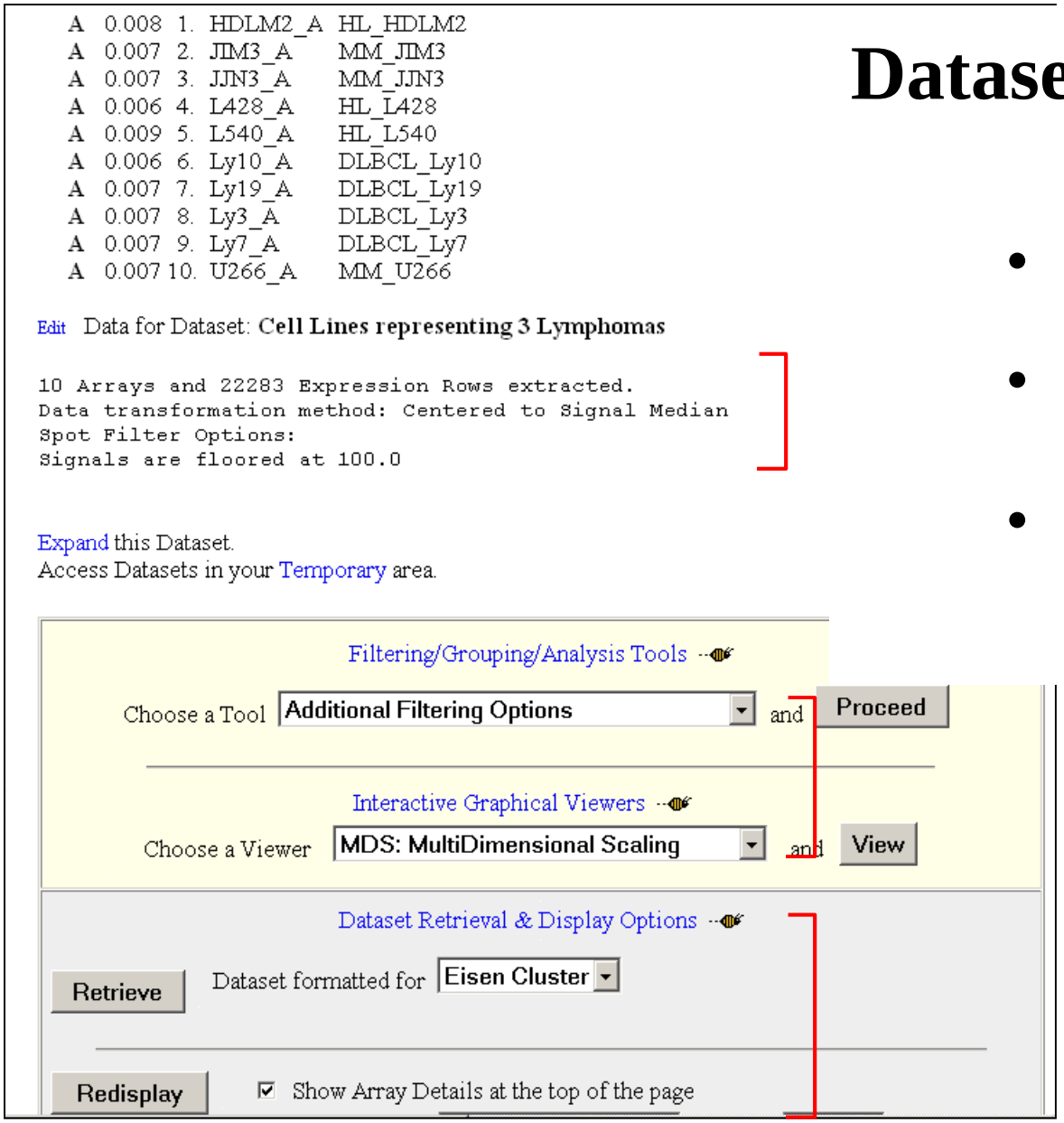
Gene expression level = (normalized) Log(Red signal / Green signal)



New or Existing Dataset:

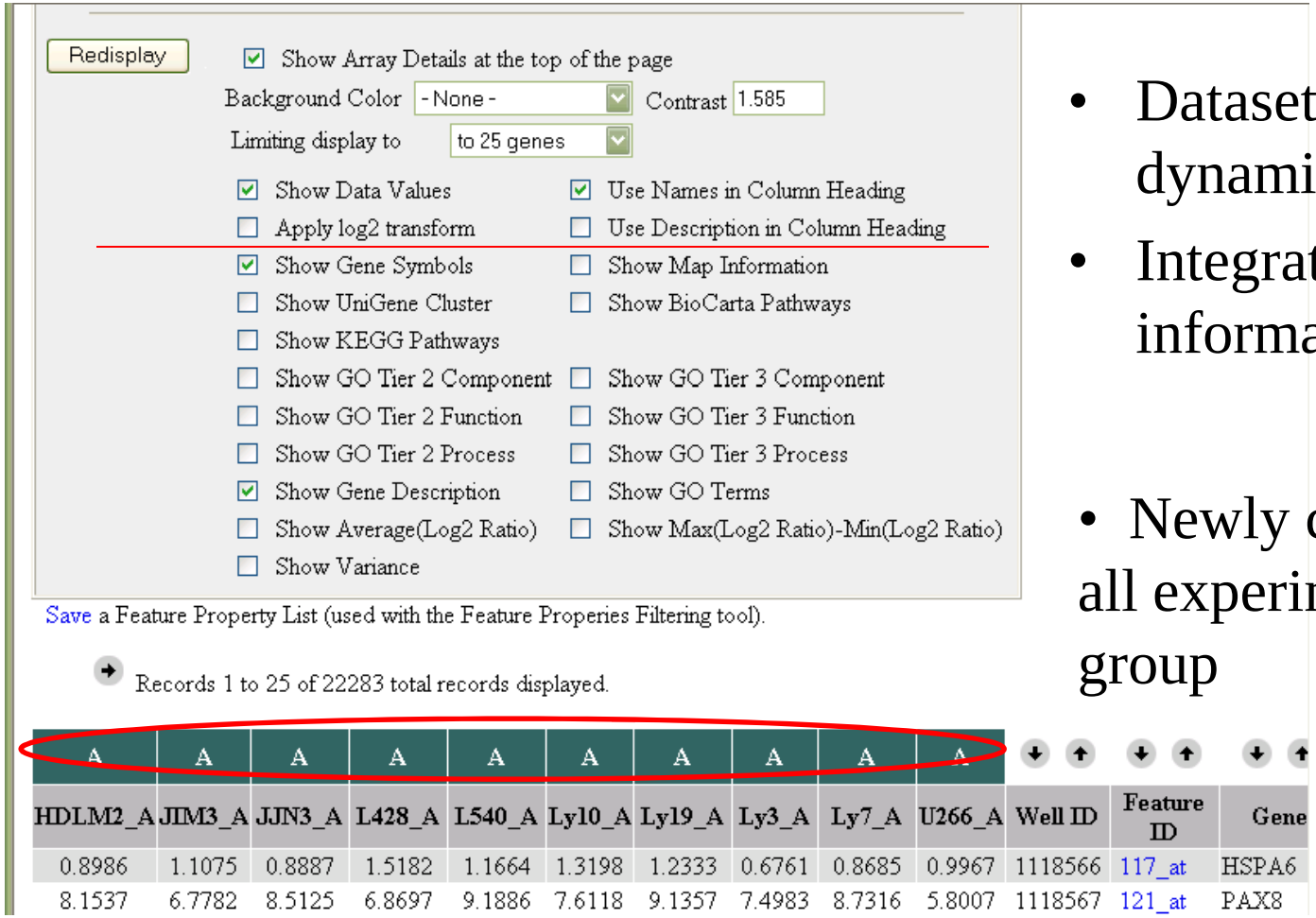
1. Create New Dataset
2. Access Existing Dataset

Dataset Display Page



- Dataset History
- Analysis Tools
- Retrieval and Display Options…

Dataset Display



A	A	A	A	A	A	A	A	A	A			
HDLM2_A	JIM3_A	JJN3_A	L428_A	L540_A	Ly10_A	Ly19_A	Ly3_A	Ly7_A	U266_A	Well ID	Feature ID	Gene
0.8986	1.1075	0.8887	1.5182	1.1664	1.3198	1.2333	0.6761	0.8685	0.9967	1118566	117_at	HSPA6
8.1537	6.7782	8.5125	6.8697	9.1886	7.6118	9.1357	7.4983	8.7316	5.8007	1118567	121_at	PAX8

- Dataset display options dynamic
- Integrated gene information

- Newly created dataset puts all experiments into a single group

mAdb Dataset Display

Group label	A	A	A	A	A
Sample name	BJAB_A_B	Daudi_A_B	Jurkat_A_B	Ly10_A_B	Ly3_A_B
genes				7.7702	
	9.7305	9.7985	9.7249	10.2981	10.1150
		8.9715			
		8.8918	9.0752	10.2200	
	8.4250	7.0224	7.8511	7.4692	7.7886
	6.9189	7.5645			7.7814
	9.3296	9.6202	9.4409	9.9652	10.0534
				7.8629	7.3505
	10.0053	9.6605	9.3872	9.9003	9.3181
	8.1908	8.2187	7.3540	8.3650	
	6.5014			7.0629	
		6.5251	6.4512		
	9.6604	10.0402	8.6991	9.9747	9.4539
	8.3781	8.8981	8.1739	8.2322	9.3807
	7.9419	7.4741	7.9301		
	8.9372	9.8243	9.4774	9.7465	10.2738
	8.2002			9.9105	9.6255
	5.0575	6.8163	5.9542		5.7388
	9.9564	9.8420	9.7677	10.1529	9.3419
	9.9284	9.6363	9.3726	9.8858	10.1808
	9.4419	9.0507	9.4075	9.9434	9.0739
	10.4035	9.7502	9.2389	10.1029	10.5434
	9.0906	9.3452	9.3869	9.6770	9.3613

Well ID	Feature ID	Gene	Description
1118566	117_at	HSPA6	heat shock 70kDa protein 6 (HSP70B')
1118567	121_at	PAX8	paired box gene 8
1118568	177_at	PLD1	phospholipase D1, phophatidylcholine-sp
1118569	179_at	PMS2L9	postmeiotic segregation increased 2-like
1118570	320_at	PEX6	peroxisomal biogenesis factor 6
1118572	564_at	GNA11	guanine nucleotide binding protein (G pr
1118573	632_at	GSK3A	glycogen synthase kinase 3 alpha
1118574	823_at	CX3CL1	chemokine (C-X3-C motif) ligand 1
1118575	1053_at	RFC2	replication factor C (activator 1) 2, 40kD
1118576	1294_at	UBE1L	ubiquitin-activating enzyme E1-like
1118577	1316_at	THRA	thyroid hormone receptor, alpha (erythro
1118579	1431_at	CYP2E1	cytochrome P450, family 2, subfamily E
1118581	1487_at	ESRRA	estrogen-related receptor alpha
1118582	1729_at	TRADD	TNFRSF1A-associated via death domain
1118584	1861_at	BAD	BCL2-antagonist of cell death
1118585	243_g_at	MAP4	microtubule-associated protein 4
1118586	266_s_at	CD24	CD24 antigen (small cell lung carcinoma
1118587	31799_at		Sapiens clone 24627 mRNA sequence
1118588	31807_at	DDX49	DEAD (Asp-Glu-Ala-Asp) box polypepti
1118589	31826_at	KIAA0674	KIAA0674 protein
1118591	31837_at	BC002942	hypothetical protein BC002942
1118592	31845_at	ELF4	E74-like factor 4 (ets domain transcripti
1118594	31861_at	IGHMBP2	immunoglobulin mu binding protein 2

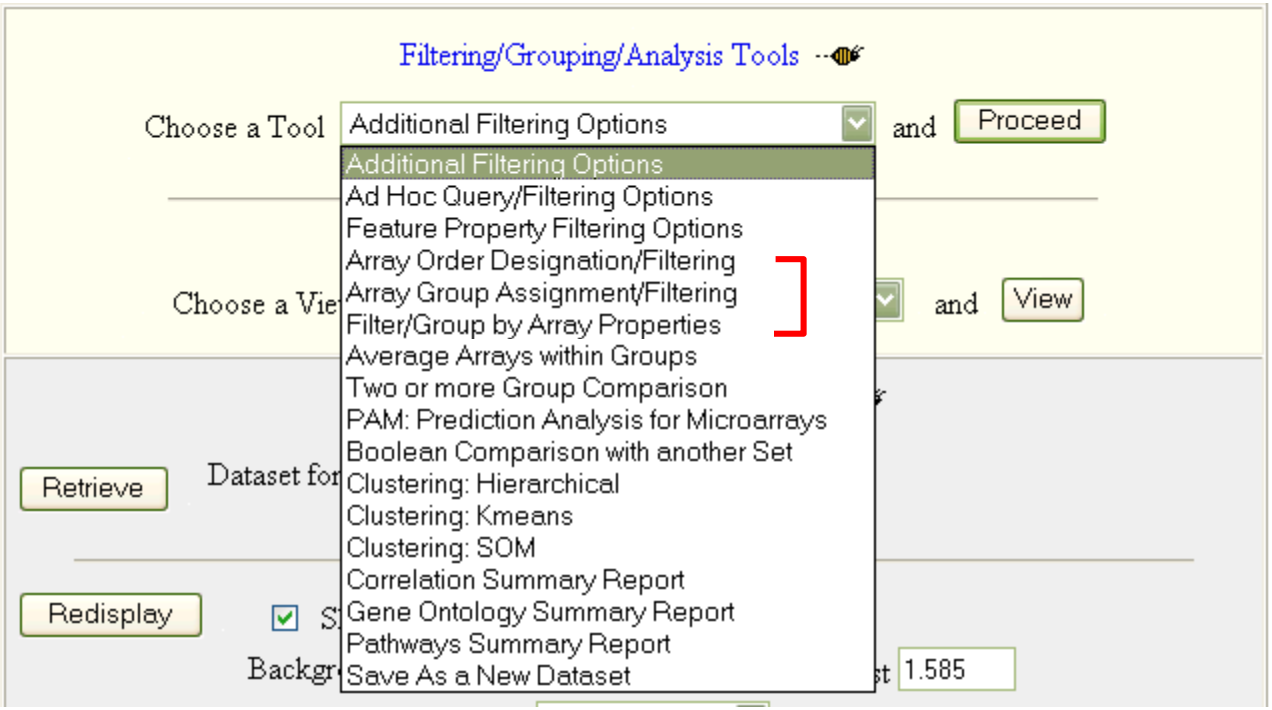
Group Examples

- Technical/Biological replicates
- Knock-outs and wild types
- Cancer vs normal samples
- Time course points
- Dosage levels

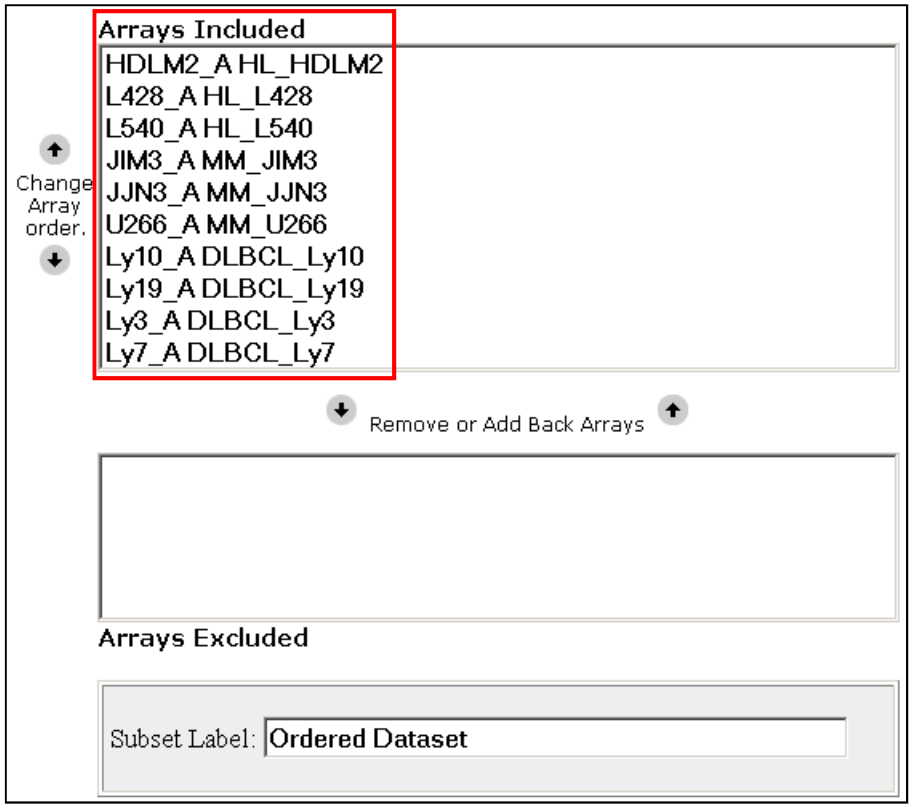
Dataset Group Assignment

- Array Order Designation/Filtering
- Array Group Assignment/Filtering
- Filter/Group by Array Properties

Dataset group assignment tools



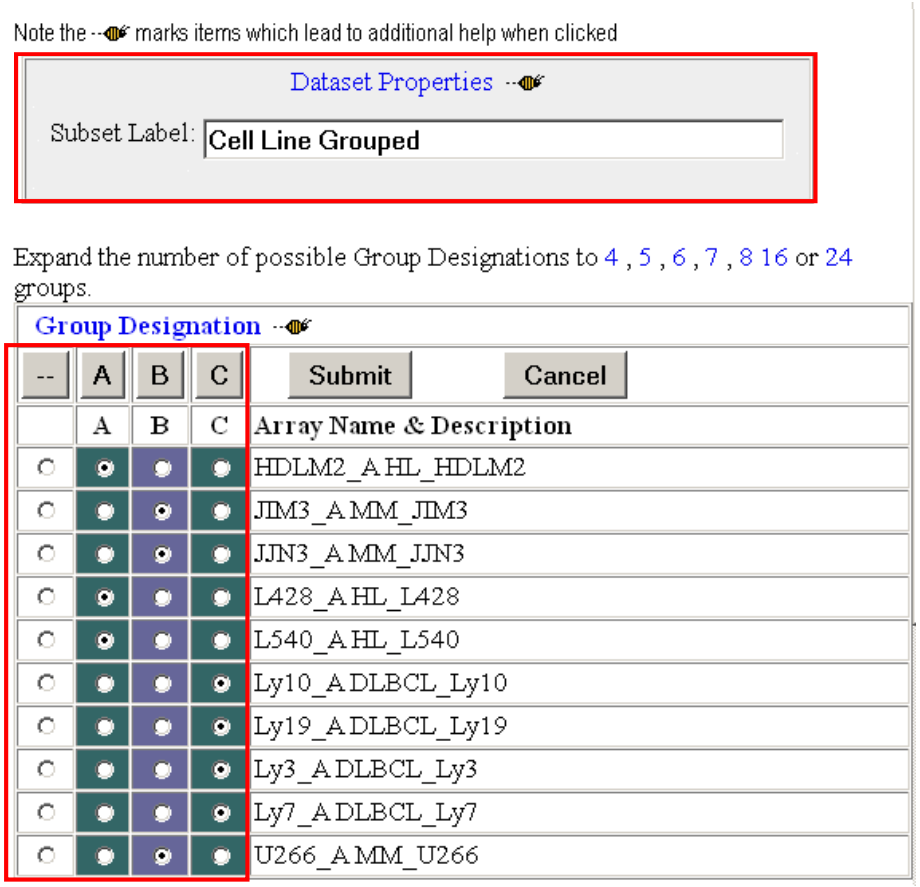
Array Order Designation/Filtering



- Order arrays in dataset
- Delete/Add back arrays in dataset
- Subsequent analysis will be ordered by groups first and then ordered within each group

- Does not group arrays

Array Group Assignment/Filtering



- One click per array for additional group
- Not convenient for large dataset
- Can not order within group

Filter/Group by Array Properties

mAdb Dataset Display

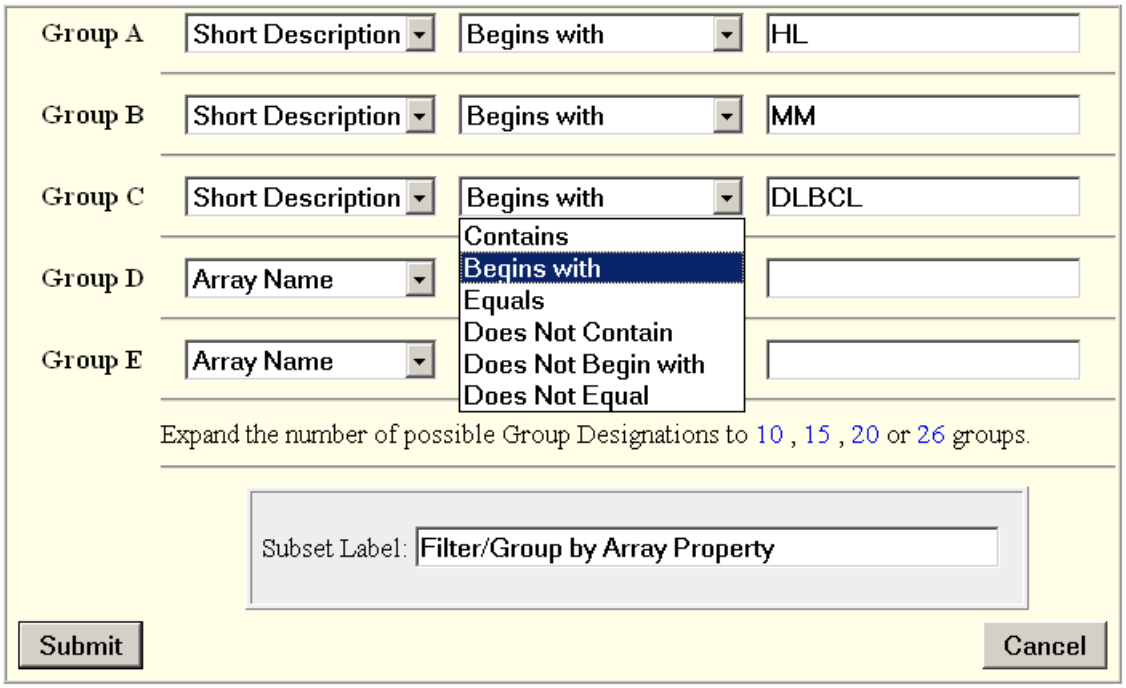
A	0.008	1.	HDLM2_A	HL_HDLM2
A	0.007	2.	JIM3_A	MM_JIM3
A	0.007	3.	JJN3_A	MM_JJN3
A	0.006	4.	L428_A	HL_L428
A	0.009	5.	L540_A	HL_L540
A	0.006	6.	Ly10_A	DLBCL_Ly10
A	0.007	7.	Ly19_A	DLBCL_Ly19
A	0.007	8.	Ly3_A	DLBCL_Ly3
A	0.007	9.	Ly7_A	DLBCL_Ly7
A	0.007	10.	U266_A	MM_U266

Edit Data for Dataset: **Cell Lines representing 3 Lymphomas**

```
10 Arrays and 22283 Expression Rows extracted.
Data transformation method: Centered to Signal Median
Spot Filter Options:
Signals are floored at 100.0
```

- Array properties include Name and Short Description
- Identify consistent pattern

Filter/Group by Array Properties



- Convenient for large dataset
- Can not order arrays within group

Group Assignment

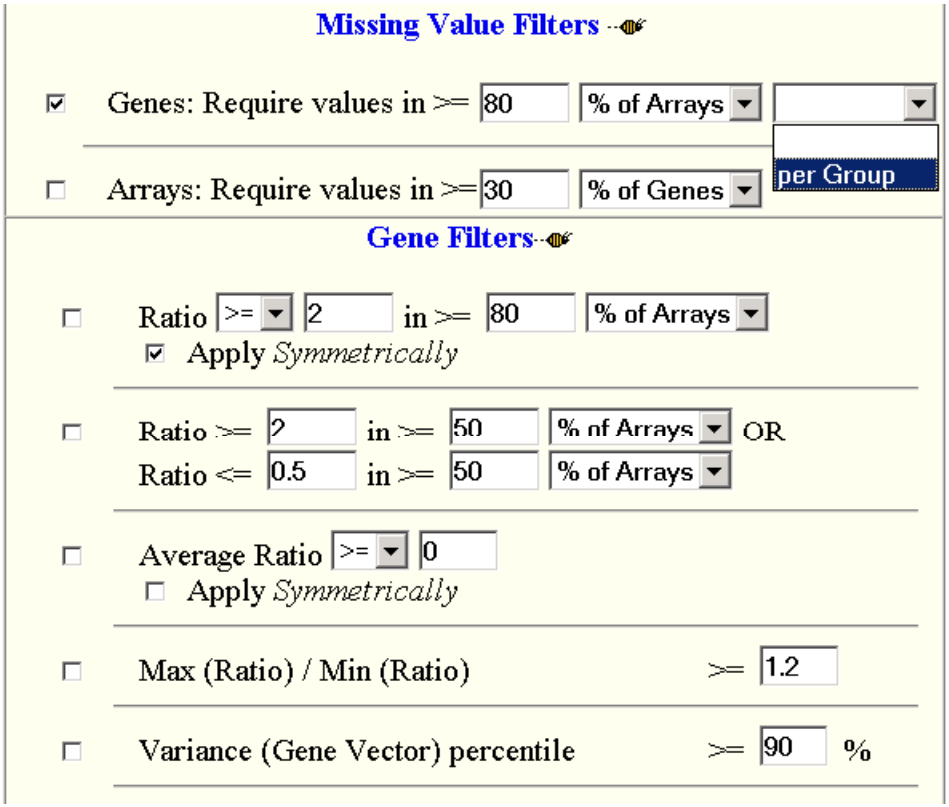
A	A	A	B	B	B	C	C	C	C			
HDLM2_A	L428_A	L540_A	JIM3_A	JJN3_A	U266_A	Ly3_A	Ly7_A	Ly10_A	Ly19_A	Well ID	Feature ID	Gene
0.8986	1.5182	1.1664	1.1075	0.8887	0.9967	0.6761	0.8685	1.3198	1.2333	1118566	117_at	HSPA6
8.1537	6.8697	9.1886	6.7782	8.5125	5.8007	7.4983	8.7316	7.6118	9.1357	1118567	121_at	PAX8
0.8042	2.2147	0.8831	0.6680	0.6954	1.4118	0.6761	0.6743	0.6046	0.7337	1118568	177_at	PLD1
4.1856	6.4728	9.8080	5.3601	6.0779	5.1954	7.1981	3.7505	7.2110	4.8481	1118569	179_at	PMS2L9
2.3557	1.6427	1.2628	2.5865	2.4068	2.0954	1.4949	2.1160	1.0713	2.5561	1118570	320_at	PEX6
1.1856	1.3852	0.9514	0.9599	0.9757	0.8588	1.2529	1.4626	1.3452	1.2318	1118571	336_at	TBXA2R
3.7746	1.6271	2.5043	1.1516	1.0508	0.6536	1.4875	1.9670	1.1227	1.1988	1118572	564_at	GNA11
4.5008	5.1783	5.5333	5.3079	7.4172	6.8863	7.1846	5.8658	6.0435	8.4519	1118573	632_at	GSK3A
4.1646	12.1329	0.8532	0.6680	0.6954	0.6536	1.1034	0.6743	1.4075	0.7337	1118574	823_at	CX3CL1
5.5663	4.3223	5.4480	1.6206	2.9270	4.4418	4.3158	3.3790	5.7775	3.3067	1118575	1053_at	RFC2
3.9173	2.4157	2.0461	1.3460	0.9437	1.1039	1.3083	2.0964	1.9933	1.9391	1118576	1294_at	UBE1L
0.7800	0.7918	0.8532	0.7715	0.6954	0.8327	0.6761	0.8483	0.8083	0.7630	1118577	1316_at	THRA
0.7800	0.6485	0.8532	0.6680	0.6954	0.6536	0.6761	0.6743	0.6046	0.7337	1118578	1320_at	PTPN21

- Group assignment information is carried into relevant analysis
- Dataset is independent from microarray platforms

Examples for using groups

- Additional Filtering per Group
- Correlation summary report
- Average arrays within groups
- Calculate statistics within groups

Filter by Group Properties



Ensures each group has sufficient number of non-missing values

Correlation Summary Report

Correlations

A	A	A	B	B	B	C	C	C	C				
#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Grp		Array Name	Array Description
#1 A	0.890	0.914	0.844	0.873	0.852	0.853	0.838	0.856	0.836	A	1.	HDLM2_A	HL_HDLM2
	#2 A	0.882	0.852	0.860	0.847	0.856	0.824	0.869	0.845	A	2.	L428_A	HL_L428
		#3 A	0.860	0.880	0.855	0.858	0.850	0.859	0.843	A	3.	L540_A	HL_L540
			#4 B	0.896	0.895	0.852	0.826	0.850	0.846	B	4.	JIM3_A	MM_JIM3
				#5 B	0.885	0.868	0.853	0.859	0.867	B	5.	JJN3_A	MM_JJN3
					#6 B	0.857	0.832	0.852	0.848	B	6.	U266_A	MM_U266
						#7 C	0.871	0.924	0.882	C	7.	Ly10_A	DLBCL_Ly10
							#8 C	0.873	0.918	C	8.	Ly19_A	DLBCL_Ly19
								#9 C	0.883	C	9.	Ly3_A	DLBCL_Ly3
									#10 C	C	10.	Ly7_A	DLBCL_Ly7

- Pair wise correlation between 2 samples in dataset
- Individual scatter plot available
- Group pattern for quality control

Visual Bivariate Data Analysis

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://madb.nci.nih.gov/cgi-bin/restricted/mAdb_correlation_report.cgi Go Links

Home Page | mAdb Gateway | Upload Status
Forums | Reference Info | Program Downloads | GeneCards

mAdb Correlation Report

View Array Summaries

Return to the input dataset.

Redisplay Background Color Scheme Green/White/Red

Color Saturation Max/Mid/Min 0.8 0.6 0.4

Note: For proper coloring Max > Mid > Min

Note: Click on the Correlation values to display the corresponding ScatterPlot

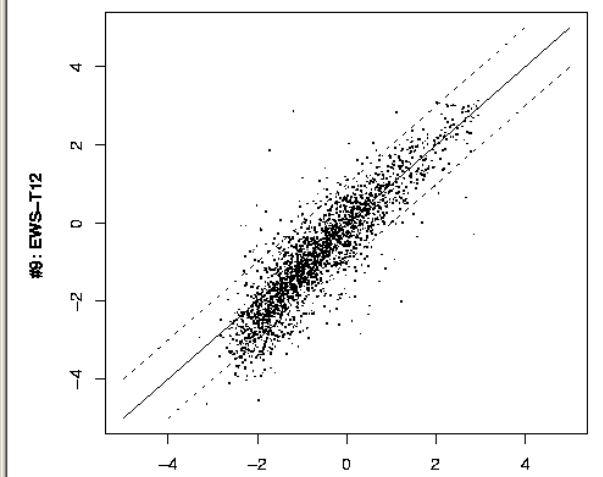
Correlations

	A #1	A #2	A #3	A #4	A #5	A #6	A #7	A #8	A #9	A #10	A #11	A #12	#	A #21
1.	#1 A	0.809	0.815	0.596	0.618	0.711	0.653	0.723	0.743	0.506	0.682	0.693	0.	.732
2.		#2 A	0.820	0.583	0.528	0.602	0.603	0.645	0.659	0.395	0.549	0.597	0.	.767
3.			#3 A	0.615	0.626	0.689	0.651	0.719	0.699	0.562	0.652	0.659	0.	.673
4.				#4 A	0.497	0.468	0.442	0.513	0.507	0.385	0.432	0.518	0	.517
5.					#5 A	0.784	0.665	0.711	0.731	0.681	0.733	0.726	0.	.492
6.						#6 A	0.727	0.804	0.802	0.659	0.831	0.807	0.730	0.523	0.548	0.501	0.609	0.635	0.641	0.629	0.551
7.							#7 A	0.756	0.751	0.639	0.764	0.799	0.690	0.632	0.560	0.516	0.511	0.636	0.646	0.643	0.523
8.								#8 A	0.888	0.638	0.803	0.808	0.787	0.563	0.588	0.546	0.561	0.633	0.643	0.615	0.573
9.									#9 A	0.623	0.815	0.828	0.725	0.604	0.639	0.623	0.591	0.664	0.672	0.660	0.629
10.										#10 A	0.712	0.679	0.608	0.444	0.330	0.261	0.344	0.409	0.423	0.453	0.290
11.											#11 A	0.881	0.761	0.591	0.519	0.475	0.535	0.646	0.656	0.694	0.496
12.												#12 A	0.711	0.639	0.562	0.533	0.534	0.668	0.676	0.698	0.538
13.													#13 A	0.476	0.466	0.413	0.570	0.588	0.598	0.580	0.529
14.														#14 A	0.552	0.530	0.418	0.799	0.807	0.777	0.497
15.															#15 A	0.891	0.625	0.663	0.668	0.641	0.796
16.																#16 A	0.641	0.642	0.645	0.606	0.821

mAdb Correlation Report - Microsoft Internet Explorer

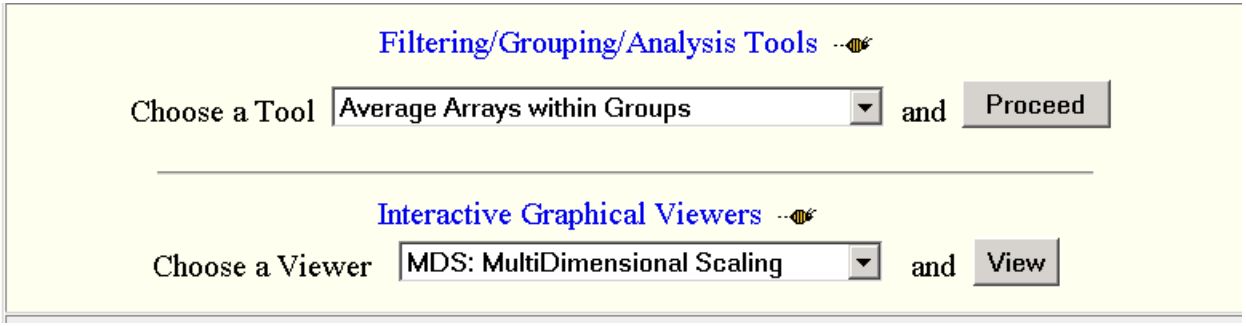
File Edit View Favorites Tools Help

Back Search Favorites Media



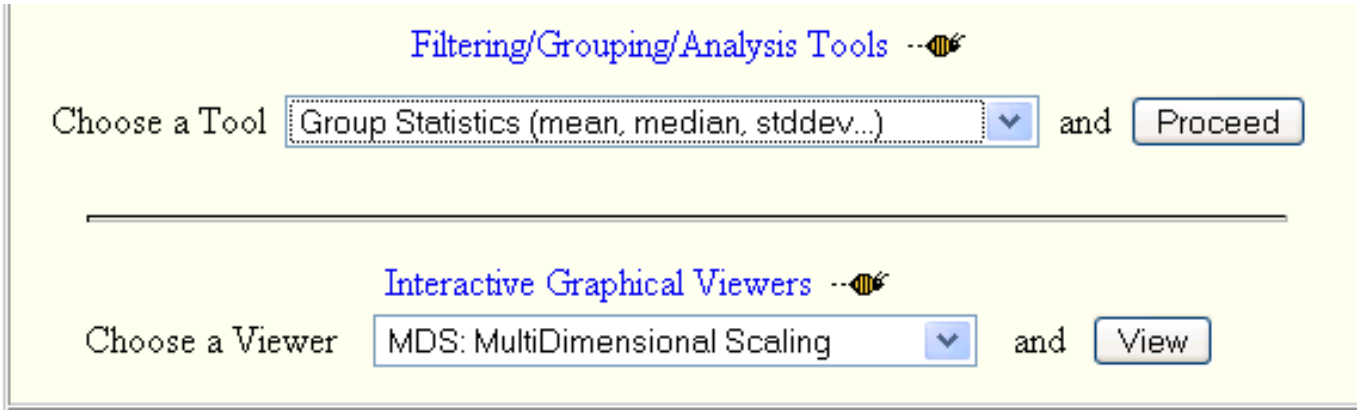
http://madb.nci.nih.gov/cgi-bin/restricted/mAdb_correlation_report.cgi#main_options_table Internet

Average Arrays within Groups



- Averages calculated using log ratios regardless of linear or log display options chosen

Calculate statistics within Groups



- All values calculated using log ratios regardless of linear or log display options chosen

Dataset I
Small Round Blue Cell Tumors
(SRBCTs)

- Khan et al. *Nature Medicine* 2001
- 4 tumor classifications
- 63 training samples, 25 testing samples, 2308 genes
- Neural network approach

Hands-on Session 1

- Lab 1- Lab 4
- Read the questions before starting, then answer them in the lab.
- Use web site: http://madb-training.cit.nih.gov
- Avoid maximizing web browser to full screen.
- Total time: 20 minutes

3. mAdb dataset analysis tools

- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Analysis Overview

Class Discovery - Unsupervised	• Clustering – Hierarchical, K-means, SOMs • Principal components Analysis (PCA) • Multidimensional Scaling (MDS)
Class Comparison - Supervised	• paired t-tests • t-test pooled (equal) variance • t-test separate (unequal) variance • Significance Analysis of Microarrays (SAM) • One way ANOVA • Wilcoxon Rank-Sum (Mann Whitney U) • Wilcoxon Matched-pairs Signed Rank • Kruskal-Wallis
Class Prediction - Supervised	Prediction Analysis for Microarrays (PAM)

Class Discovery Example

- Discover cancer subtypes by gene expression profiles
- Identify genes which have different expression patterns in different groups

- Tools: Cluster Analysis, PCA and MDS

Class Comparisons Example

- Find genes that are differentially expressed among cancer groups
- Find genes up/down regulated by drug treatment

- Tools:
  - Group comparison
  - Statistics Results filtering

Class Prediction Example

- Identify an expression profile which correlates with survival in certain cancers
- Identify an expression profile which can be used to diagnose different types of lymphomas

- Tools: Prediction Analysis for Microarrays (PAM)

3. mAdb dataset analysis tools

- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Class Discovery

- Dataset with large amount of data
- Dataset not organized
- Visualization with Clustering, PCA, MDS

Cluster Analysis

- Organize large microarray dataset into meaningful structures
- Visualize and extract expression patterns

What to Cluster?

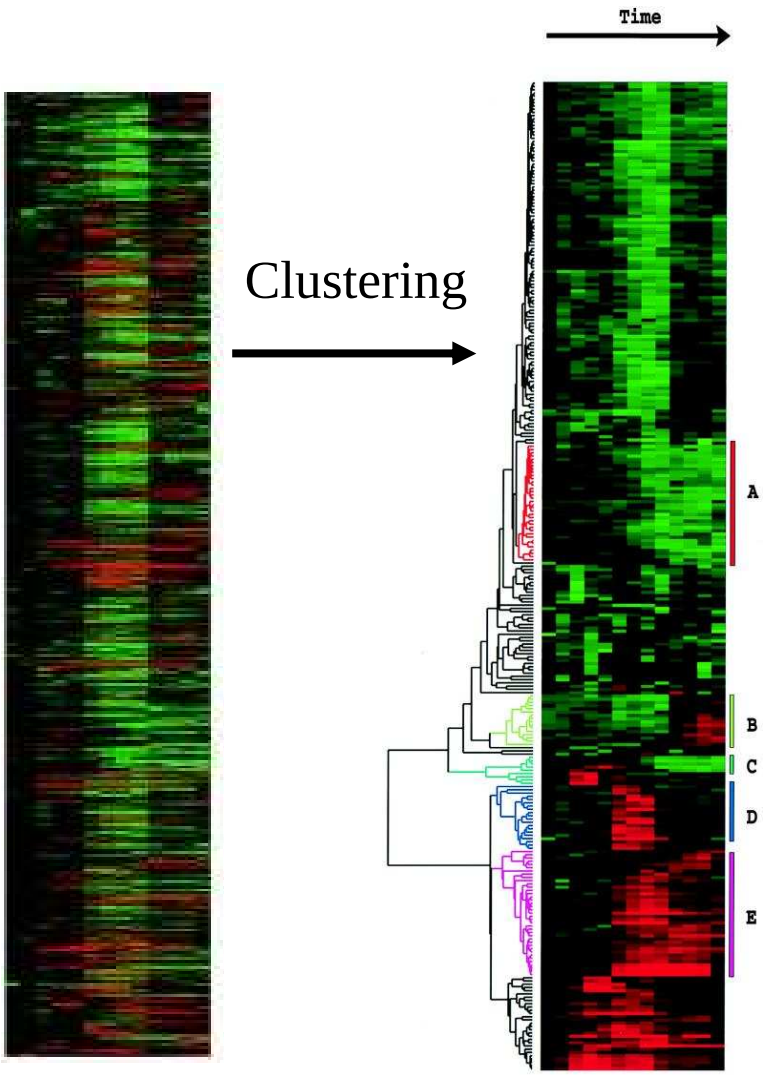
Genes - identify groups of genes that have correlated expression profiles

Samples - put samples into groups with similar overall gene expression profiles

Clustering Methods

- Hierarchical clustering
- Partitional clustering
  - K-means
  - Self-Organizing Maps (SOM)

Cluster Example on Genes



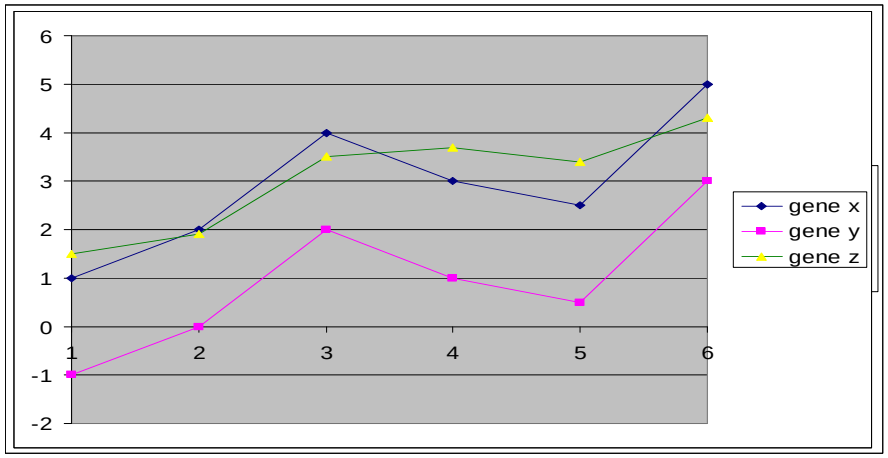
Much easier to look at large blocks of similarly expressed genes

Dendogram helps show how ‘closely related’ expression patterns are

A. Cholesterol syn.
B. Cell cycle
C. Immediate-early response
D. Signaling
E. Tissue remodeling

2 Steps

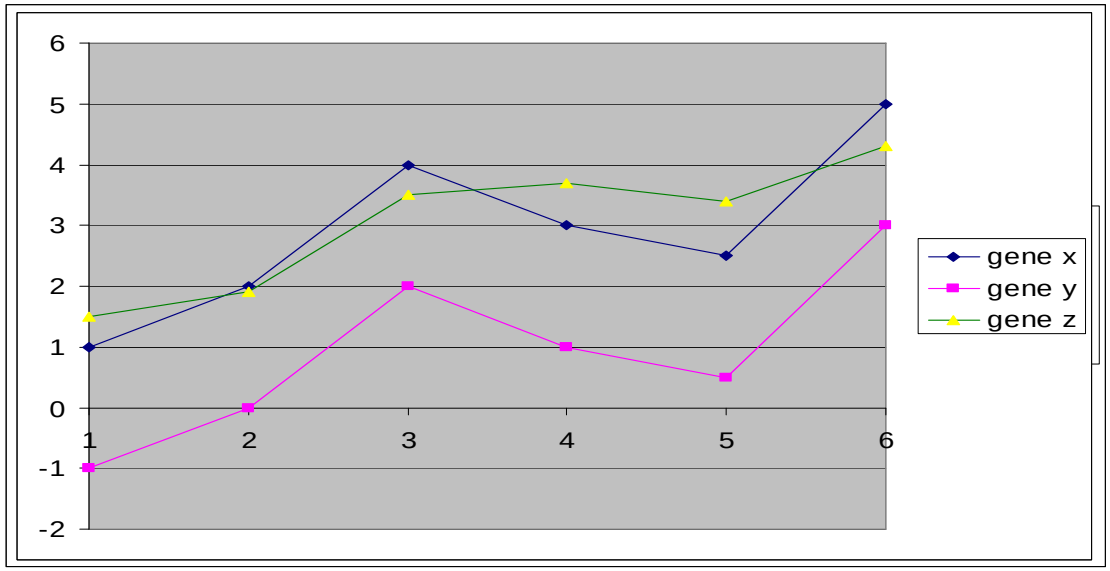
– Pick a distance method
  - Correlation
  - Euclidian



– Pick the linkage method
  - Average linkage
  - Complete linkage
  - Single linkage

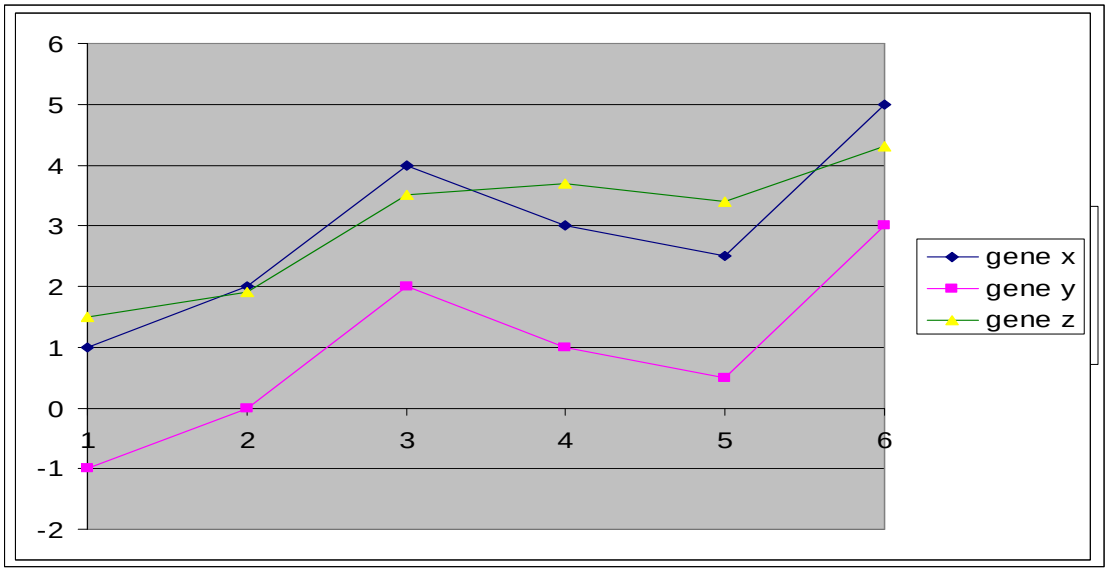
Correlation

- Compares shape of expression curves (-1 to 1)
- Can detect inverse relationships (absolute correlation)

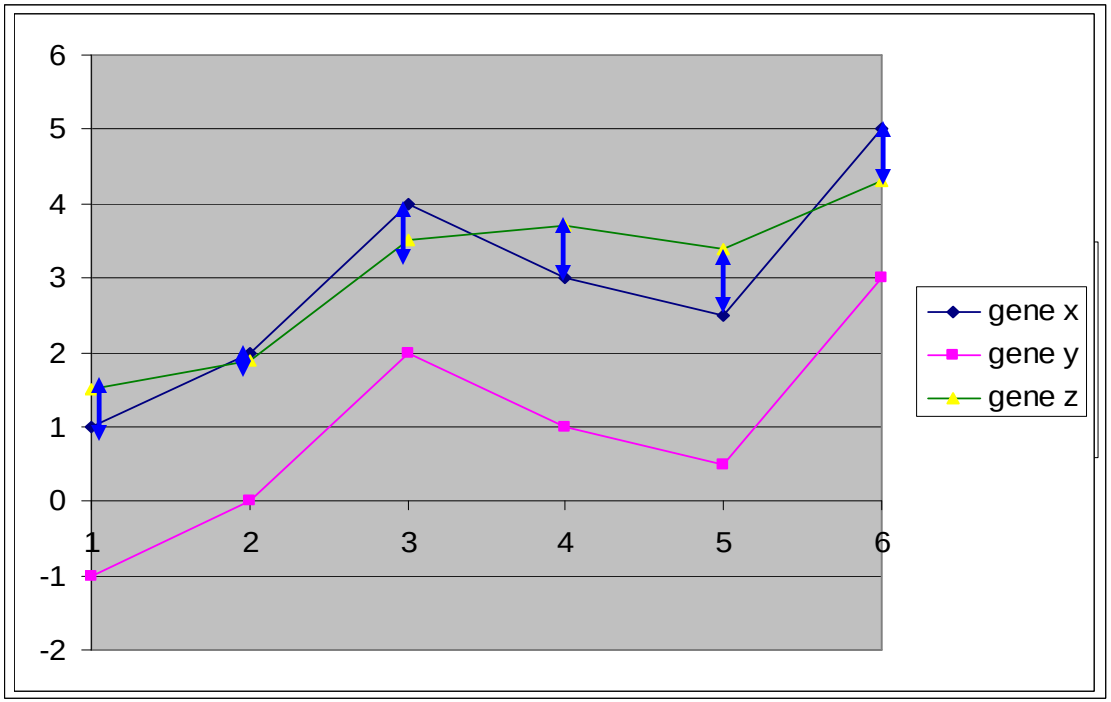


Two Flavors of correlation

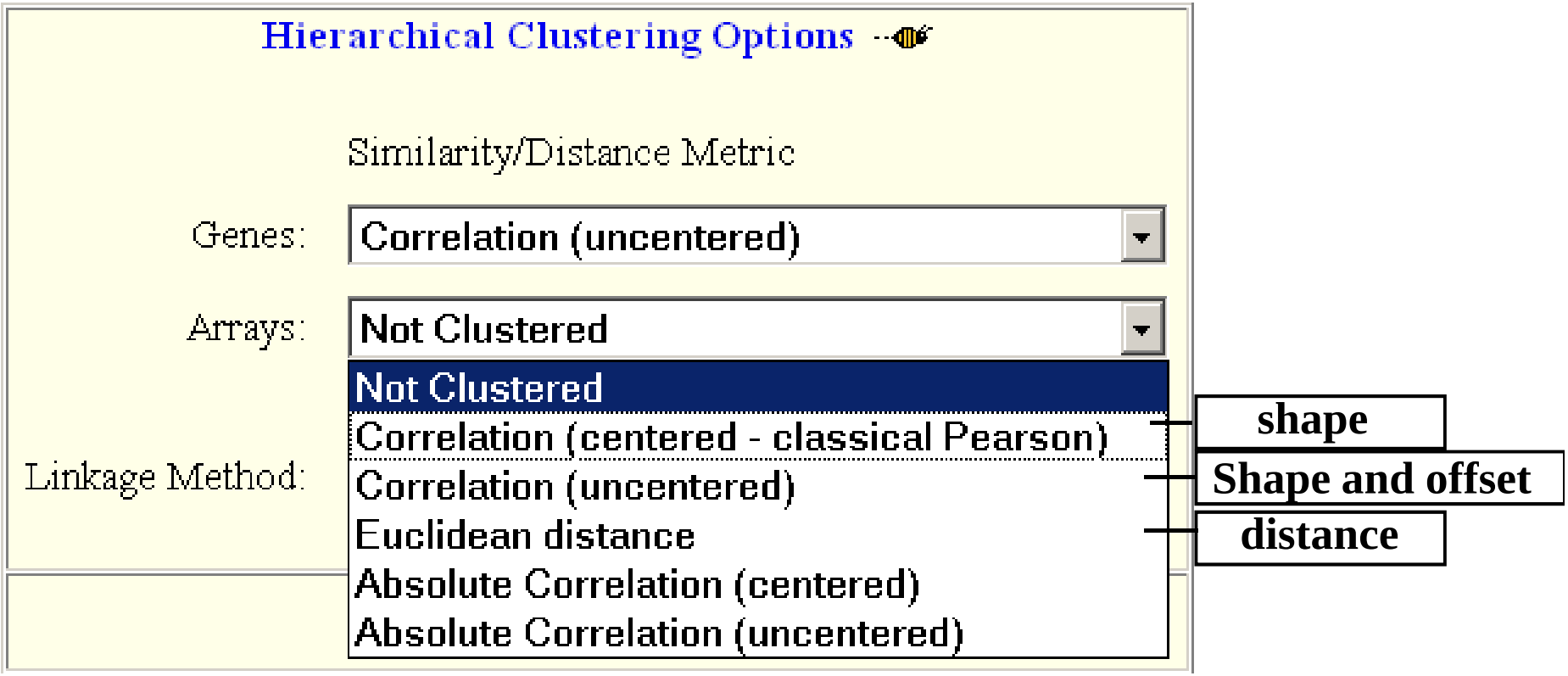
- Correlation (centered-classical Pearson)
- Correlation (un-centered)
  - assume the mean of the data is 0, penalize if not
  - Measures both similarity of shape and the offset from 0



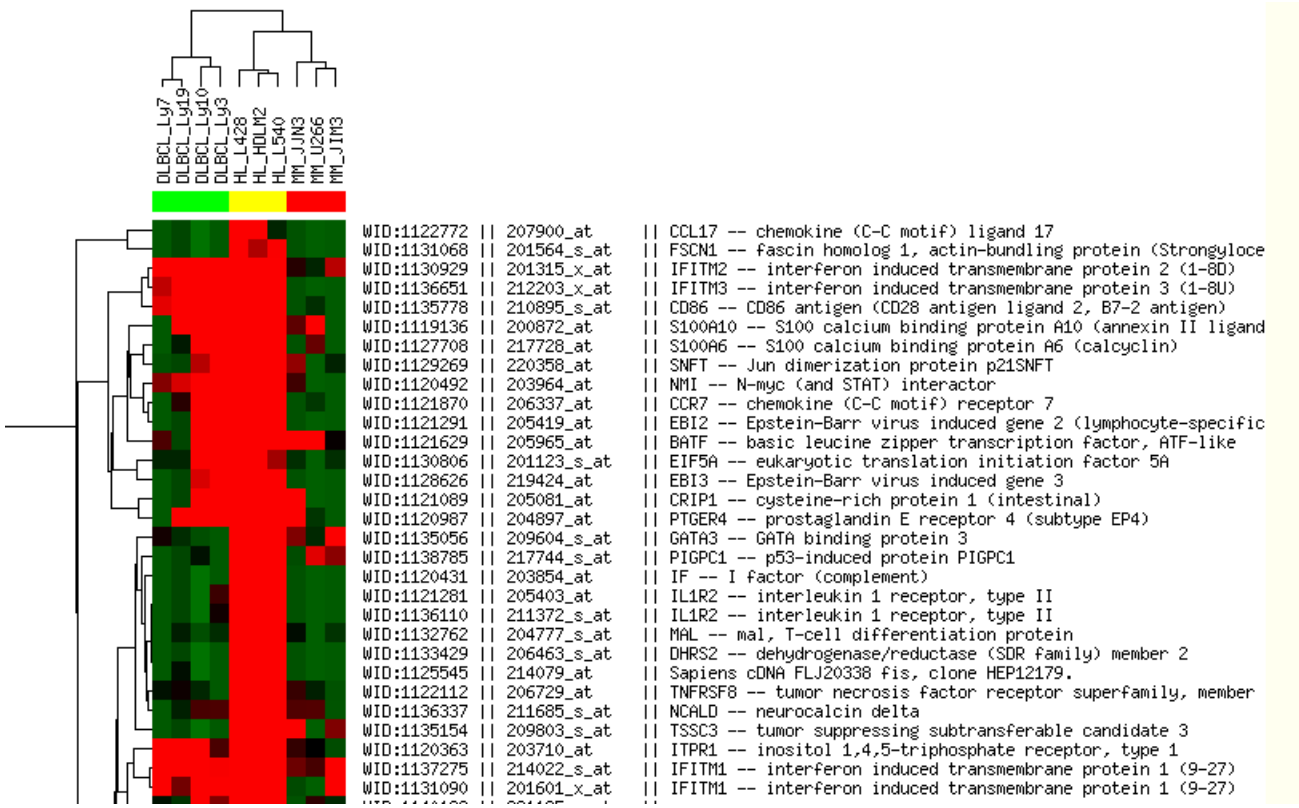
Euclidean Distance



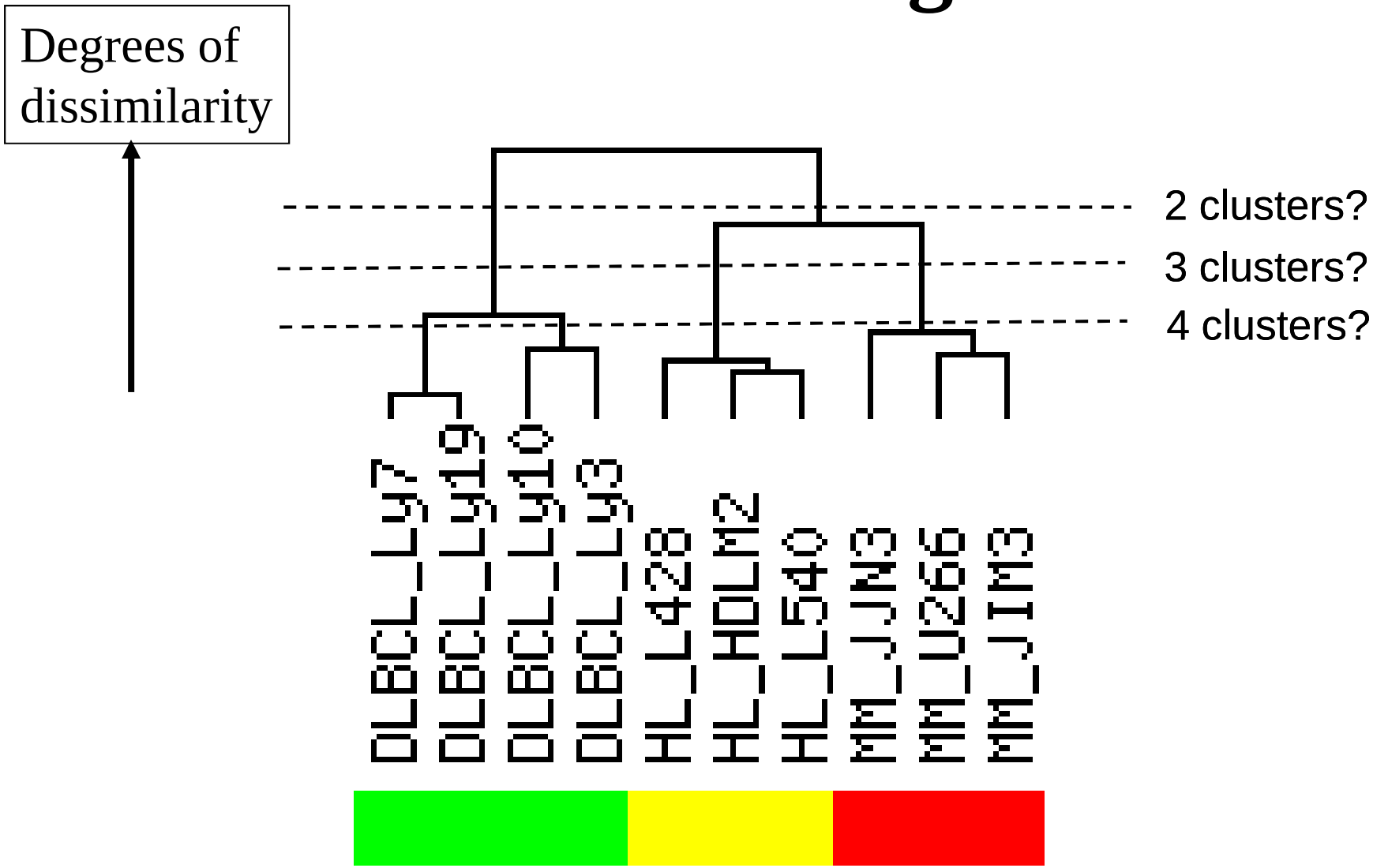
Similarity/Distance Metric Summary



Hierarchical Clustering Example



Tree Cutting



Hierarchical Clustering Summary

- Detection of patterns for both genes and samples
- Good visualization with tree graphs

- Dataset size limitations
- No partition in results, require tree cutting

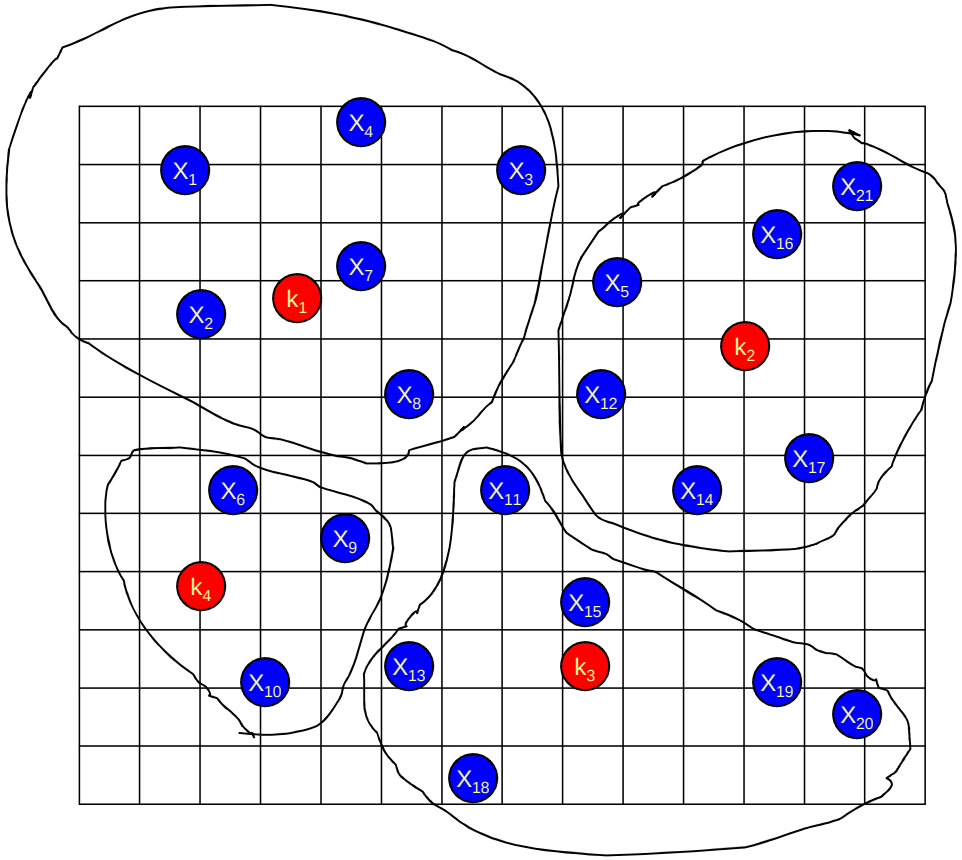
Partitional clustering : K-means

- Partition data into K clusters, with number K supplied by user.
- Produce cluster membership as results.

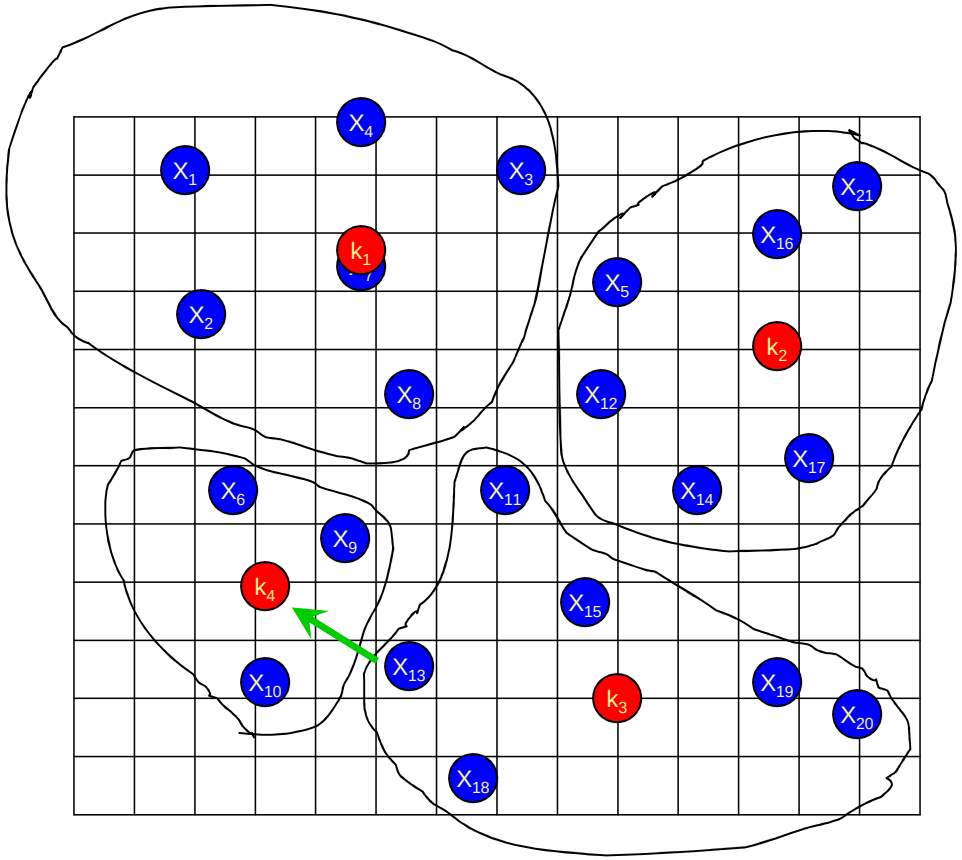
K-means Algorithm

- Divide observations into K clusters.
- Use cluster averages (means) to represent clusters
- Maximize the inter-cluster distance
  Minimize intra-cluster distance.

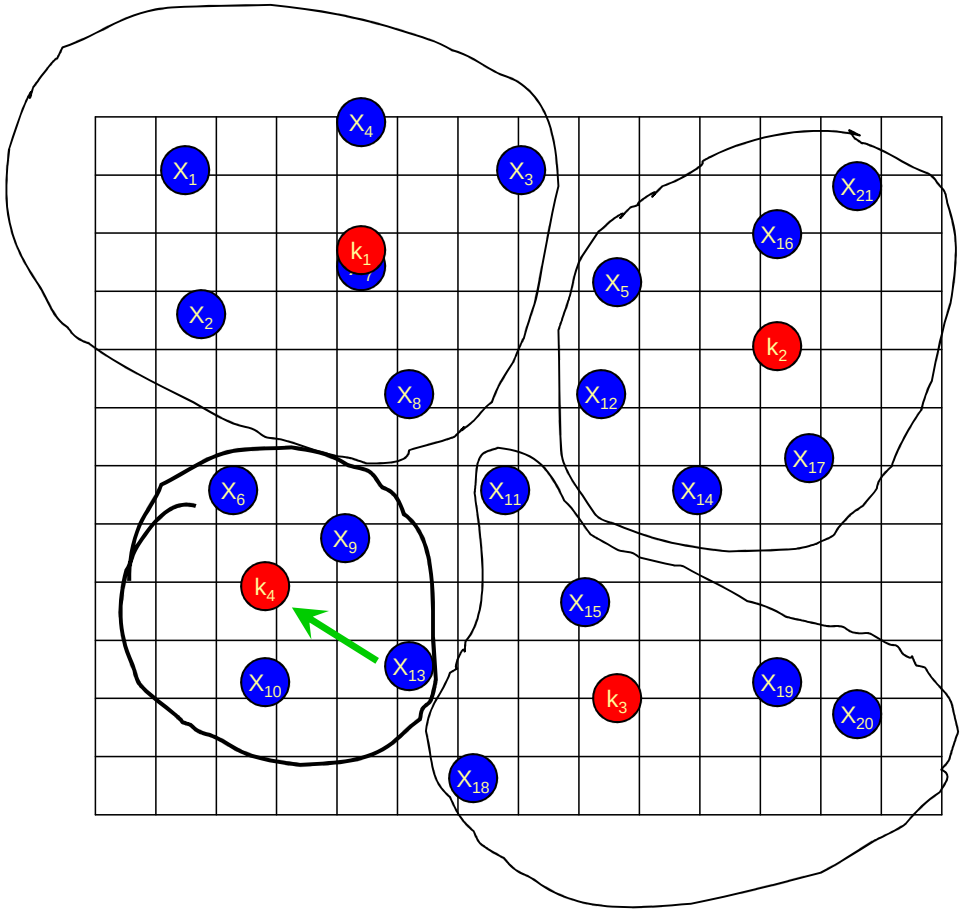
K-means Algorithm



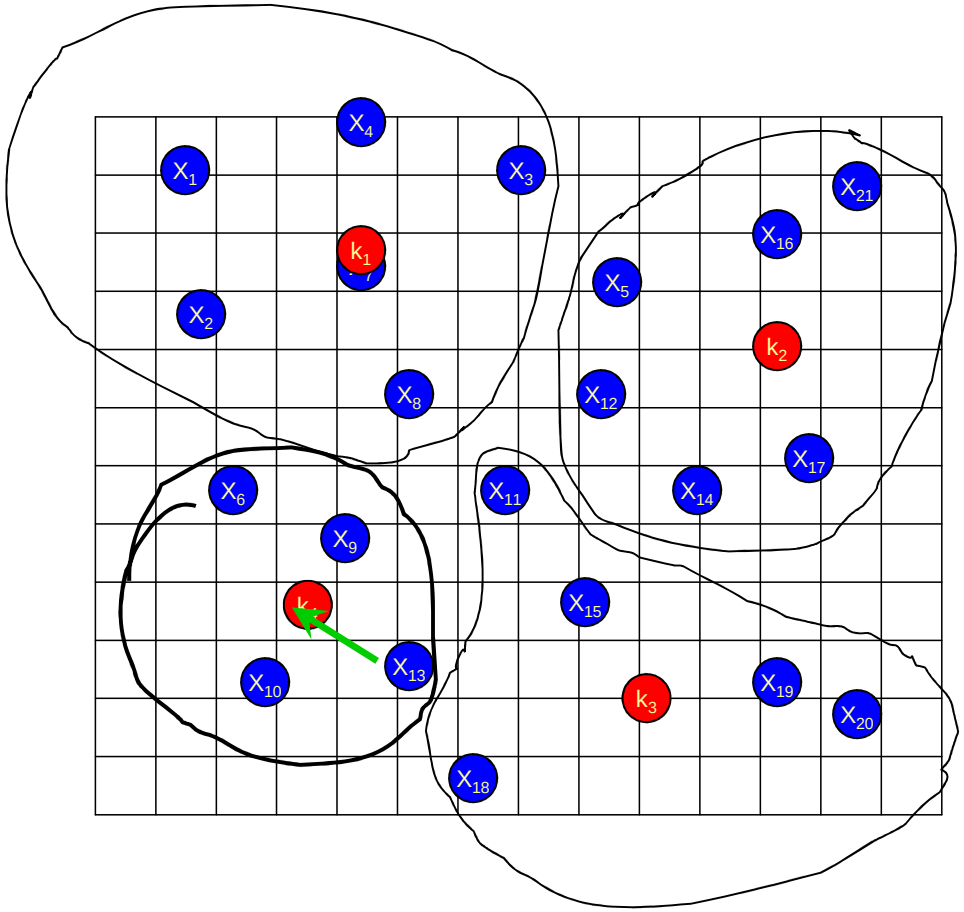
K-means Algorithm



K-means Algorithm



K-means Algorithm



mAdb K-means Options

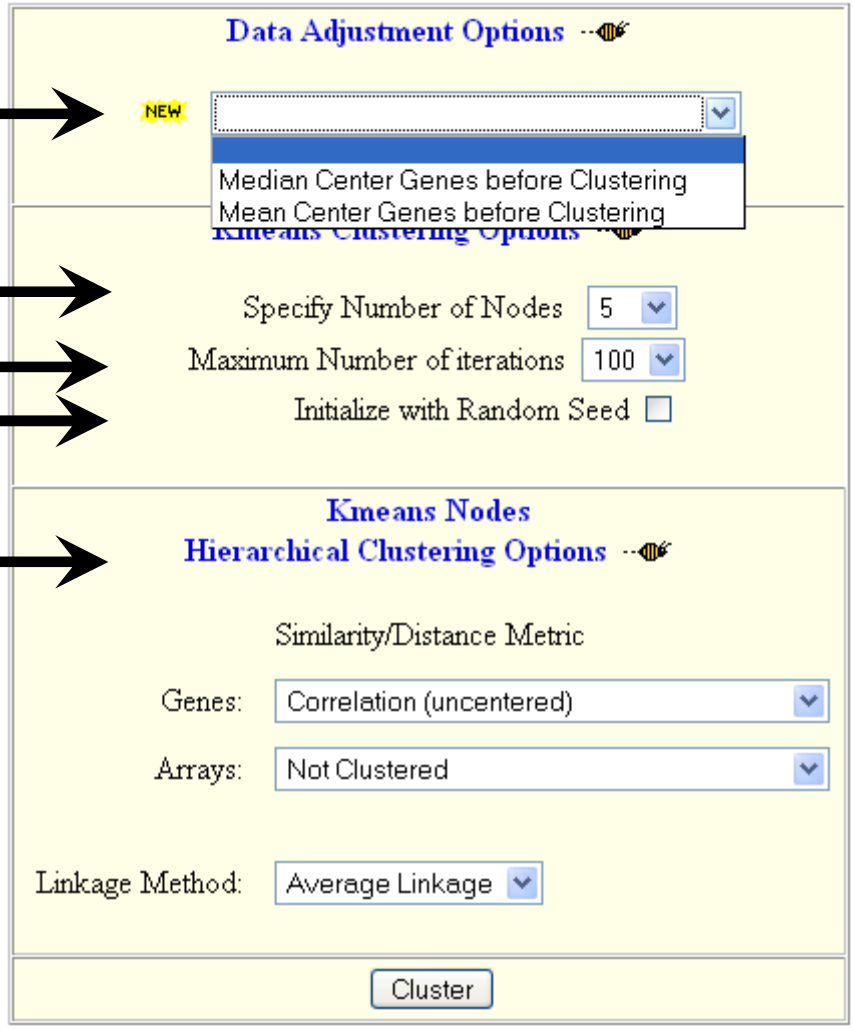
Adjust data for analysis

Set number of clusters

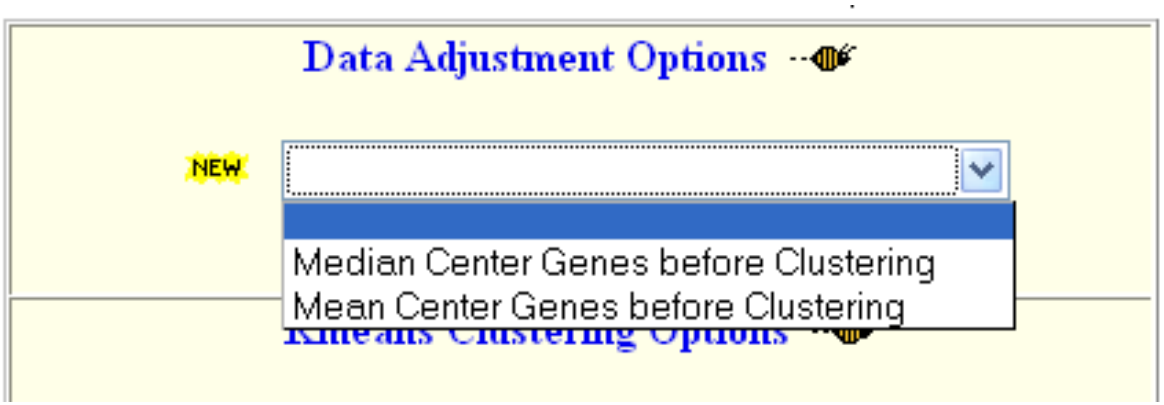
Set number of iteration

Activate random seed

Hierarchical clustering within node



Data Adjustment Options



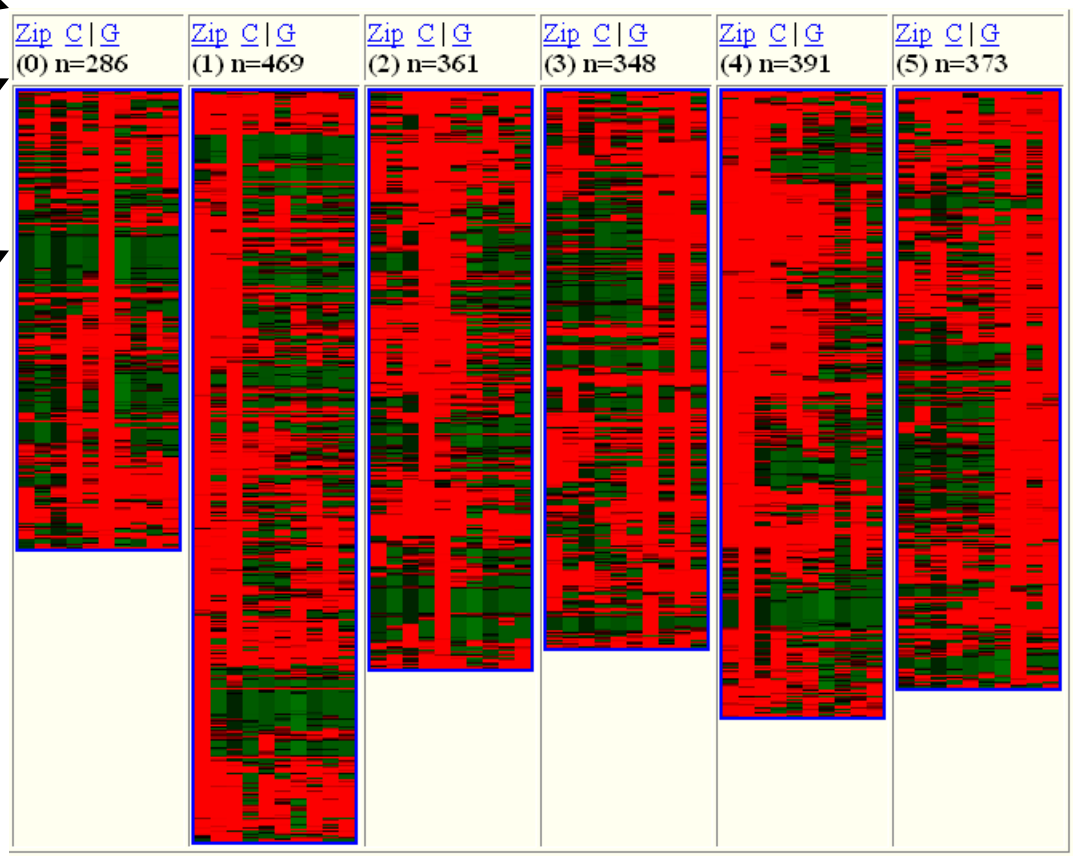
- Adjusts data rows so median/mean will be zero
- Used only for analysis – not saved in dataset
- Center genes to compare relative values among genes
- Not appropriate if clustering arrays
- Not appropriate if using Euclidean distance/similarity metric

K-means Clustering Example

Save as input to TreeView

Create new subset of genes

Show hierarchical clustering



Summary

- Fast algorithm
- Partitions features into smaller, manageable groups
- mAdb allows hierarchical clustering within each K-mean cluster

- Must supply reasonable number of K
- No relationship among partitions

Self-Organizing Maps (SOM)

- Partitions data into 2 dimensional grid of nodes
- Clusters on the grid have topological relationships

- 2 numbers for the dimension of grid supplied by user

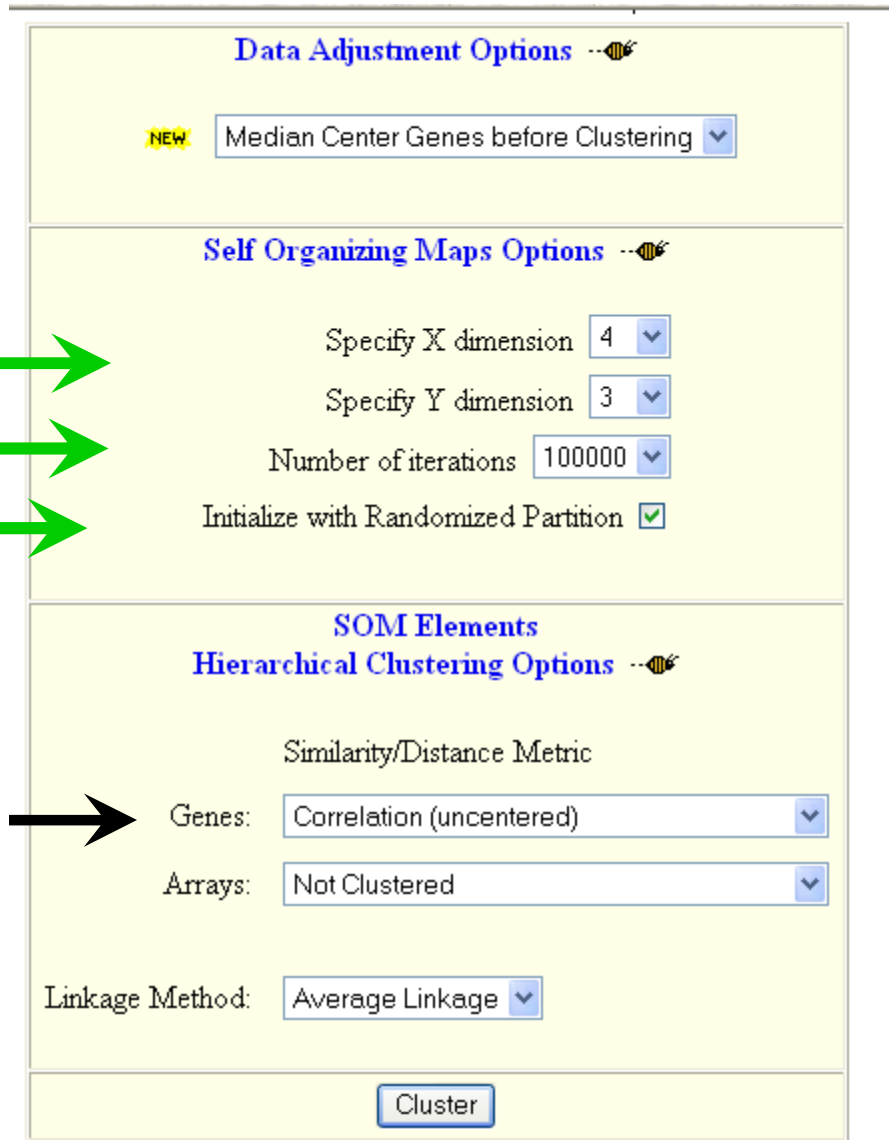
mAdb SOM options

Set number of clusters (X, Y)

Set number of iteration

Activate Randomized Partition

Hierarchical within SOM clusters

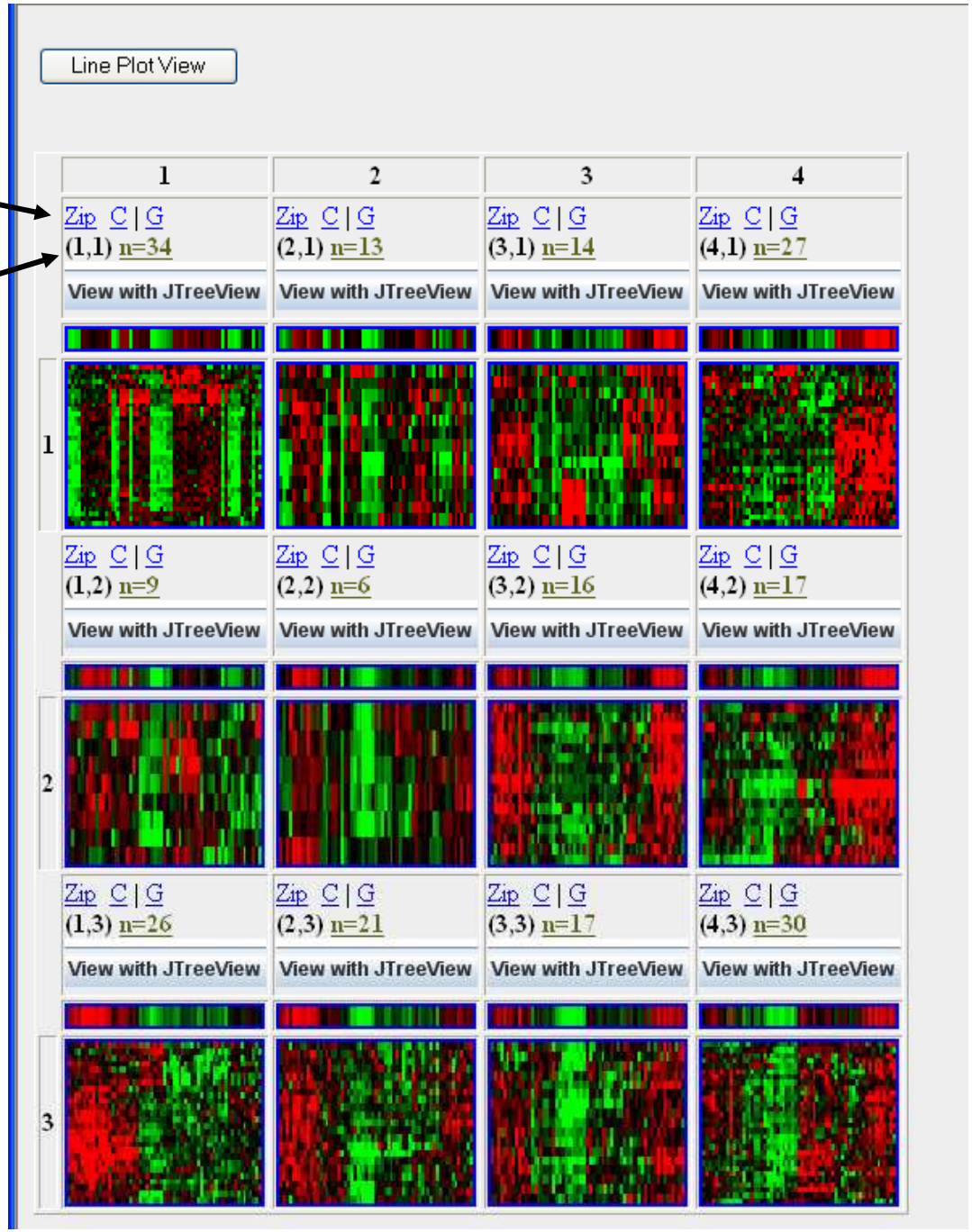


Heat map View

Save as input to TreeView

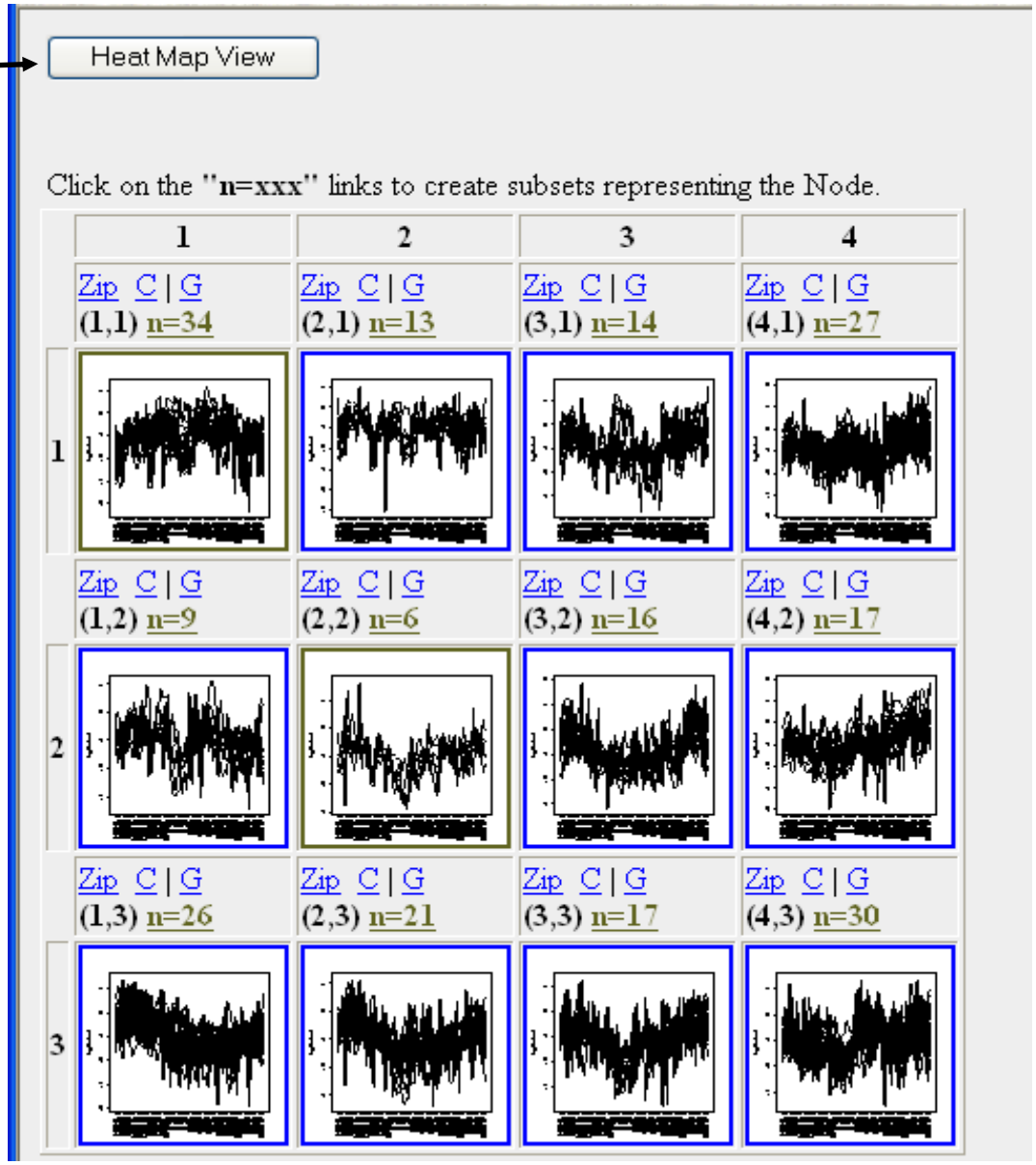
Create new subset of genes

Show hierarchical clustering



Line Plot View

Toggle back to Heat Map View



SOM Summary

- Neighboring partitions similar to each other
- Partitions features into smaller groups
- mAdb allows hierarchical clustering within each SOM cluster

- Results may depend on initial partitions

Summary of mAdb Clustering Tools

	Hierarchical	K-means	SOM
Relationship visualization	Tree Structure	partition Membership	Partition 2-D topology
Data Size	Small	Large	Large
Performance	Slow	Fast	Middle
Cluster Type	Gene/Array	Gene	Gene

Cluster Analysis

- Normalization is important
- Reduce data points by variance
- Use K-mean or SOM to partition dataset
- Use biological information to interpret results

Hands-on Session 2

- Lab 5 - lab 6 (Lab 7 optional)
- Total time: 15 minutes

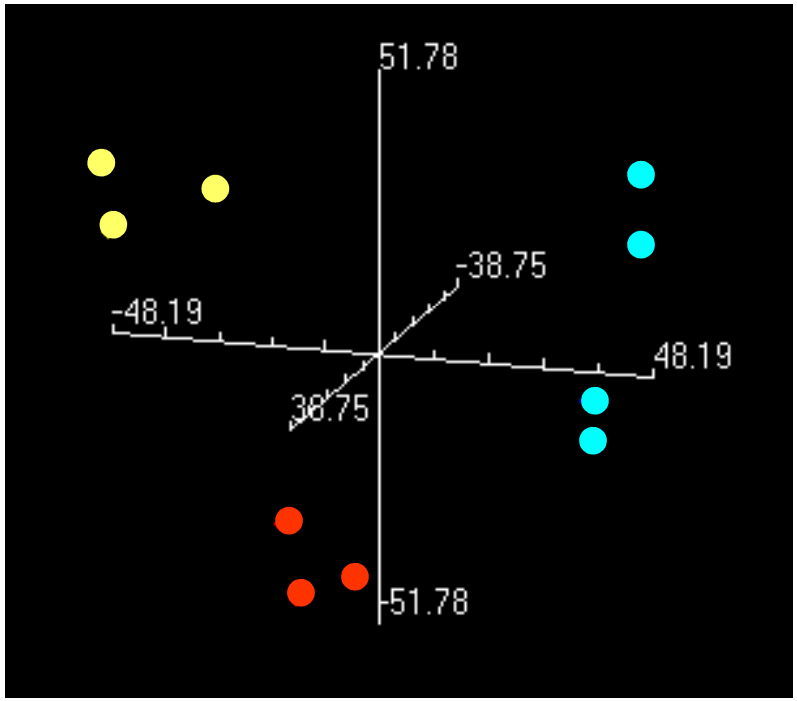
Principal Component Analysis

- How different samples are from each other
- Project high-dimensional data into lower dimensions, which captures most of the variance
- Display data in 2D or 3D plot to reveal the data pattern

Principal Component Analysis

- Hypothesis - there exist unobservable or "*hidden*" variables (complex traits) which have given rise to the ***correlation*** among the observed objects (genes or microarrays or patients)
- The Principal Components (PC) Model is a straightforward model that seeks to achieve this objective

PCA 3D plot



- Axes represent the first 3 components
- The first 3 components should explain most of the variance
- Formation of clusters
- Relationship of clusters.

Basic Idea of PCA is a Data Reduction Method Based on Analysis of Correlation Pattern(s) That Can Exist Among the Observed Random Variables (i.e. Expression values of Genes).

Raw Data

Array	1	2	…	m
Gene 1	a_{11}	a_{12}	…	a_{1m}
Gene 2	a_{21}	a_{22}	…	a_{2m}
Gene ...	⋮	⋮	⋮	⋮
Gene n	a_{n1}	a_{n2}	…	a_{nm}

n is the number of genes (gene probes); m is the number of arrays (experiments)

A Structure of Correlation Matrix is the **Major Object for PCA**

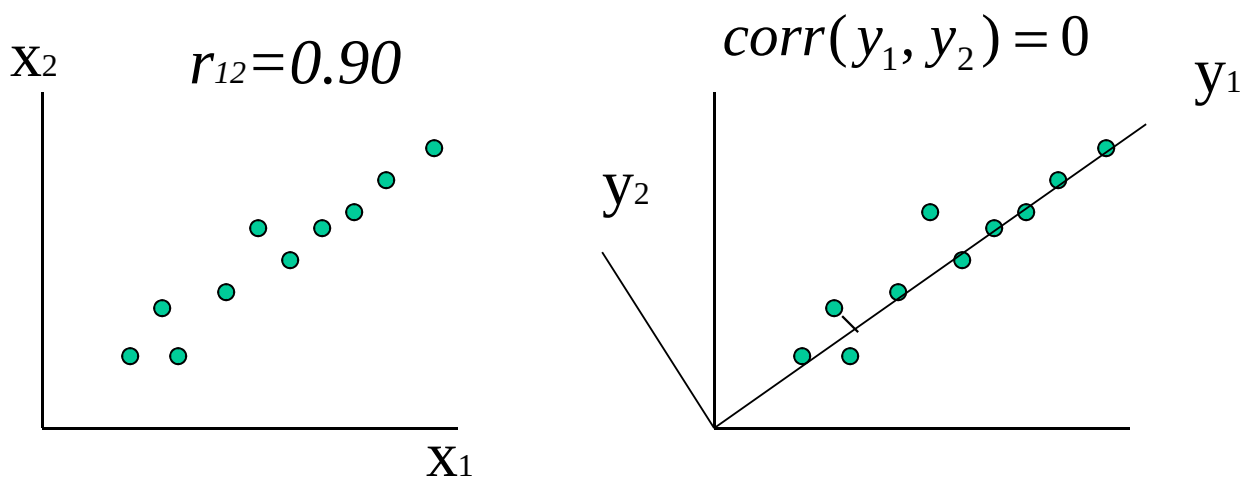
Correlation Matrix	Gene 1	Gene 2	…	Gene n
Gene 1	1	r_{12}	…	r_{1n}
Gene 2	r_{21}	1	…	r_{2n}
Gene ...	⋮	⋮	⋮	⋮
Gene n	r_{n1}	r_{n2}	…	1

A correlation matrix is a symmetric matrix of correlation coefficients

($-1 \le r_{ij} \le 1$ and $r_{ij} = r_{ji}; i, j = 1,2,...,n; r_{ii} = 1$)

The Results of PCA are a small set of the orthogonal (independent) Variables Grouping of the Variables

From a purely mathematical viewpoint the purpose of PCA is to transform **n** correlated random variables to an orthogonal set which reproduces the original variance/covariance structure.

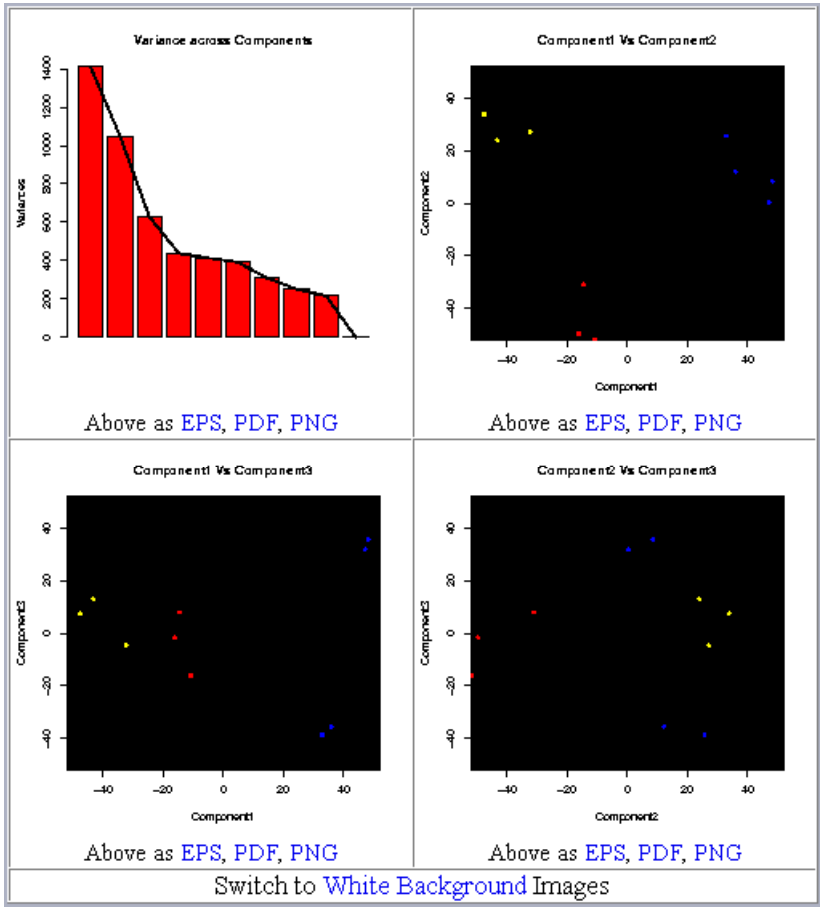


(The First) Principal Component y_1 can “explain” the major fraction (~90%) of a dispersion of variables x_1 and x_2 for all of the 10 observed objects.

Sample:Small Round Blue Cell Tumors (SRBCTs)

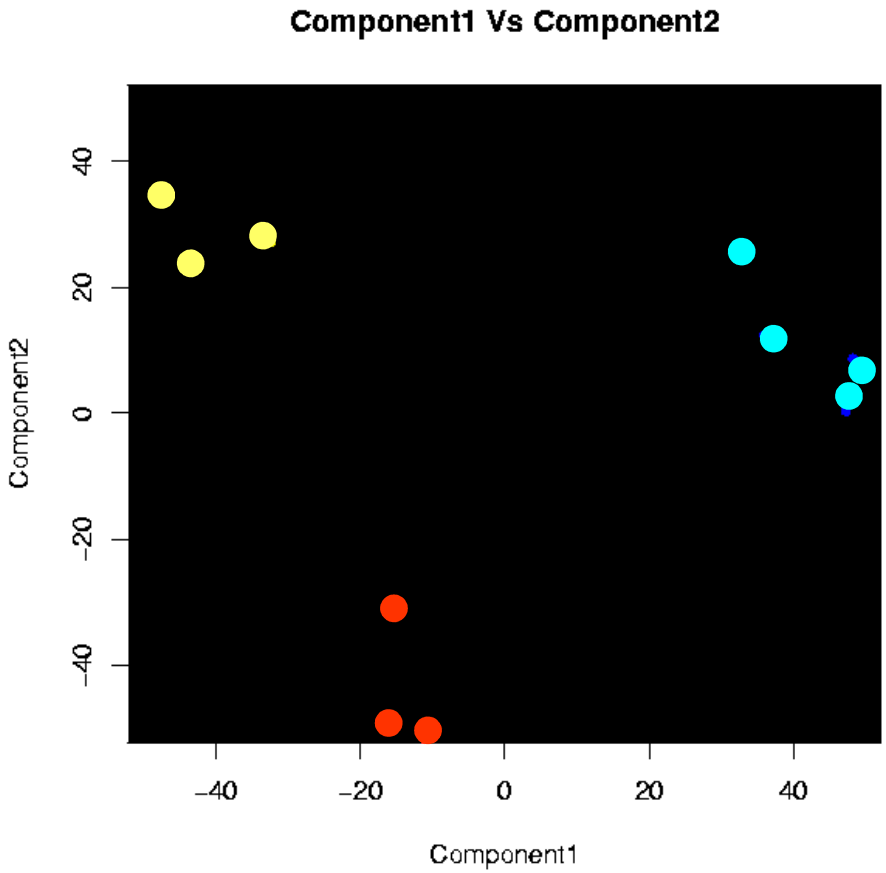
- 63 Arrays representing 4 groups
  - BL (Burkitt Lymphoma, n1=8)
  - EWS (Ewing, n2=23)
  - NB (neuroblastoma, n3=12)
  - RMS (rhabdomyosarcoma, n4=20)
- There are 2308 features (distinct gene probes)

PCA Detailed Plot



- ”Scree” plot
- 2-D plots

PCA 2-D plots



- First 2 components separate 3 groups well

MDS overview
(Multidimensional Scaling)

- An alternative for PCA
- Non-linear projection methodology
- Tolerates missing values

Summary of PCA and MDS

- Dimension reduction tools
- Graphic representation to help explain patterns
- Quality control for experimental variance

Hands-on Session 3

- Lab 8
- Total time: 15 minutes

- Next class tomorrow at 1:00 pm