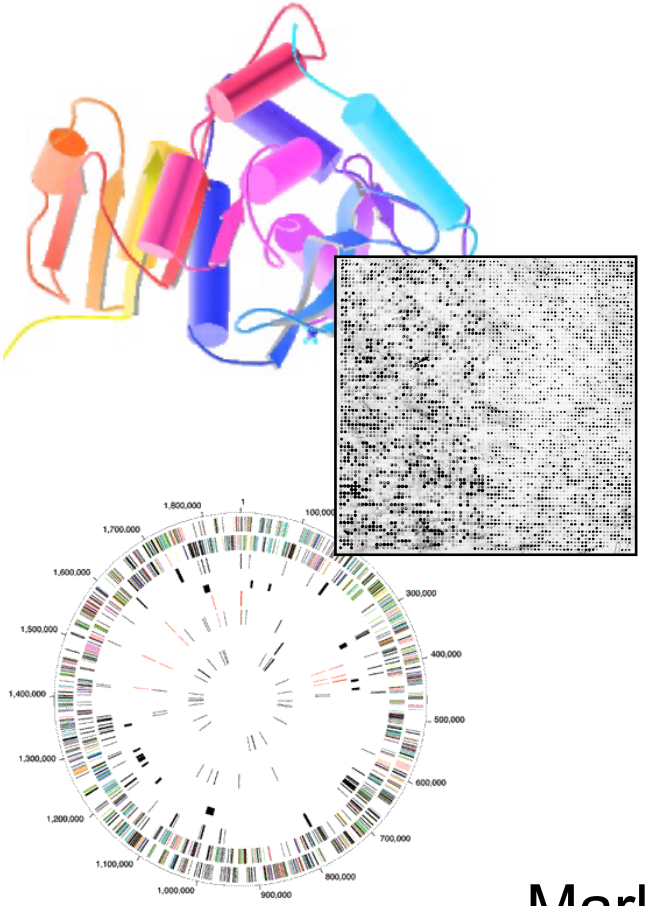
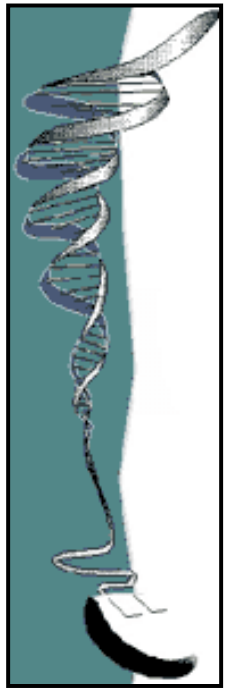
BIOINFORMATICS
Datamining #2





Mark Gerstein, Yale University

gersteinlab.org/courses/452

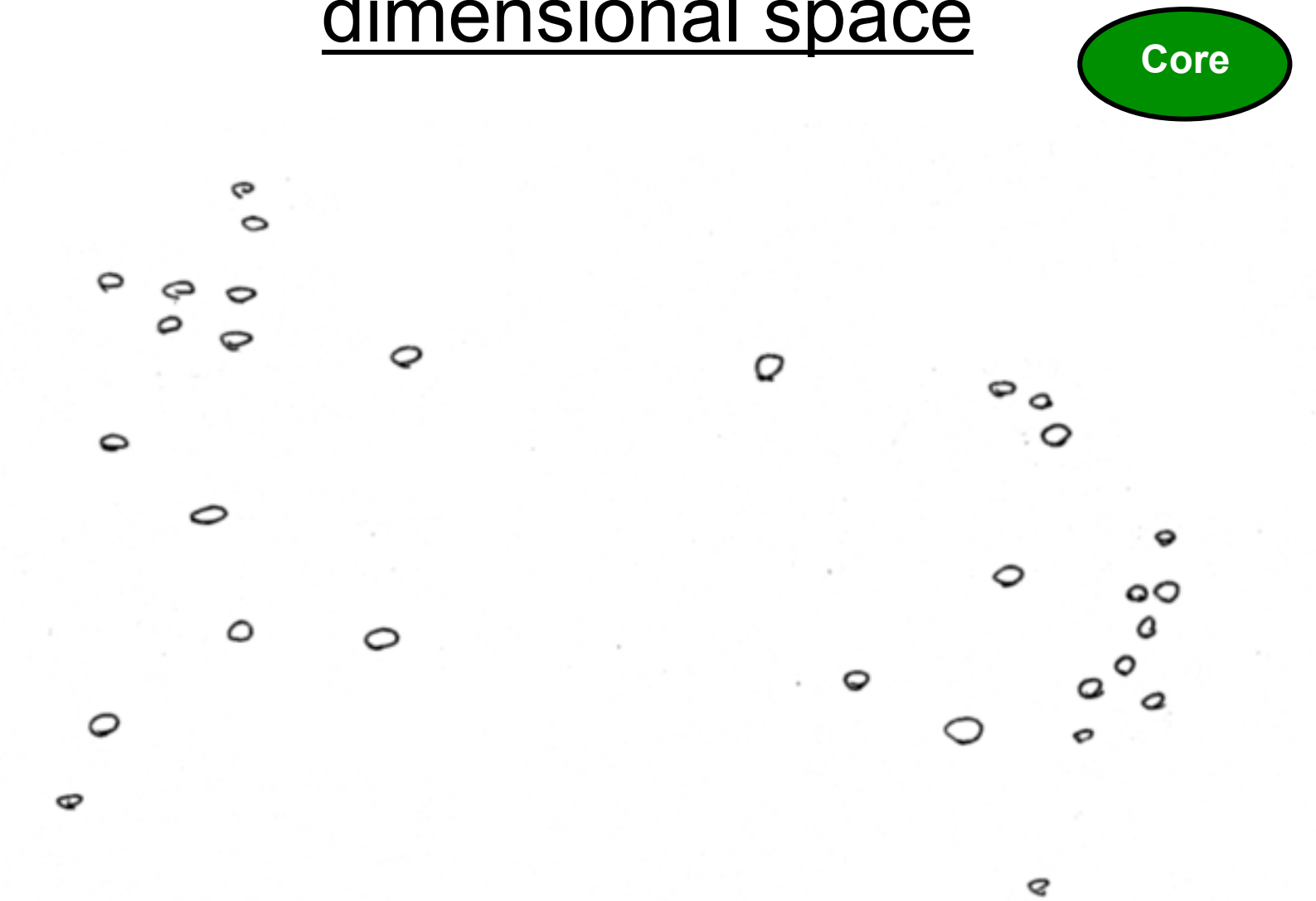
Spectral Methods Outline & Papers

- Simple background on PCA (emphasizing lingo)
- More abstract run through on SVD
- Application to
  - ◊ J Qian et al. (2001). "Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions." J Mol Biol 314: 1053-66.
  - ◊ O Alter et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." PNAS vol. 97: 10101-10106
  - ◊ Y Kluger et al. (2003). "Spectral biclustering of microarray data: coclustering genes and conditions." Genome Res 13: 703-16.

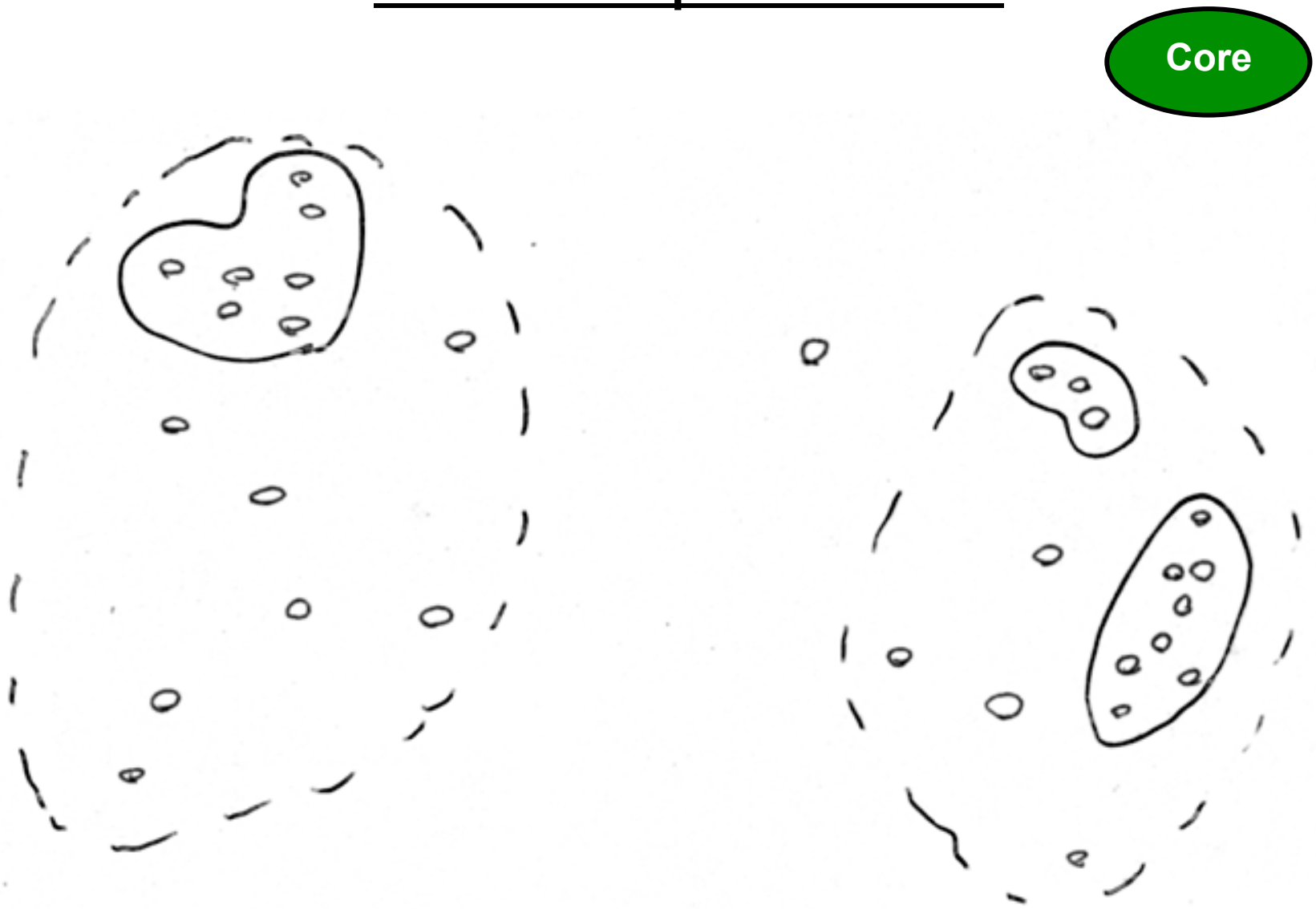
Typical Predictors and Response for Yeast

Basics		Predictors																			Response		
		Sequence Features													Genomic Features								
		seq. length	Amino Acid Composition							How many times does the sequence have these motif features?						Abs. expr. Level (mRNA copies / cell)		Prot. Abun-dance	Cell cycle timecourse				Function		Localization
Yeast Gene ID	Sequence		A	C	D	E	F	W	Y	farn site	NLS	hdel motif	nuc2	signalp	tms1	Gene-Chip expt. from RY Lab	sage tag freq.	(1000 copies /cell)	t=0	t=1	t=15	t=16	function ID(s) (from MIPS)	function description	5-compartment
YAL001C	MNIFEMLRI	1160	.08	.02	.06			.01	.04	0	1	0	1	0	0	0.3	0	?	5	3	4	5	04.01.01;04.03	TFIIIC (transcription initi	N
YAL002W	KVFGRCELA	1176	.09	.02	.06			.01	.04	0	0	0	0	0	1	0.2	?	?	8	4	4	3	06.04;08.13	vacuolar sorting protein,	C
YAL003W	KMLQFNLRW	206	.08	.02	.06			.01	.04	0	0	0	0	0	0	19.1	19	23	70	73	98	126	05.04;30.03	translation elongation fac	N
YAL004W	RPDFCLEPP	215	.08	.02	.06			.01	.04	0	0	0	0	0	0	?	0	?	18	12	4	6	01.01.01	0	N
YAL005C	VINTFDGVA	641	.08	.02	.06			.01	.04	0	0	0	0	0	1	13.4	16	17	39	38	8	14	06.01;06.04;08	heat shock protein of HS	????
YAL007C	KKAVINGEQ	190	.08	.02	.06			.01	04	0	0	0	0	1	4	2.2	8	?	15	20	16	17	99	????	????
YAL008W	HPETLVKVK	198	.08	.02	.06			.01	.04	0	0	0	0	0	3	1.2	?	?	9	6	2	3	99	????	????
YAL009W	PTLEWFLSH	259	.08	.02	.06			.01	.04	0	2	0	0	0	3	0.6	?	?	6	2	3	5	03.10;03.13	meiotic protein	????
YAL010C	MEQRITLKD	493	.08	.02	.06			.02	.04	0	0	0	0	0	1	0.3	?	?	11	6	6	6	30.16	involved in mitochondrial	????
YAL011W	KSFPEVVGK	616	.08	.02	.06			.01	.04	0	8	0	1	0	0	0.4	?	?	6	5	5	6	30.16;99	protein of unknown funct	????
YAL012W	GVQVETISP	393	.08	.02	.06			.01	.04	0	0	0	0	0	1	8.9	4	6.7	29	26	23	29	01.01.01;30.03	cystathionine gamma-lya	C
YAL013W	RTDCYGNVN	362	.08	.02	.06			.01	.04	0	0	0	0	0	0	0.6	?	?	7	9	6	10	01.06.10;30.03	regulator of phospholipid	N
YAL014C	GDVEKGKKI	202	.08	.02	.06			.01	.04	0	0	0	0	0	0	1.1	?	?	12	13	9	12	99	????	N
YAL015C	MTPAVTTYK	399	.08	.02	.06			.01	.04	0	1	0	0	0	0	0.7	0	1	19	18	12	13	11.01;11.04	DNA repair protein	N
YAL016W	KKPLTQEQL	635	.08	.02	.06			.01	.04	0	0	0	0	0	1	3.3	5	?	15	20	16	16	03.01;03.04;03	ser/thr protein phosphata	????

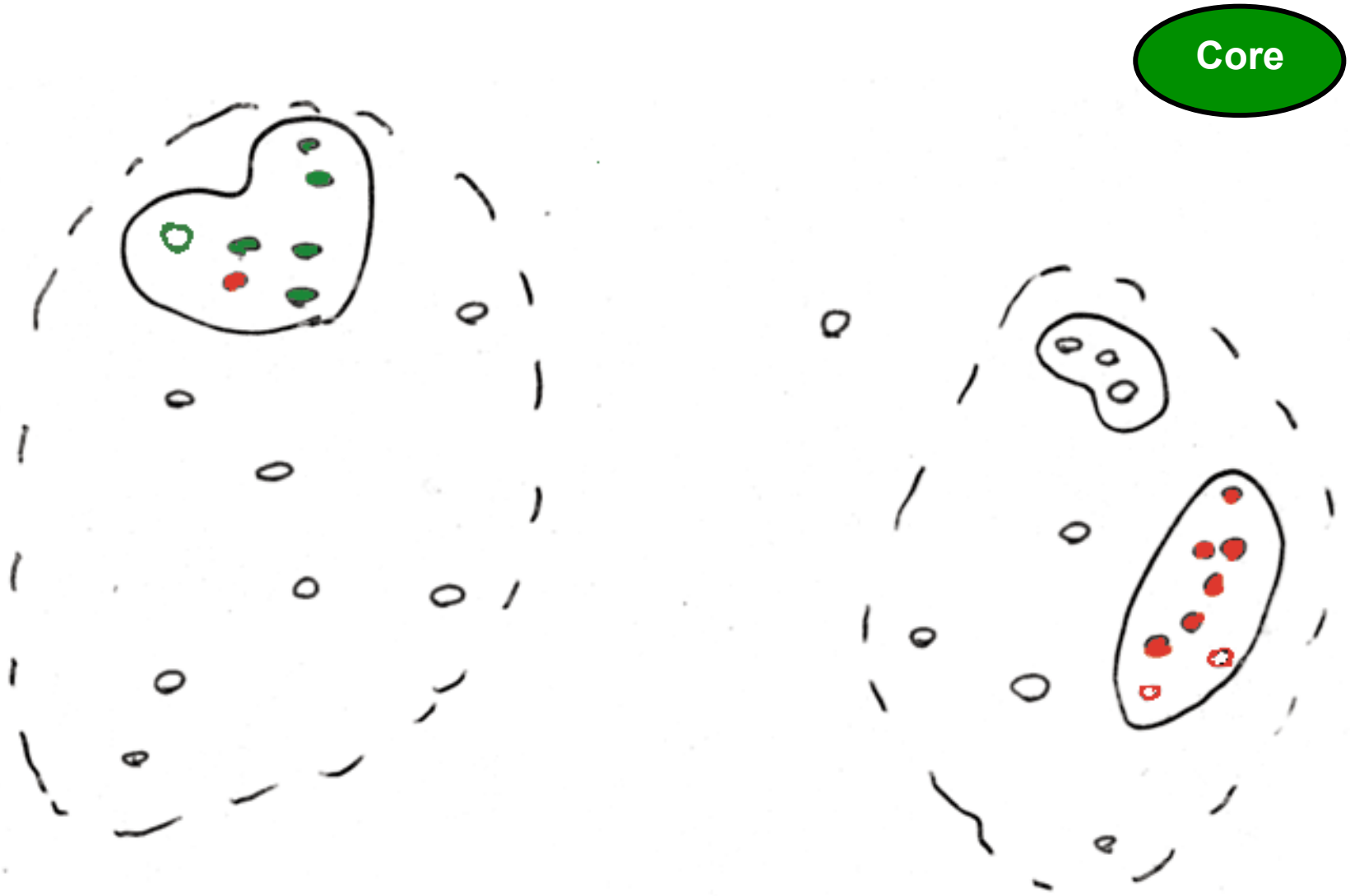
Represent predictors in abstract high dimensional space



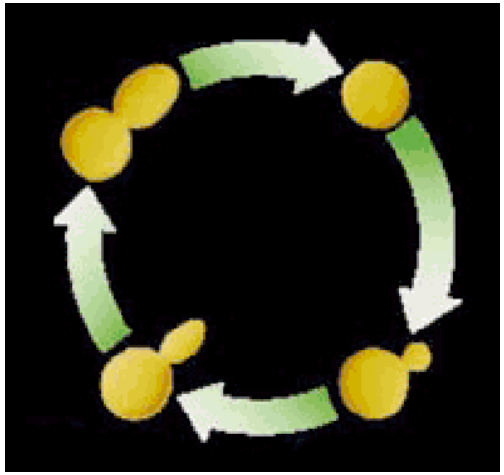
“cluster” predictors



Use clusters to predict Response

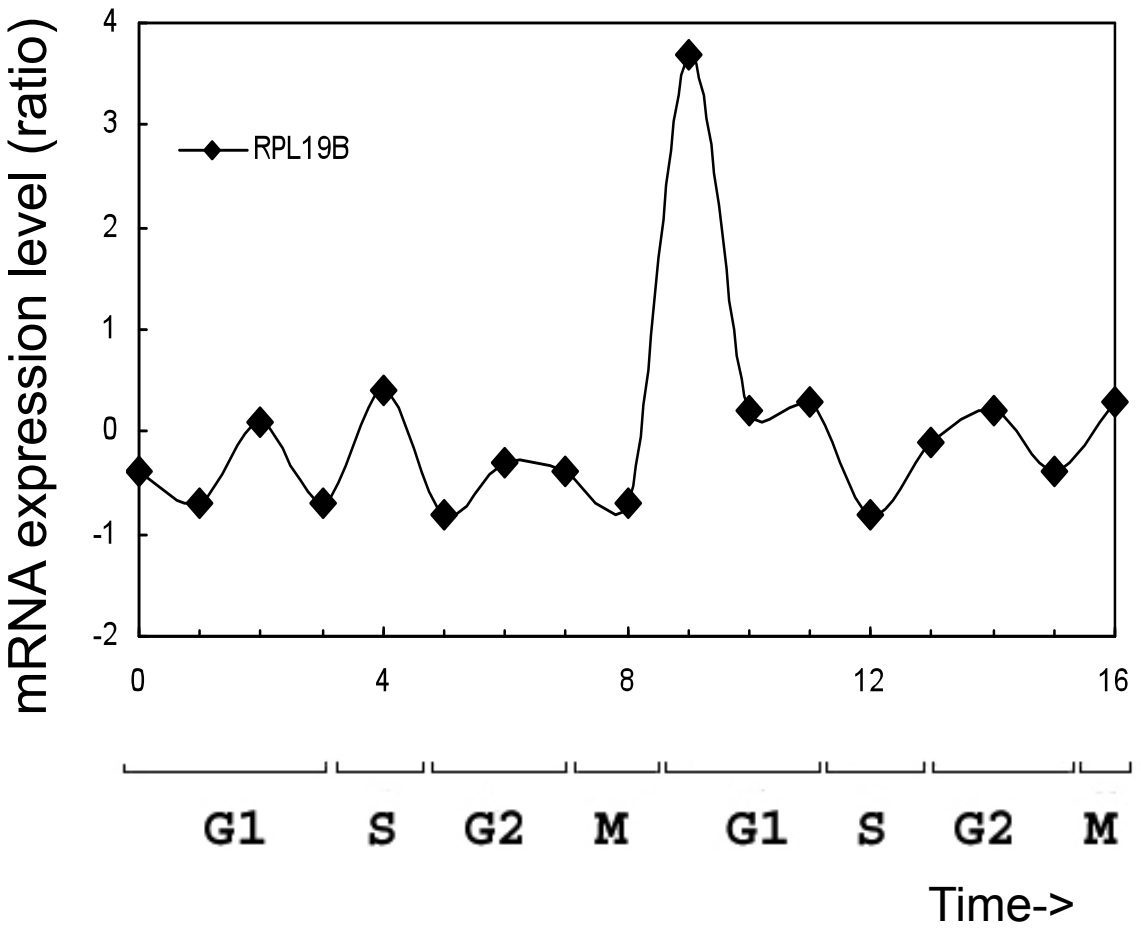


Clustering the yeast cell cycle to uncover interacting proteins



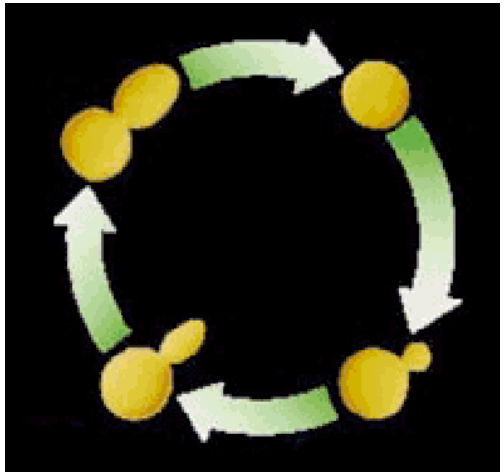
Extra

[Brown, Davis]

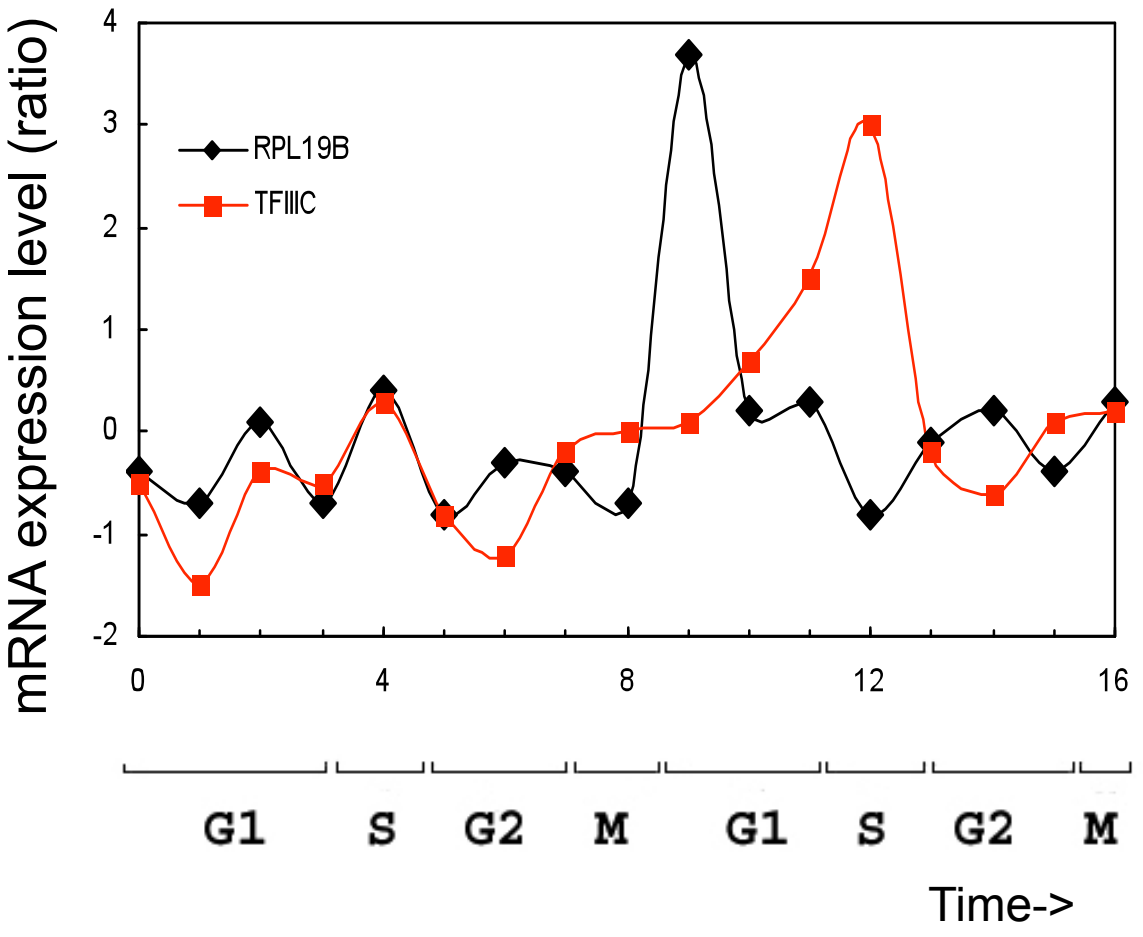


Microarray timecourse of
1 ribosomal protein

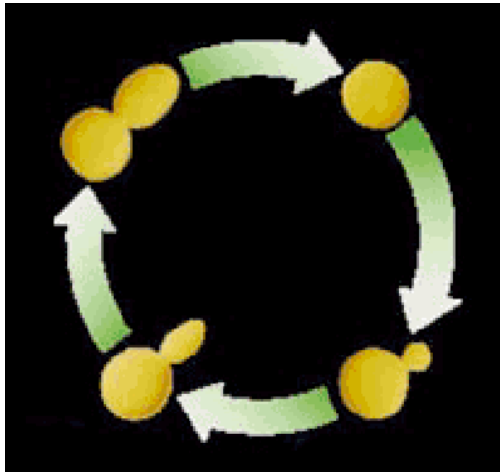
Clustering the yeast cell cycle to uncover interacting proteins



Extra

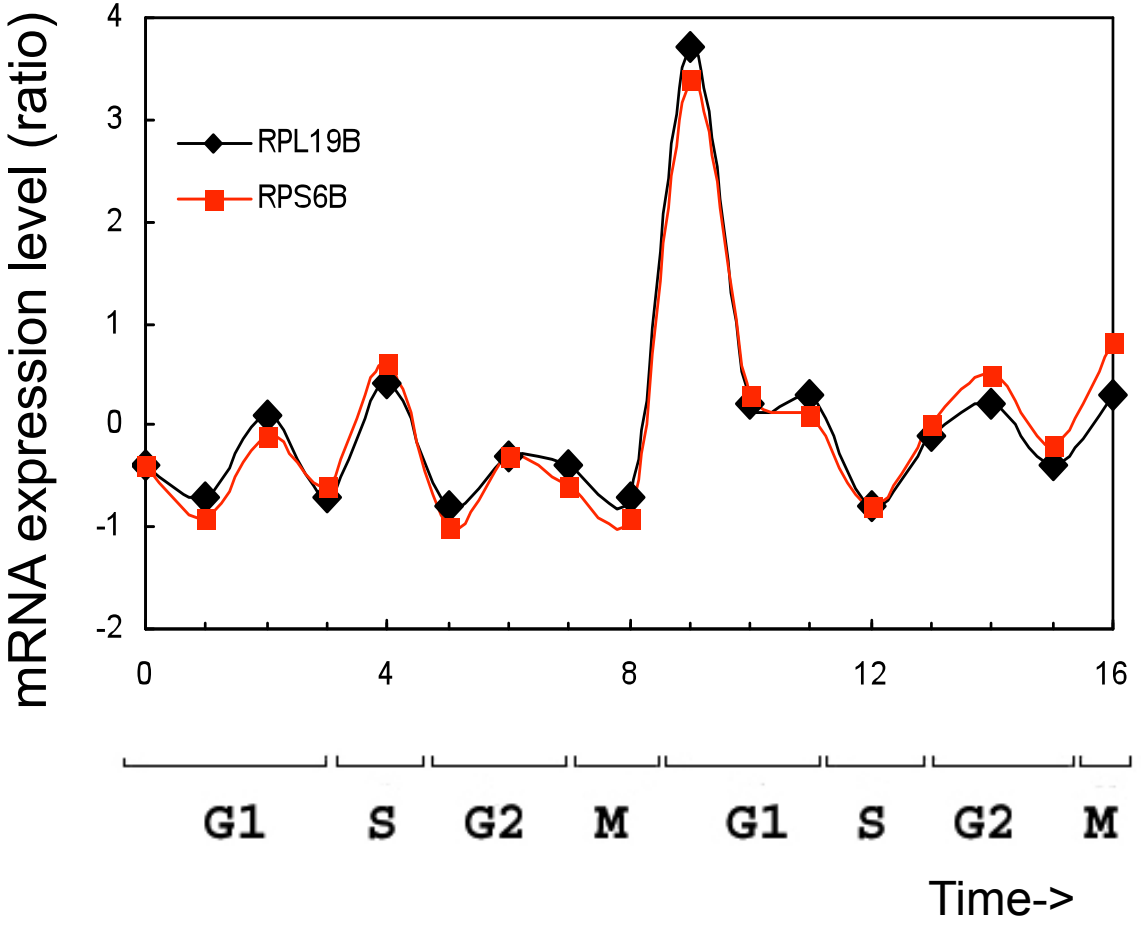


Clustering the yeast cell cycle to uncover interacting proteins



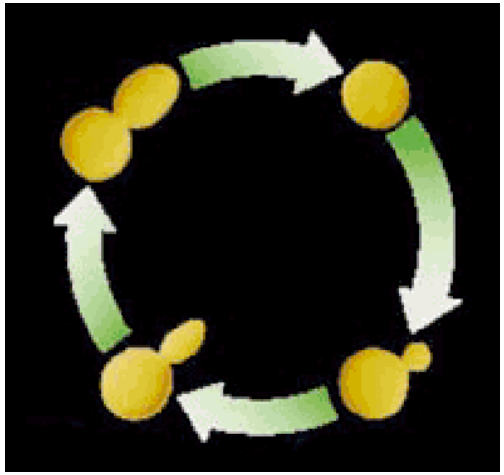
Extra

[Botstein; Church, Vidal]

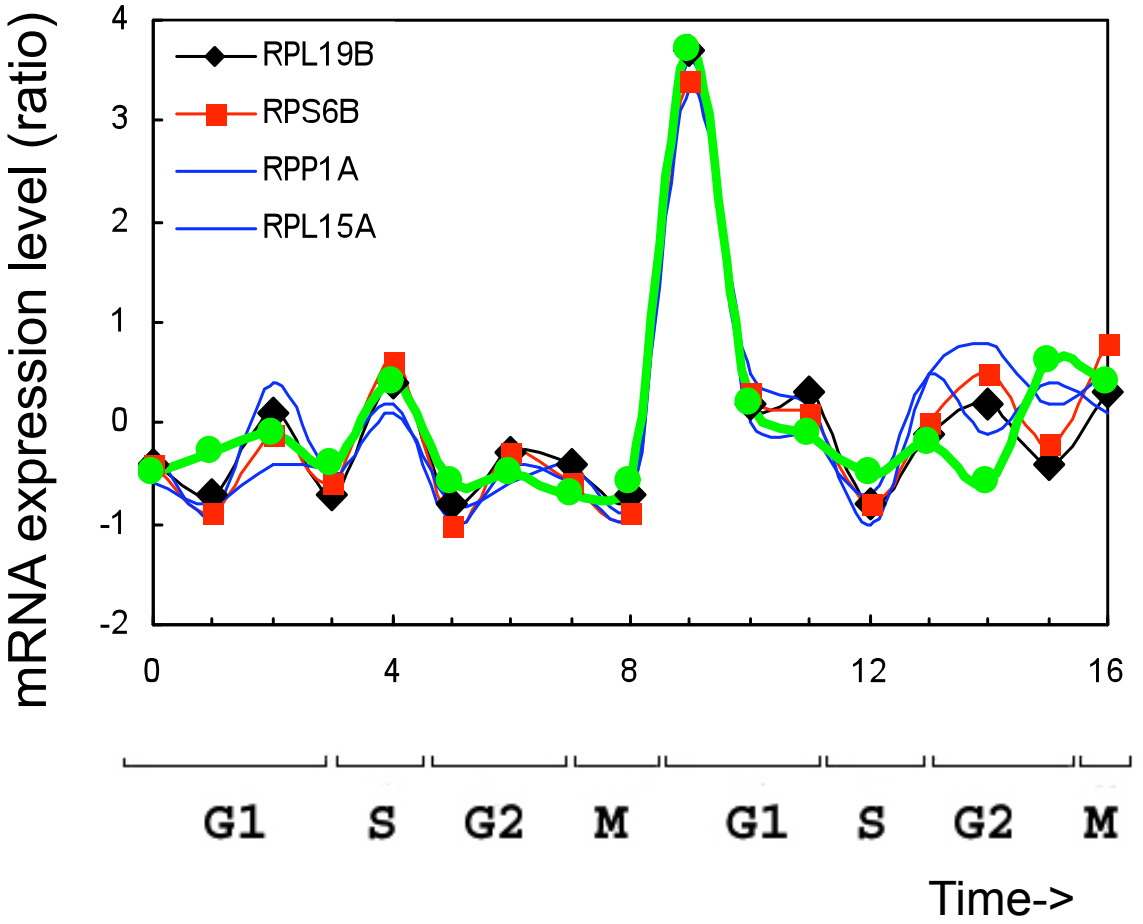


Close relationship from 18M
(2 Interacting Ribosomal Proteins)

Clustering the yeast cell cycle to uncover interacting proteins

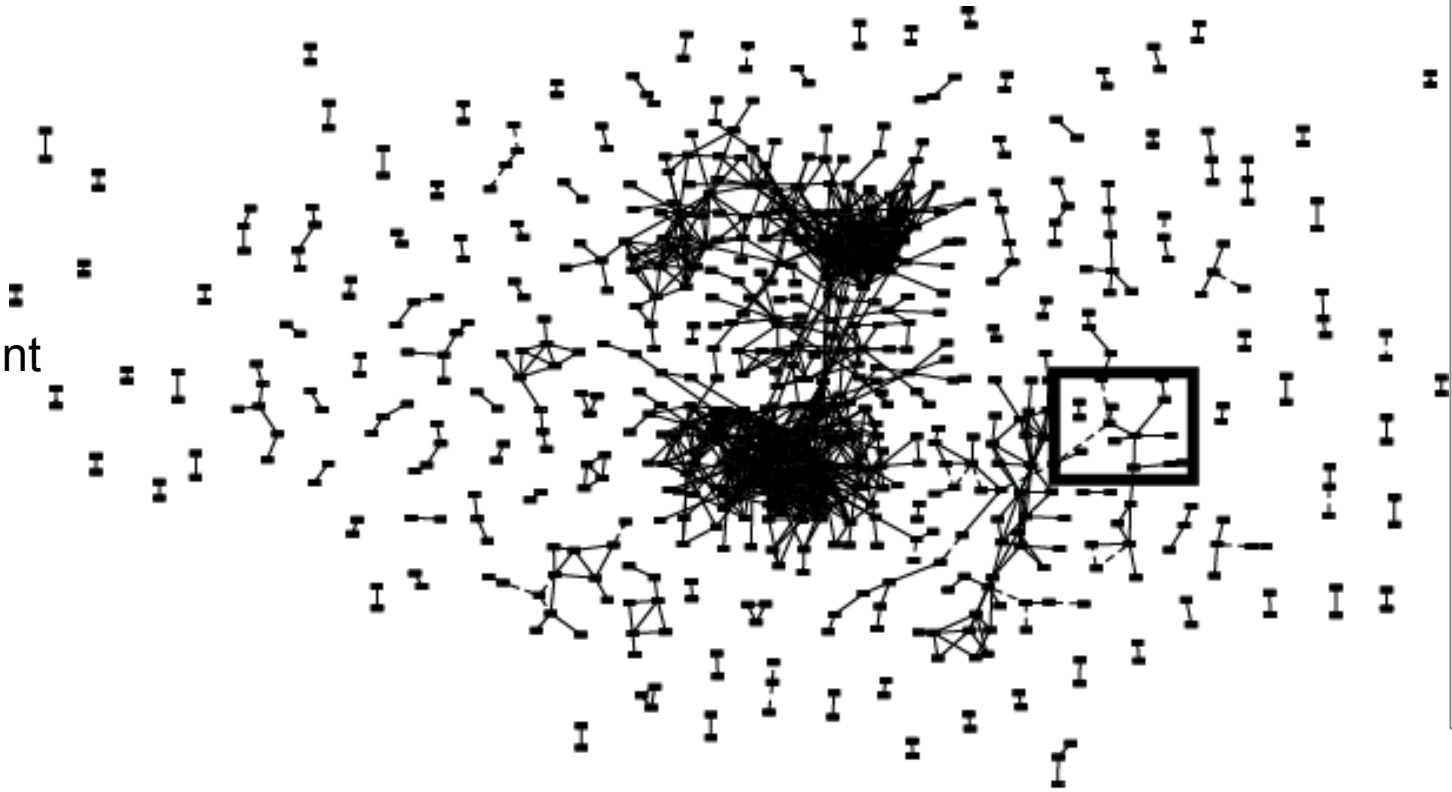


Extra



Global Network of Relationships

Core

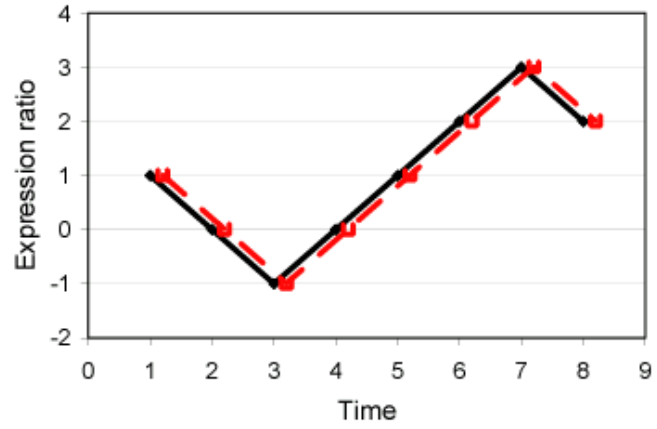


~470K significant relationships from **~18M** possible

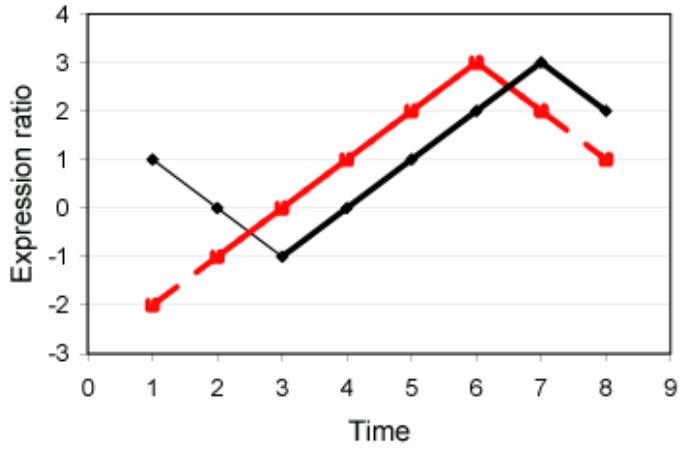
Intuition in terms of Adj. Matrix

Simultaneous

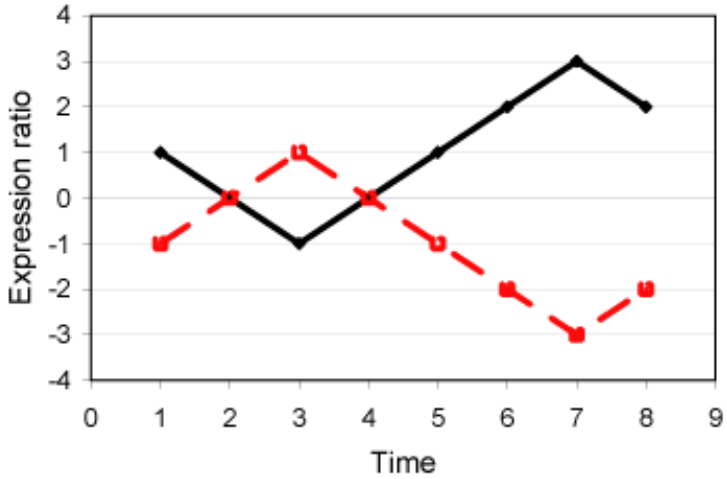
Traditional Global Correlation



Time-Shifted



Inverted



Local Clustering algorithm identifies further (reasonable) types of expression relation-ships

Core

[Church]

Local Alignment

Suppose there are n (1, 2, …, n) time points:

- The expression ratio is normalized in "Z-score" fashion;
- Score matrix: $S_{i,j} = S(x_i, y_j) = x_i \bullet y_j$;

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Suppose there are n (1, 2, …, n) time points:

- Sum matrices $E_{i,j}$ and $D_{i,j}$:

$$E_{i,j} = \max(E_{i-1,j-1} + S_{i,j}, 0);$$

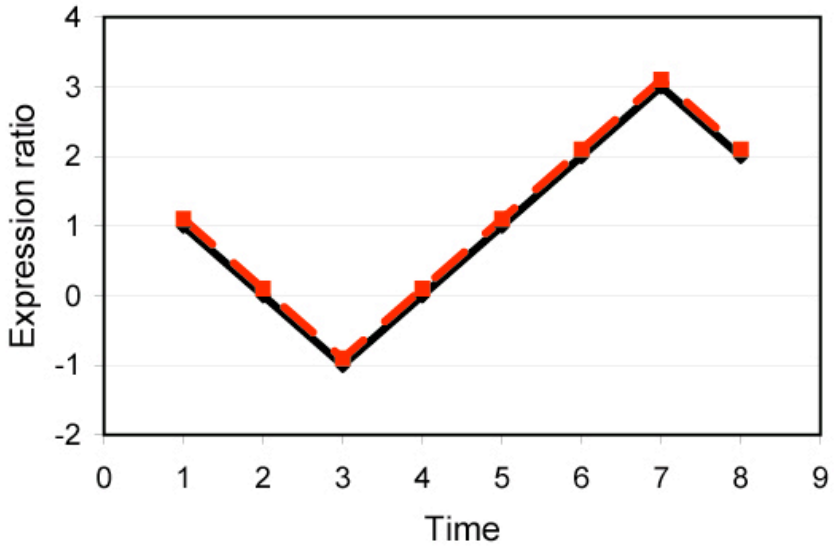
$$D_{i,j} = \max(D_{i-1,j-1} - S_{i,j}, 0);$$

- Match Score = $\max(E_{i,j}, D_{i,j})$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Simultaneous



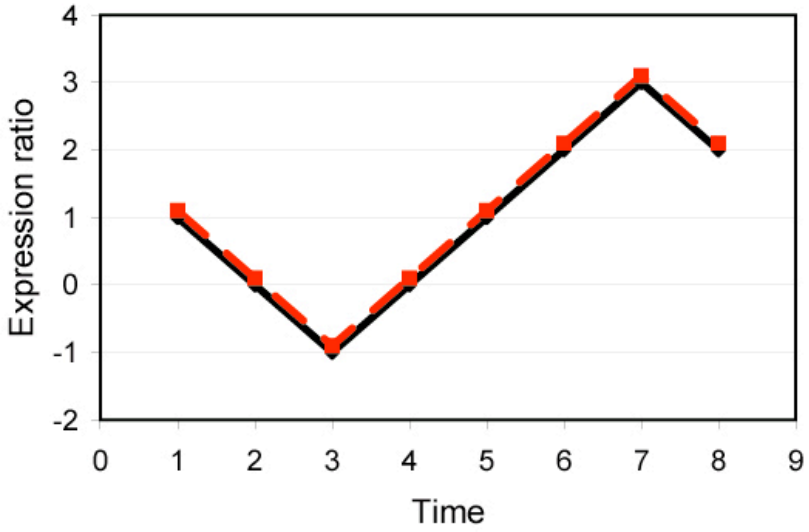
		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	-1	0	1	2	3	2
0	0	0	0	0	0	0	0	0	0
-1	0	-1	0	1	0	-1	-2	-3	-2
0	0	0	0	0	0	0	0	0	0
1	0	1	0	-1	0	1	2	3	2
2	0	2	0	-1	0	2	4	6	4
3	0	3	0	-3	0	3	6	9	6
2	0	1	0	-2	0	2	4	6	4

$$S_{i,j} = x_i \bullet y_j$$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Simultaneous



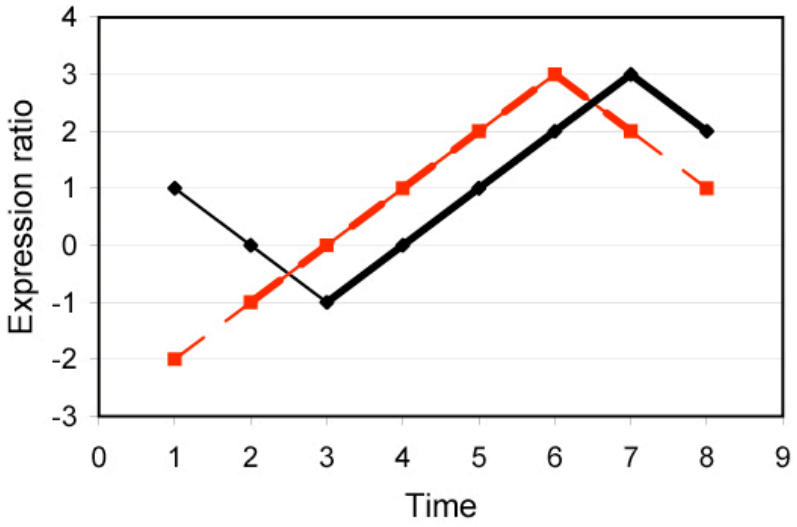
		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$$E_{i,j} = \max(E_{i-1,j-1} + x_i \cdot y_j, 0)$$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Time-Shifted



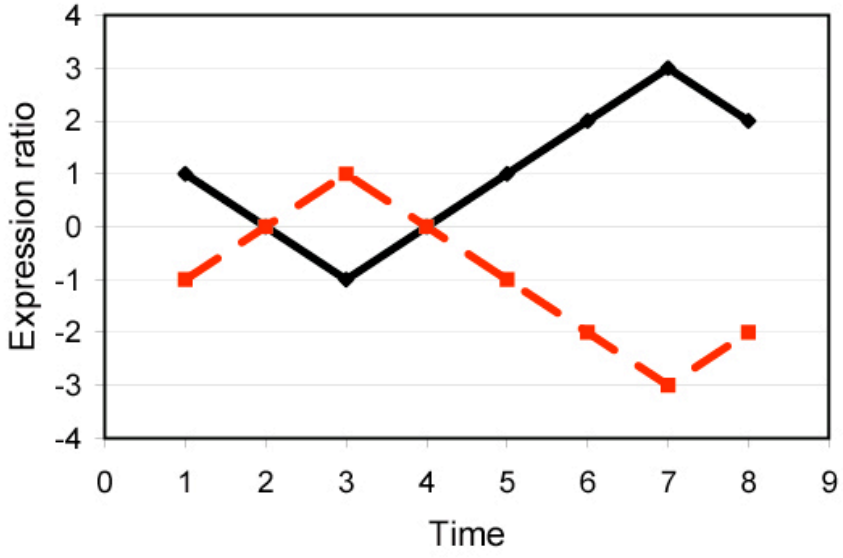
		-2	-1	0	1	2	3	2	1
	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	3	2	1
0	0	0	0	0	0	1	2	3	2
-1	0	2	1	0	0	0	0	0	2
0	0	0	2	1	0	0	0	0	0
1	0	0	0	2	2	2	3	2	1
2	0	0	0	0	4	6	8	7	4
3	0	0	0	0	3	10	15	14	10
2	0	0	0	0	2	7	16	19	16

$$E_{i,j} = \max(E_{i-1,j-1} + x_i \cdot y_j, 0)$$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Local Alignment

Inverted

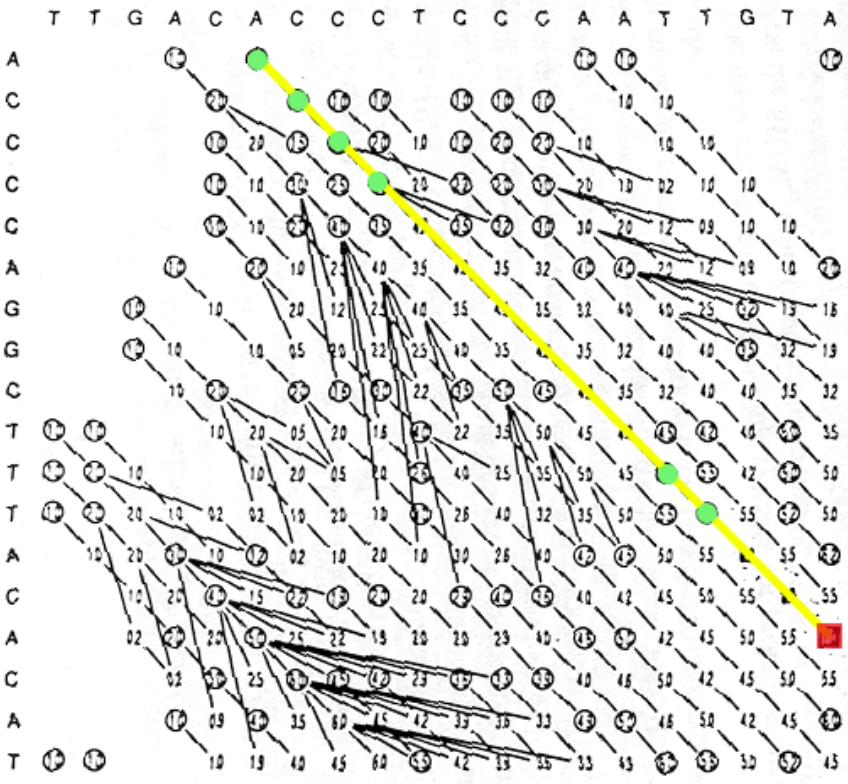


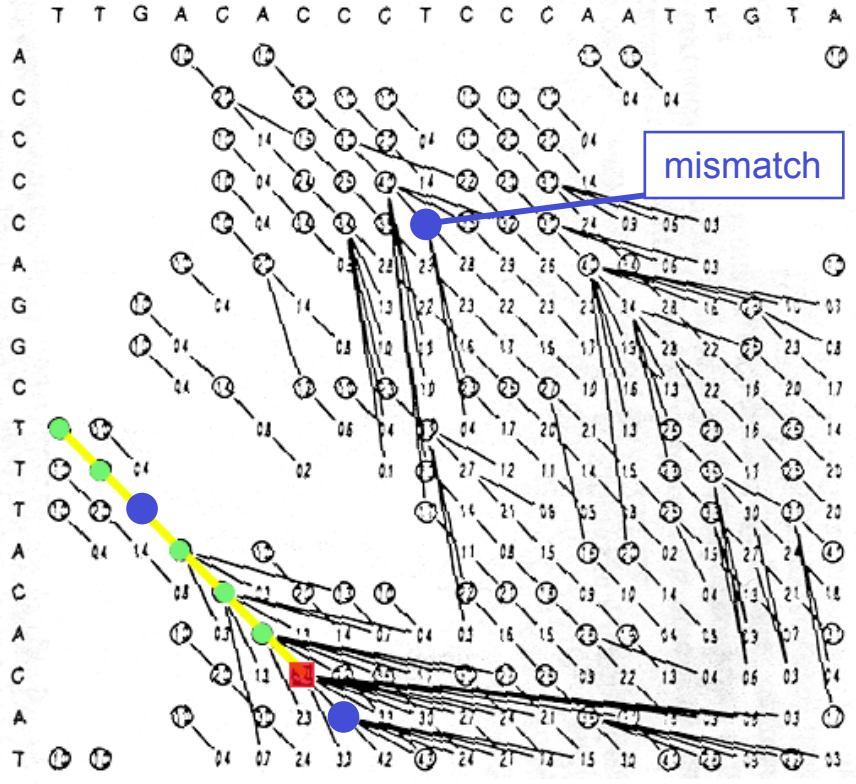
		-1	0	1	0	-1	-2	-3	-2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	2
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$$D_{i,j} = \max(D_{i-1,j-1} - x_i \cdot y_j, 0)$$

Qian J. et al. Beyond Synexpression Relationships: Local clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. J. Mol. Biol. (2001) 314, 1053-1066

Global (NW) vs Local (SW)Alignments





TTGACACCCTCCCAATTGTA...

| | | | | | |

.....ACCCCAGGC**TTTACAC**AT

1234444444566 67

T T G A C A C C...

| | - | | | | | -

T T T A C A C A...

1 2 1 2 3 4 5 4

0 0 4 4 4 4 4 8

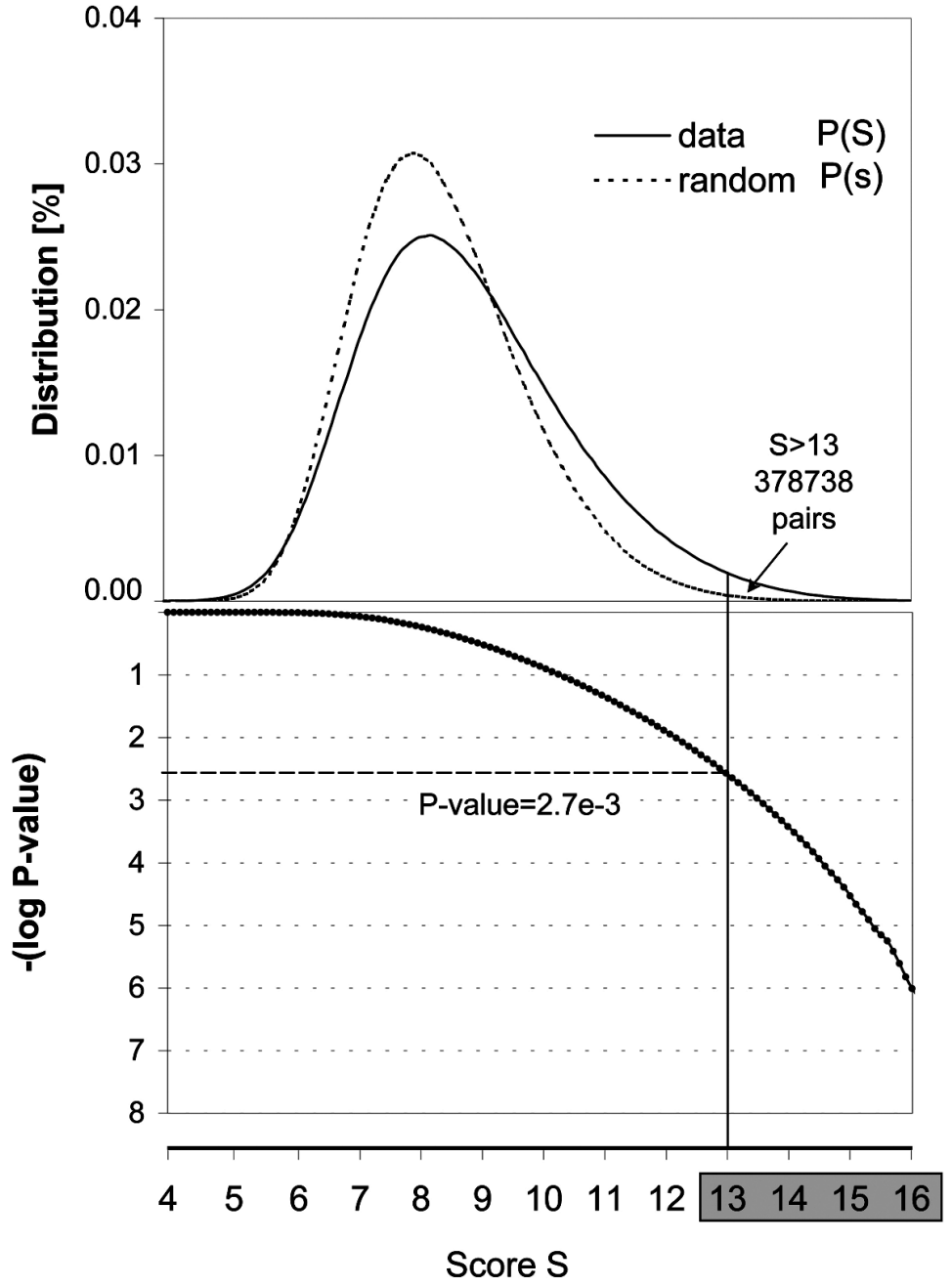
Match Score = +1

Gap-Opening=-1.2, Gap-Extension=-.03

for local alignment Mismatch = -0.6

Adapted from D J States & M S Boguski, "Similarity and Homology," Chapter 3 from Gribskov, M. and Devereux, J. (1992). Sequence Analysis Primer. New York, Oxford University Press. (Page 133)

Statistical Scoring



Examples time-shifted relationships

Suggestive

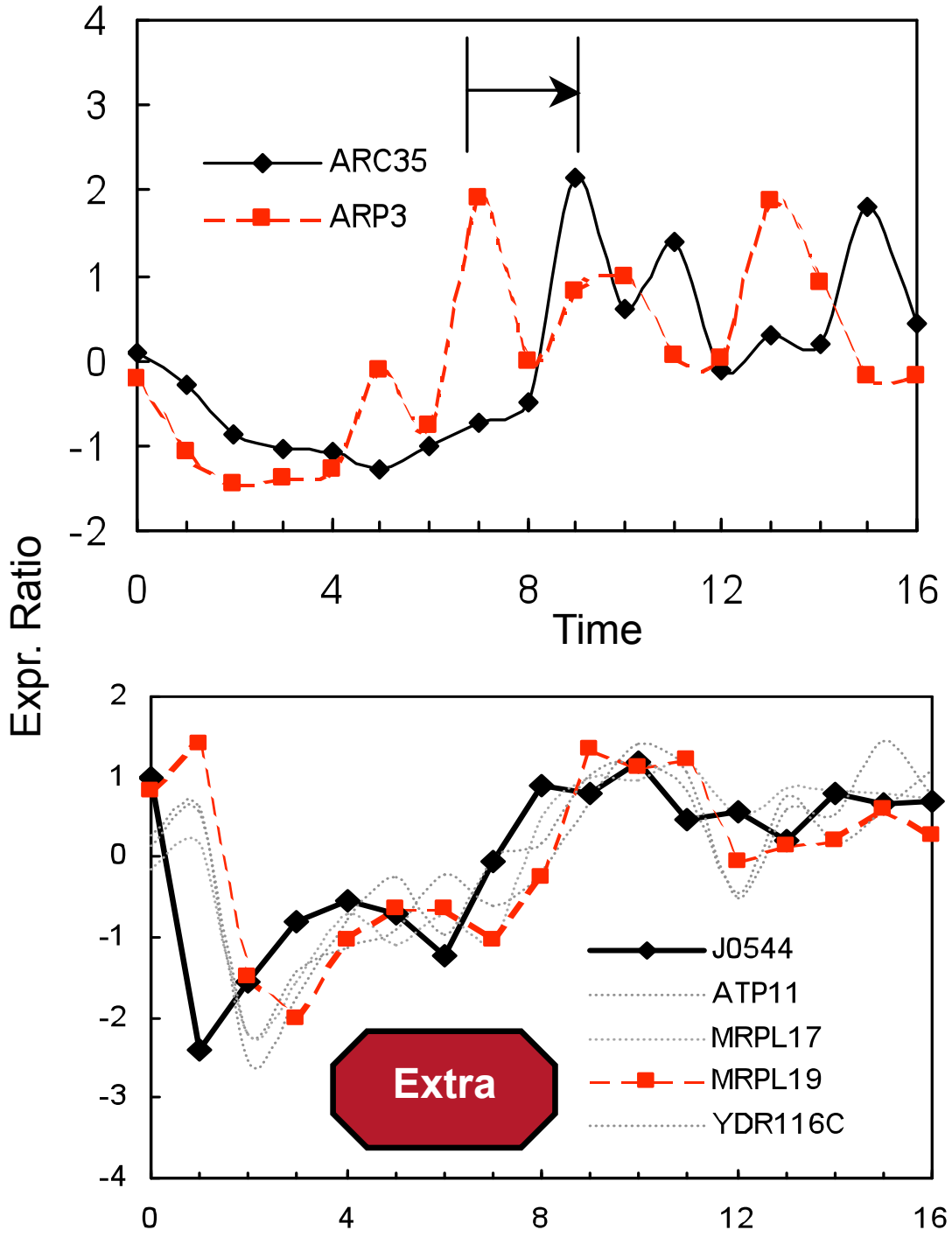
ARP3 : in actin remodelling cplx.

ARC35 : in same cplx. (required late in cell cycle)

Predicted

J0544 : **unknown** function

MRPL19: mito.ribosome



Examples time-shifted relationships

Suggestive

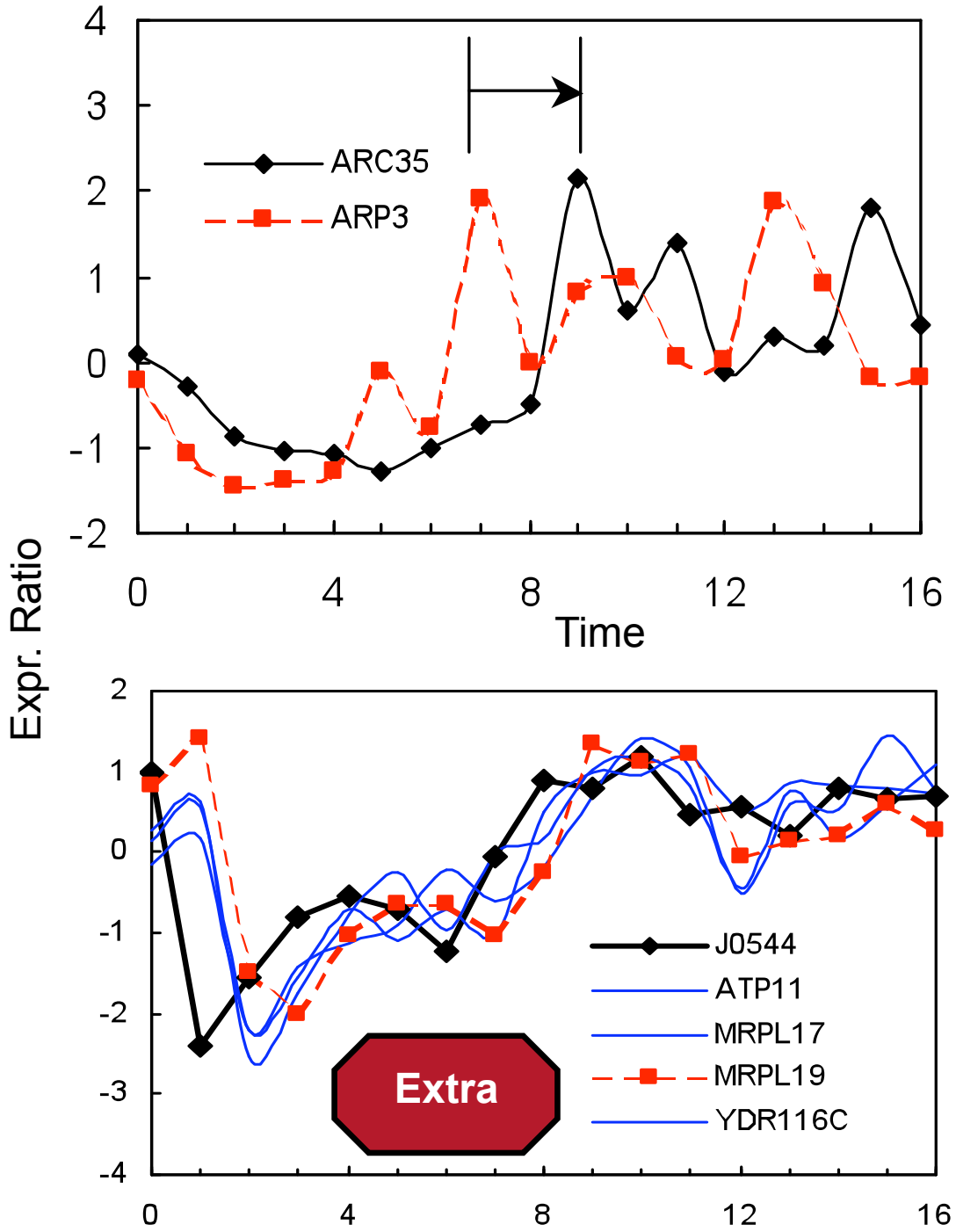
ARP3 : in actin remodelling cplx.

ARC35 : in same cplx. (required late in cell cycle)

Predicted

J0544 : **unknown** function

MRPL19: mito.ribosome

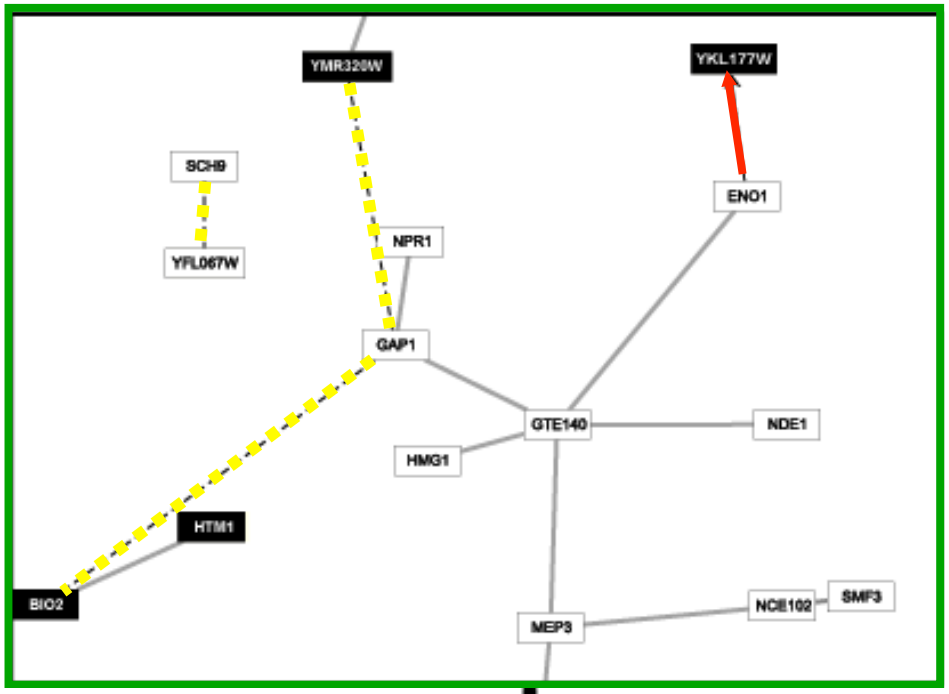


Global Network of 3 Different Types of Relationships

Simultaneous

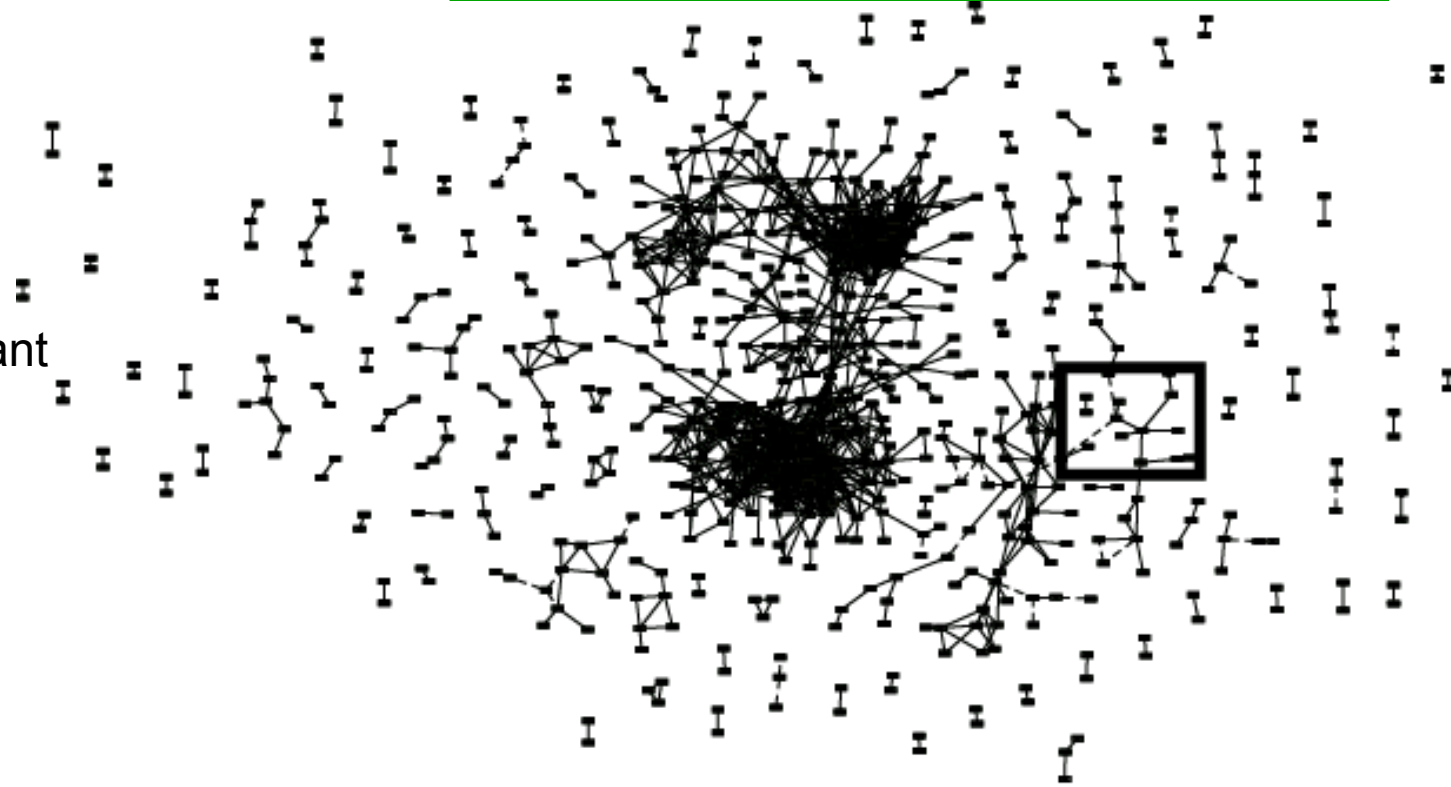
Inverted

Shifted



Extra

~470K significant relationships from **~18M** possible

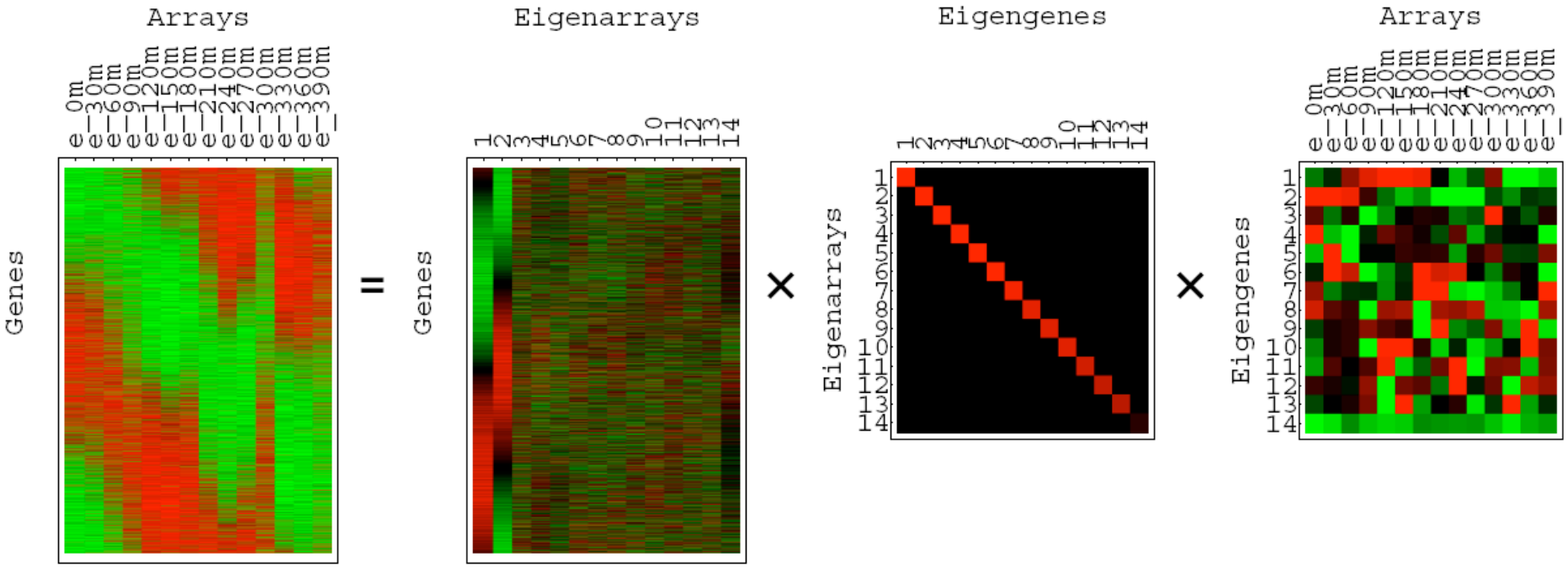


Large-scale Datamining

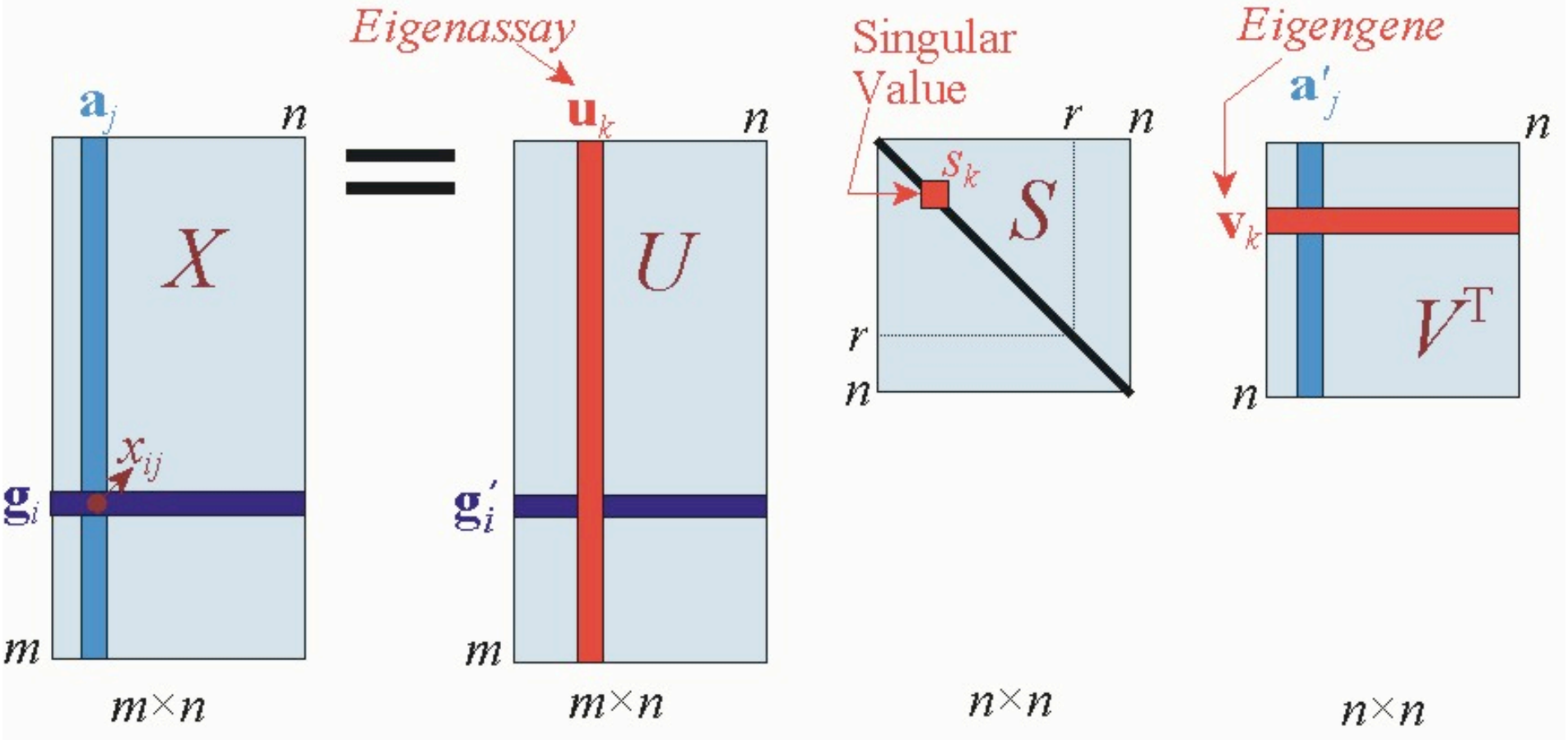
- Gene Expression
  - ◊ Representing Data in a Grid
  - ◊ Description of function prediction in abstract context
- Unsupervised Learning
  - ◊ clustering & k-means
  - ◊ Local clustering
- Supervised Learning
  - ◊ Discriminants & Decision Tree
  - ◊ Bayesian Nets
- Function Prediction EX
  - ◊ Simple Bayesian Approach for Localization Prediction

Intuition on interpretation of SVD in terms of genes and conditions

SVD for microarray data (Alter et al, PNAS 2000)

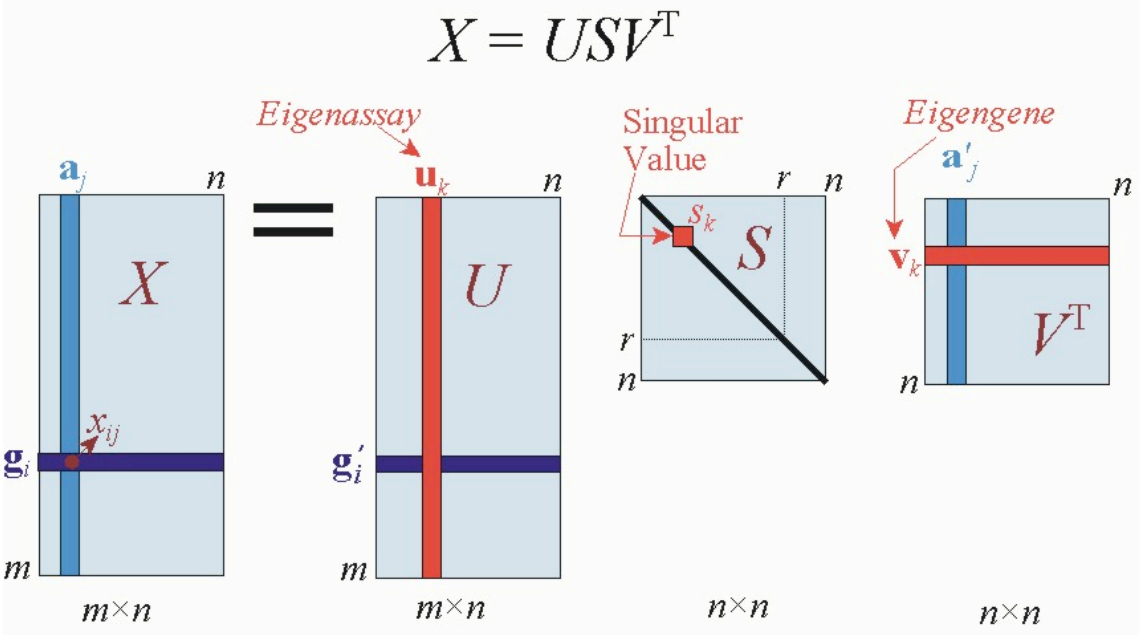


$$X = USV^{\mathrm{T}}$$

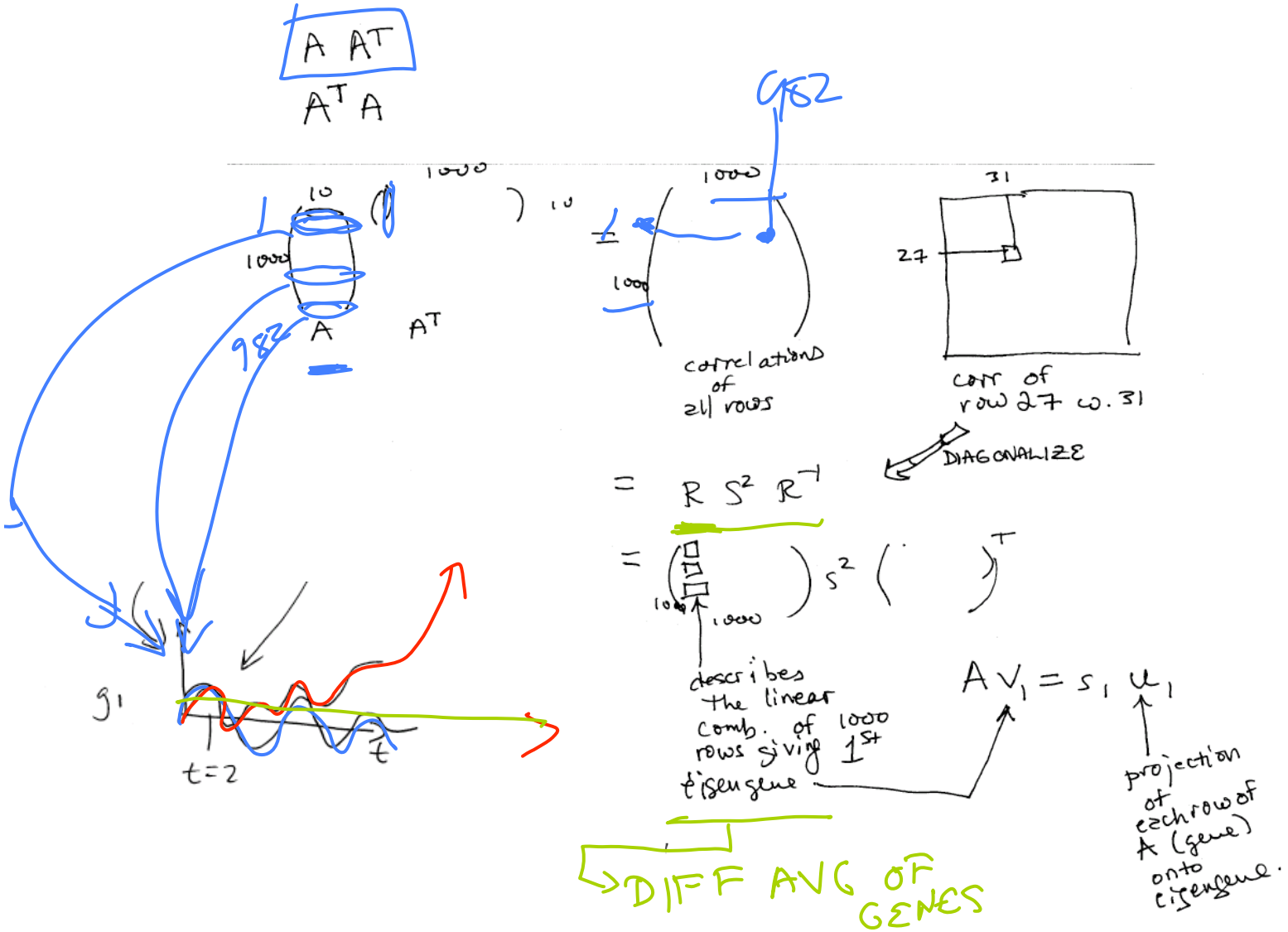


Notation

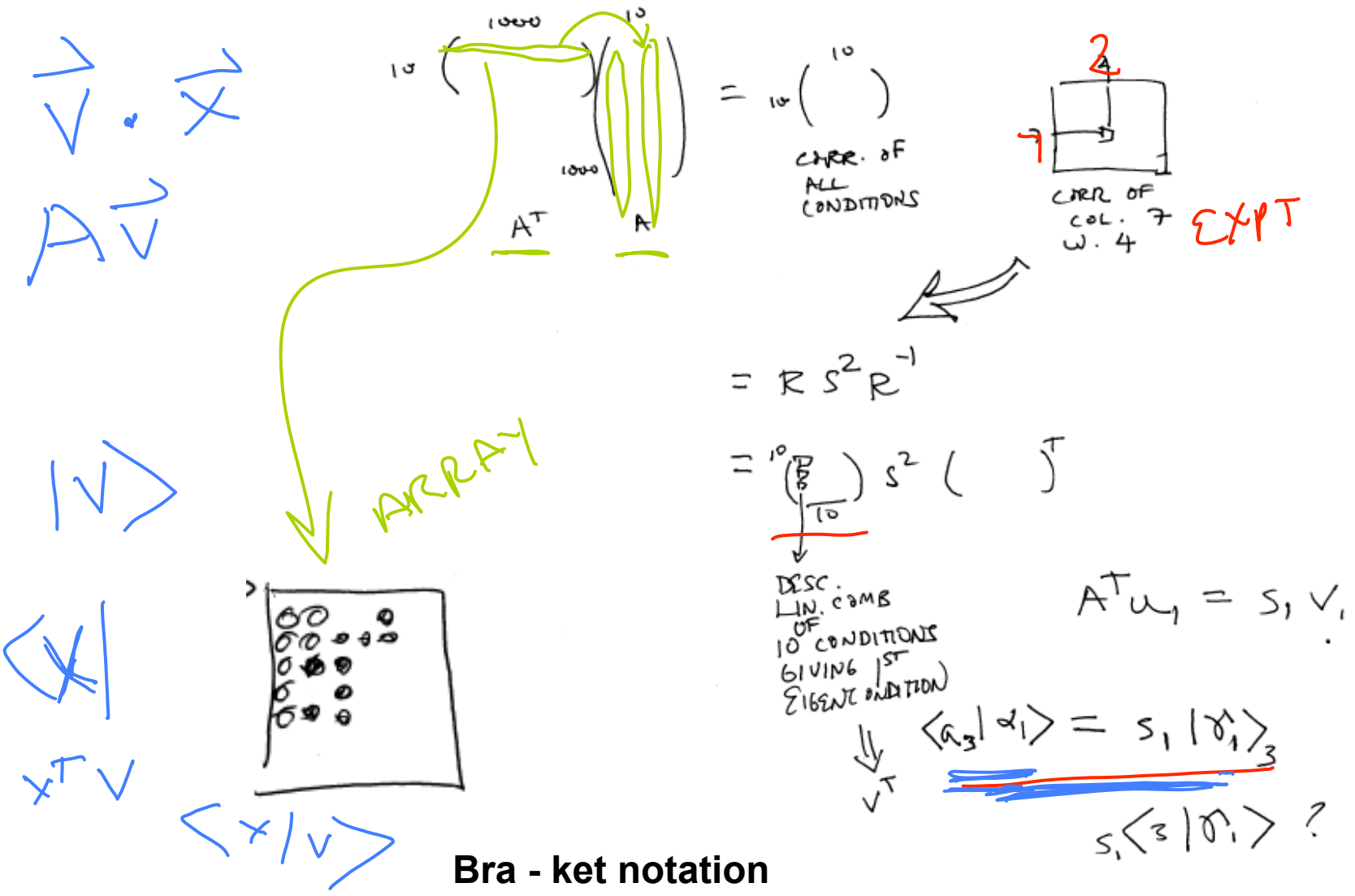
- m=1000 genes
  - row-vectors
  - 10 eigengene (v_i) of dimension 10 conditions
- n=10 conditions (assays)
  - column vectors
  - 10 eigenconditions (u_i) of dimension 1000 genes



Understanding Eigengenes (v_i) in terms PCA on (large) gene-gene correlation matrix

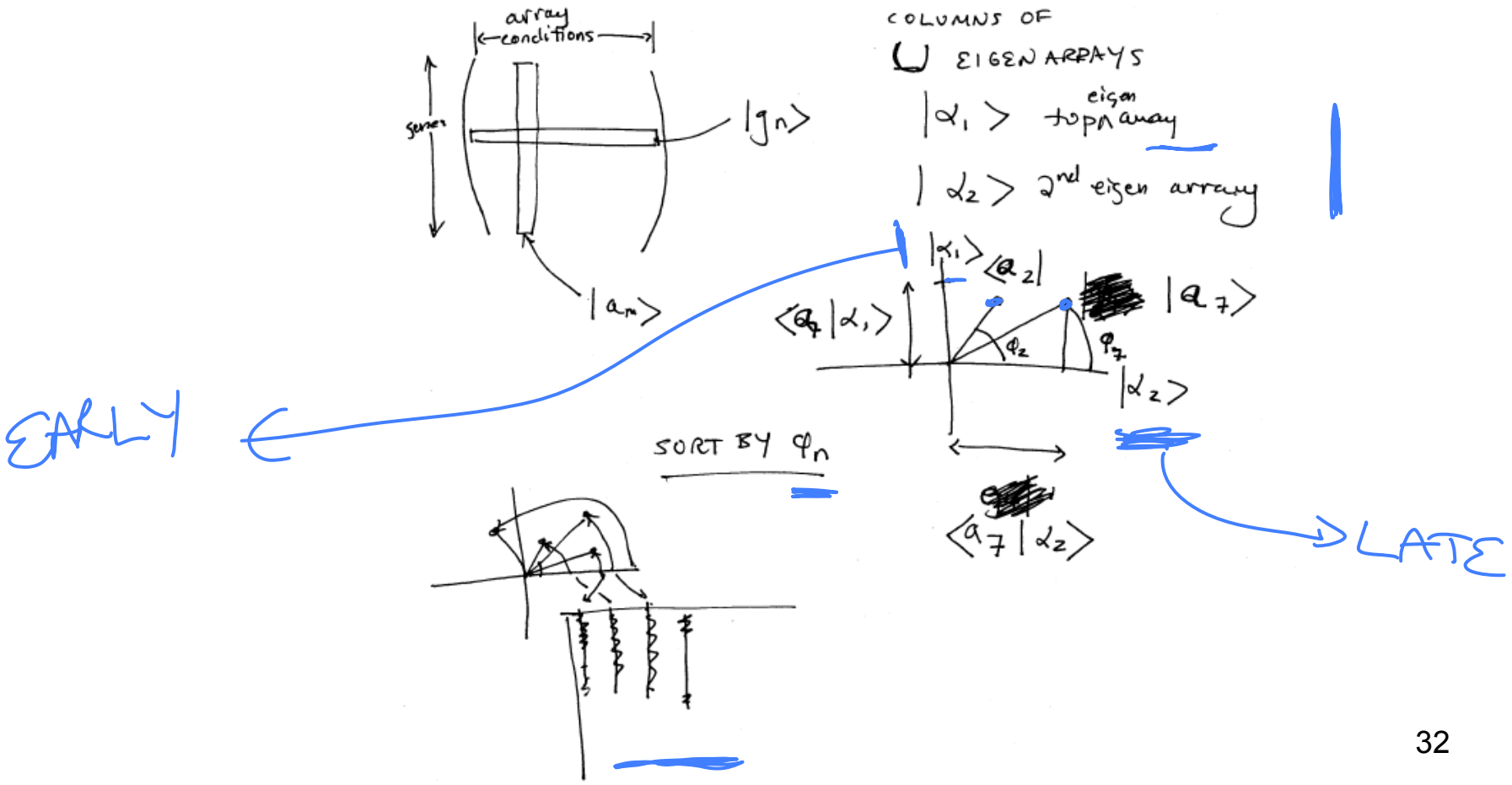


Understanding Eigenconditions (u_i) in terms of PCA on (small) condition-condition correlation matrix



Bra - ket notation

Plotting Experiments in Low Dimension Subspace



Close up on Eigengenes

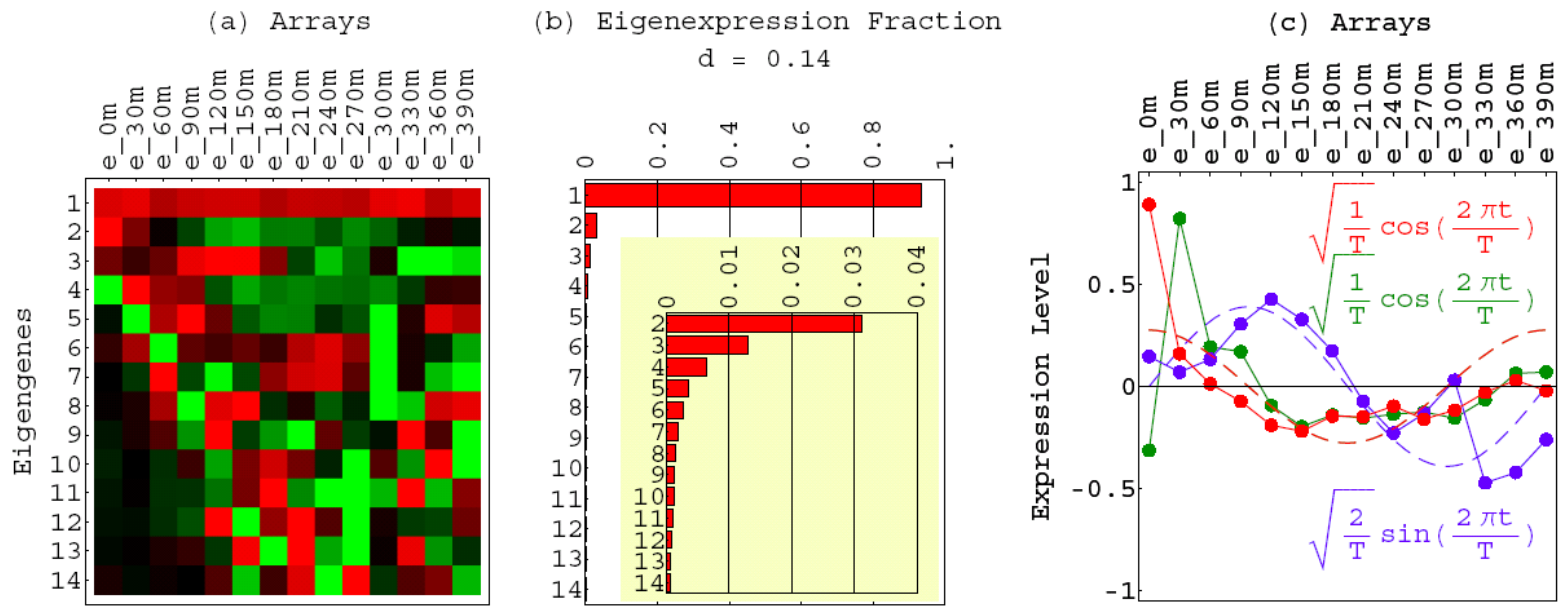
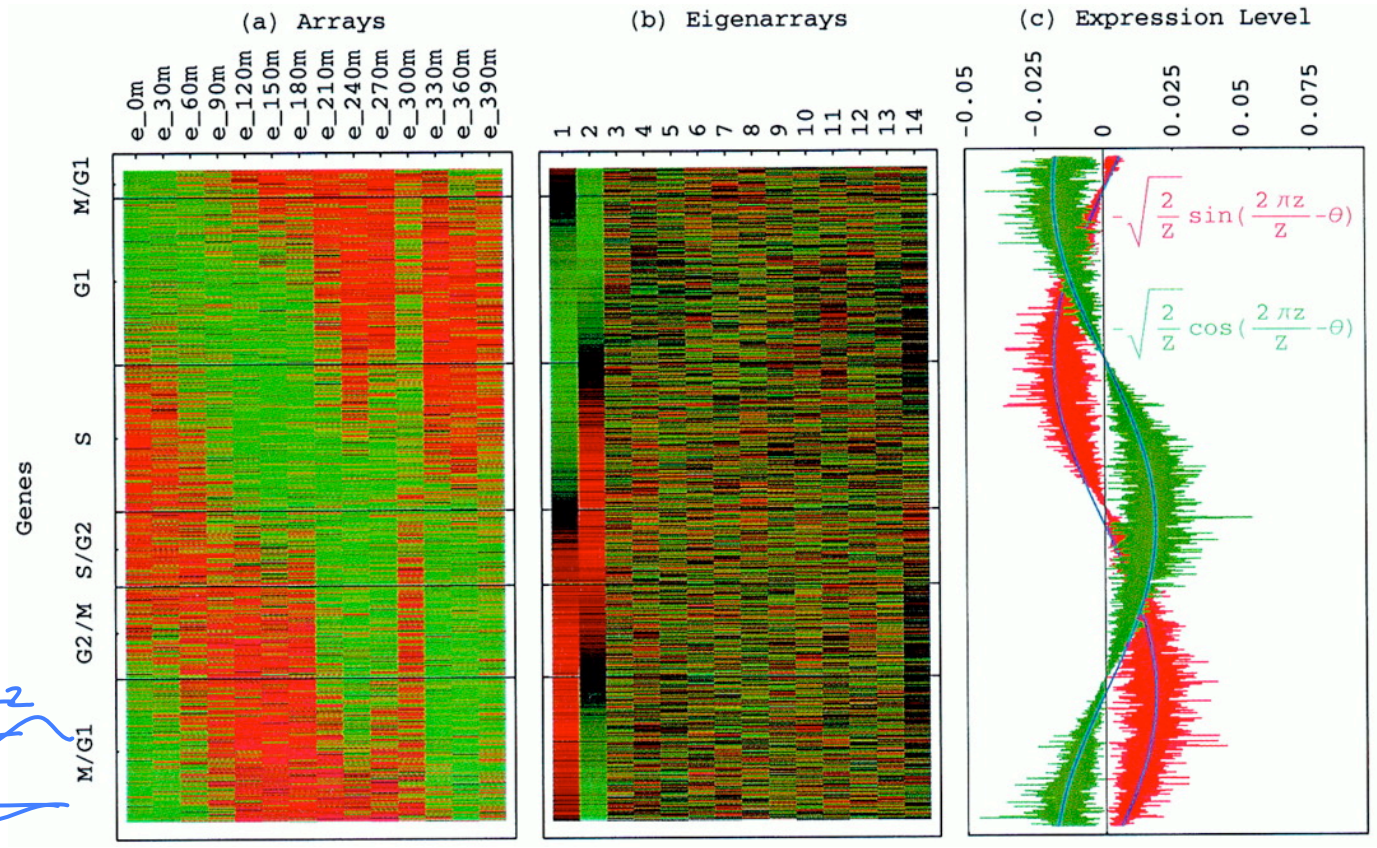


Fig. 8. Elutriation eigengenes. (a) Raster display of $\hat{v}^T$, the expression of 14 eigengenes in 14 arrays, with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene, $|\gamma_1\rangle$. (b) Bar chart of the fraction of eigenexpression p_l of each eigengene $|\gamma_l\rangle$, showing more than 90% of the overall relative expression in $|\gamma_1\rangle$, about 3%, 1.5%, and 0.5% in $|\gamma_2\rangle$, $|\gamma_3\rangle$, and $|\gamma_4\rangle$, respectively, and a low entropy $d = 0.14 \ll 1$. (c) Line-joined graphs of the expression levels of $|\gamma_2\rangle$ (red), $|\gamma_3\rangle$ (blue), and $|\gamma_4\rangle$ (green) in the 14 arrays, and dashed graphs of normalized cosine (blue) and sine (red) of period T.

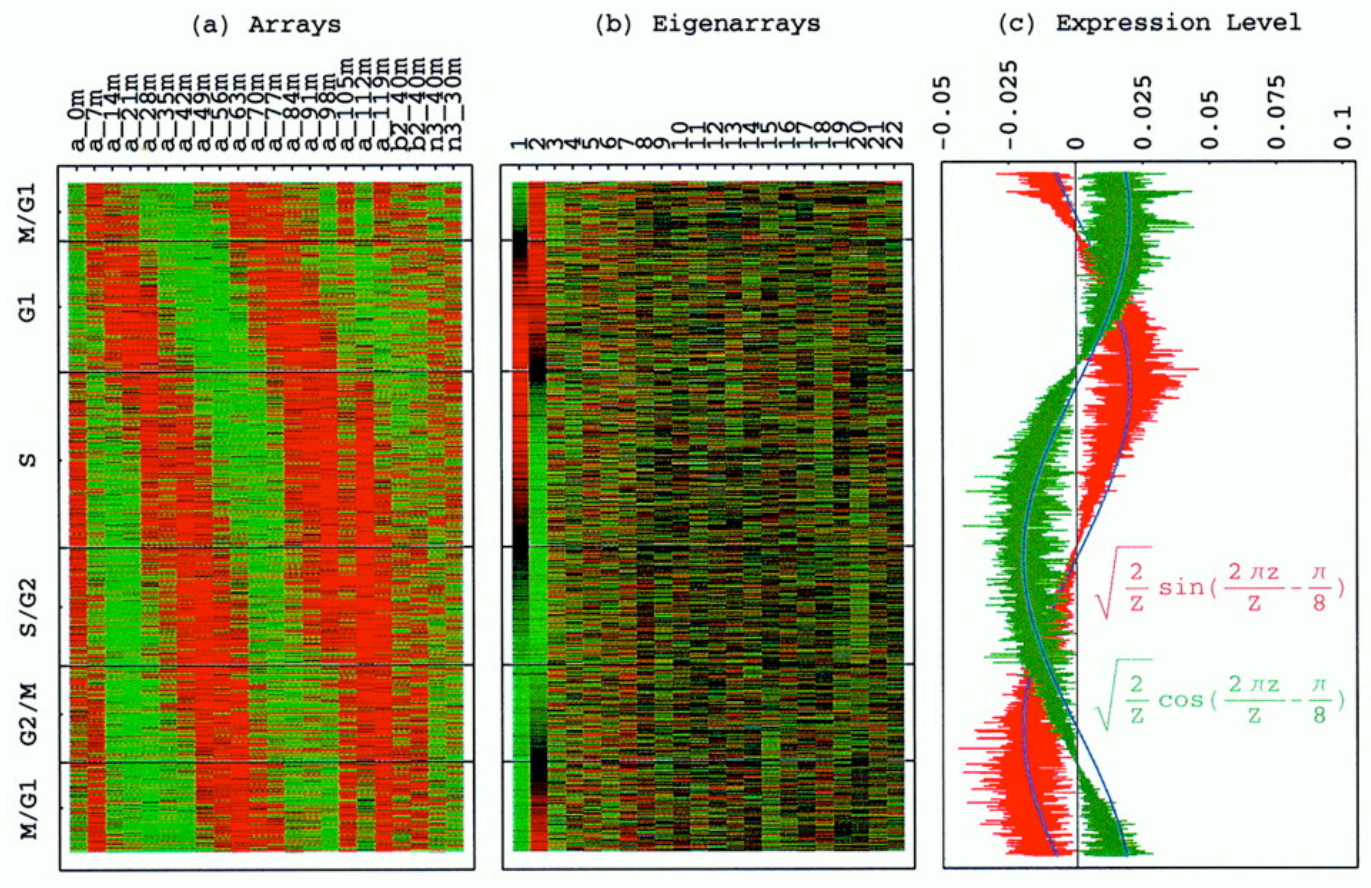
Genes sorted by correlation with top 2 eigengenes



Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Fig. 3. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized elutriation. (*a*) Normalized elutriation expression of the sorted 5,981 genes in the 14 arrays, showing traveling wave of expression. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z \equiv N - 1 = 5{,}980$ and phase $\theta \approx 2\pi/13$ (blue), respectively.

Same thing different experiment: Genes sorted by relative correlation with first two eigengenes for alpha-factor experiment



Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Normalized elutriation expression in the subspace associated with the cell cycle

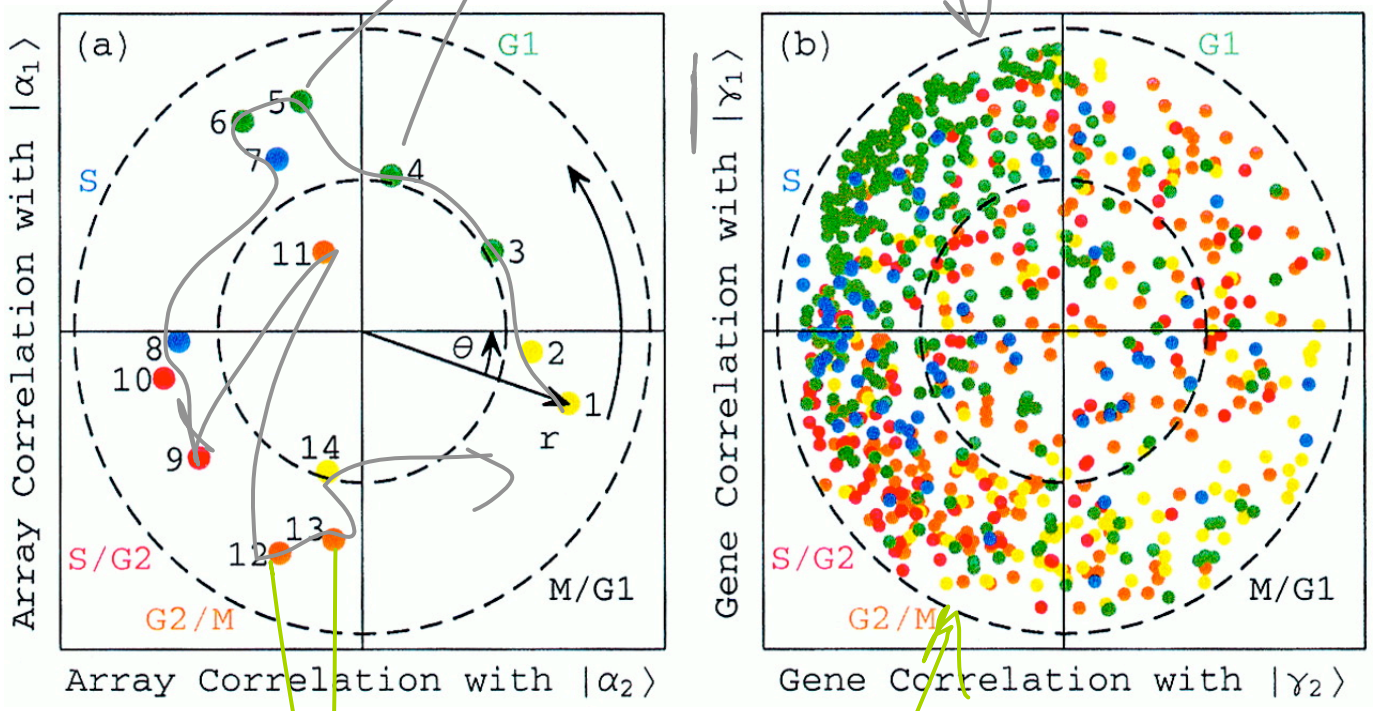
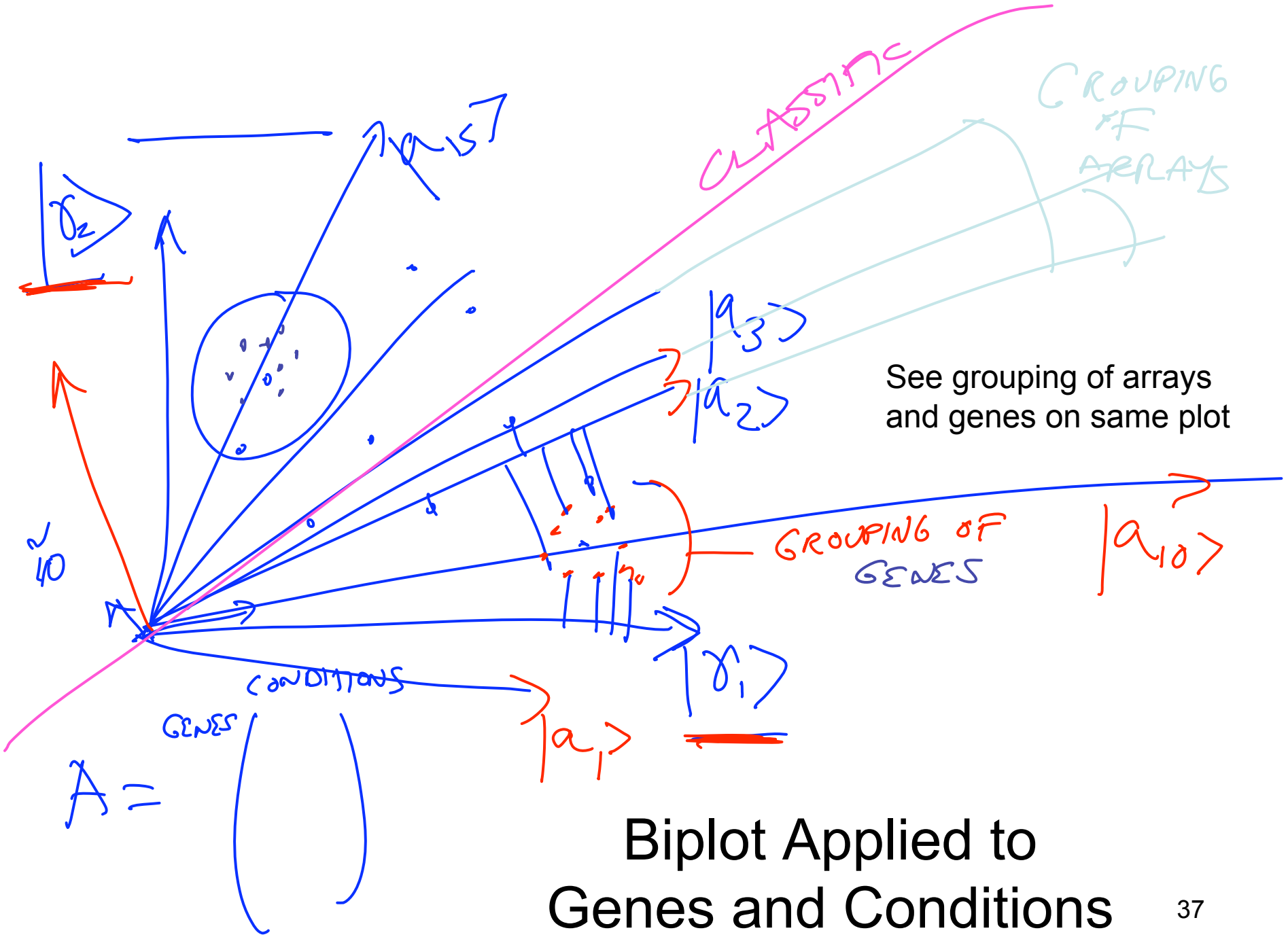


Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (a) Array correlation with $|\alpha_1\rangle_N$ along the *y*-axis vs. that with $|\alpha_2\rangle_N$ along the *x*-axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 784 cell cycle regulated genes, color-coded according to the classification by Spellman *et al.* (3).

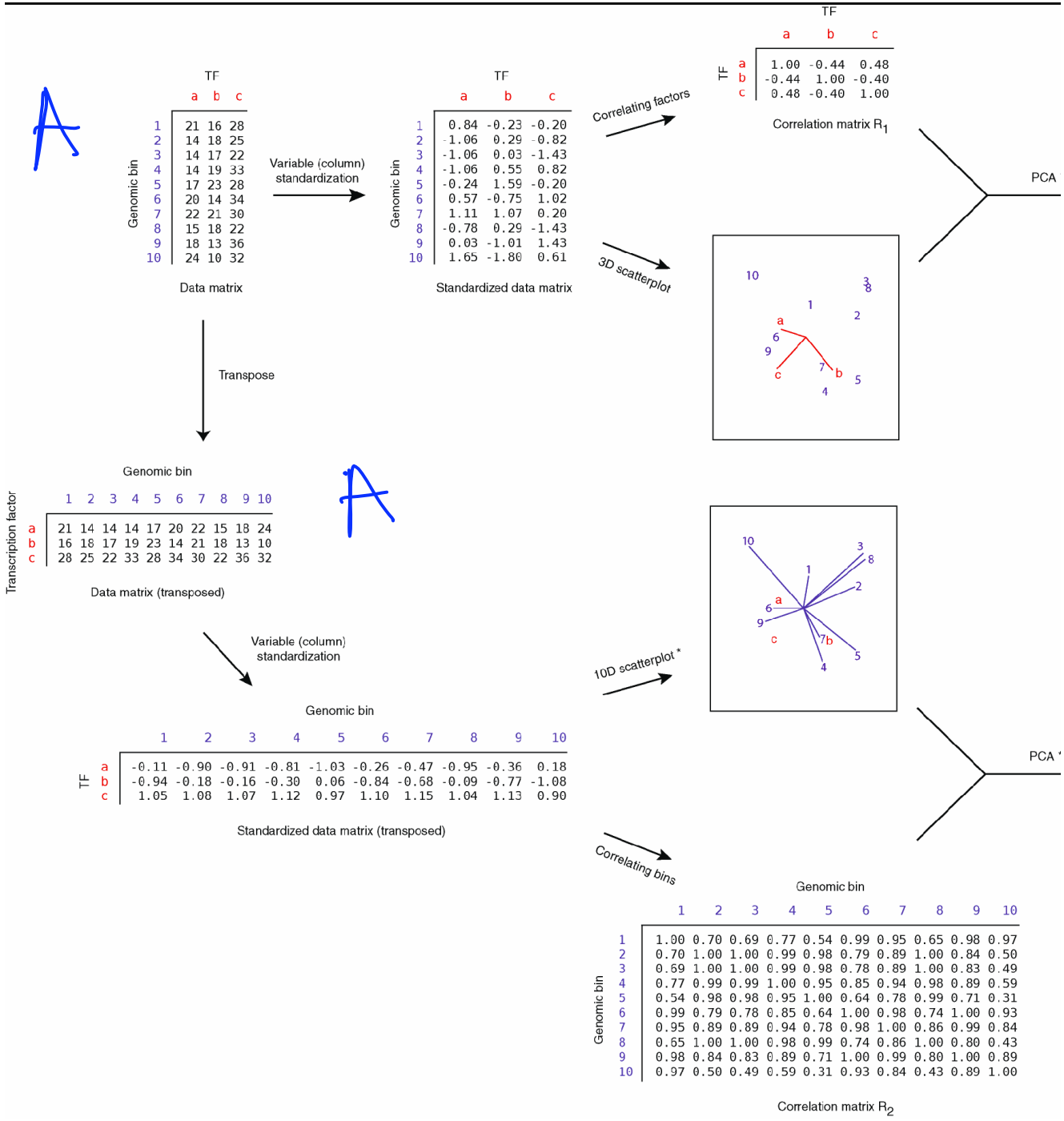
Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106



See grouping of arrays and genes on same plot

Biplot Applied to Genes and Conditions

Biplot Ex

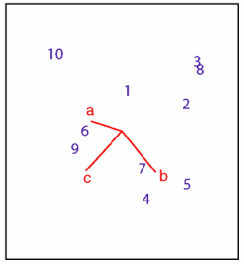

TF
a b c
Genomic bin
1 21 16 28
2 14 18 25
3 14 17 22
4 14 19 33
5 17 23 28
6 20 14 34
7 22 21 30
8 15 18 22
9 18 13 36
10 24 10 32
Data matrix
Variable (column) standardization
1 0.84 -0.23 -0.20
2 -1.06 0.29 -0.82
3 -1.06 0.03 -1.43
4 -1.06 0.55 0.82
5 -0.24 1.59 -0.20
6 0.57 -0.75 1.02
7 1.11 1.07 0.20
8 -0.78 0.29 -1.43
9 0.03 -1.01 1.43
10 1.65 -1.80 0.61
Standardized data matrix
Correlating factors
a b c
a 1.00 -0.44 0.48
b -0.44 1.00 -0.40
c 0.48 -0.40 1.00
Correlation matrix R1
PCA
3D scatterplot
Transpose
Transcription factor
Genomic bin
1 2 3 4 5 6 7 8 9 10
a 21 14 14 14 17 20 22 15 18 24
b 16 18 17 19 23 14 21 18 13 10
c 28 25 22 33 28 34 30 22 36 32
Data matrix (transposed)
Variable (column) standardization
TF
a -0.11 -0.90 -0.91 -0.81 -1.03 -0.26 -0.47 -0.95 -0.36 0.18
b -0.94 -0.18 -0.16 -0.30 0.06 -0.84 -0.68 -0.09 -0.77 -1.08
c 1.05 1.08 1.07 1.12 0.97 1.10 1.15 1.04 1.13 0.90
Standardized data matrix (transposed)
10D scatterplot *
PCA '
Correlating bins
Genomic bin
1 1.00 0.70 0.69 0.77 0.54 0.99 0.95 0.65 0.98 0.97
2 0.70 1.00 1.00 0.99 0.98 0.79 0.89 1.00 0.84 0.50
3 0.69 1.00 1.00 0.99 0.98 0.78 0.89 1.00 0.83 0.49
4 0.77 0.99 0.99 1.00 0.95 0.85 0.94 0.98 0.89 0.59
5 0.54 0.98 0.98 0.95 1.00 0.64 0.78 0.99 0.71 0.31
6 0.99 0.79 0.78 0.85 0.64 1.00 0.98 0.74 1.00 0.93
7 0.95 0.89 0.89 0.94 0.78 0.98 1.00 0.86 0.99 0.84
8 0.65 1.00 1.00 0.98 0.99 0.74 0.86 1.00 0.80 0.43
9 0.98 0.84 0.83 0.89 0.71 1.00 0.99 0.80 1.00 0.89
10 0.97 0.50 0.49 0.59 0.31 0.93 0.84 0.43 0.89 1.00
Correlation matrix R2

Biplot Ex #2

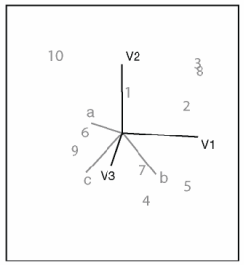
TF

TF	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

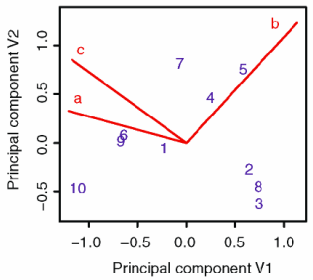
Correlation matrix R_1

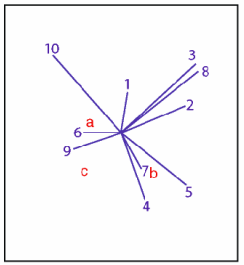


PCA *

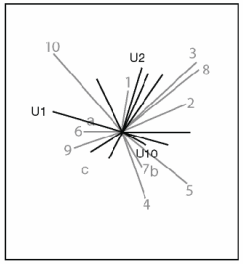


Projection *

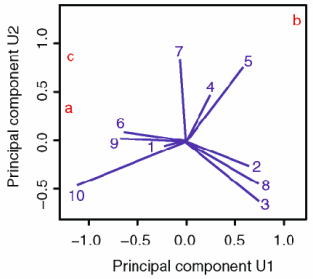




PCA *



Projection *



The same rank-2 approximation of the original data matrix

Genomic bin

2	3	4	5	6	7	8	9	10
.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

Correlation matrix R_2

*

10D scatterplots are used here for illustrative purpose only.

PCA: the correlation matrix is eigen-decomposed; then the principal components are added to the original space.

Projection: the points and axes in the original space are projected onto the plane defined by the top two principal components.

Biplot Ex #3

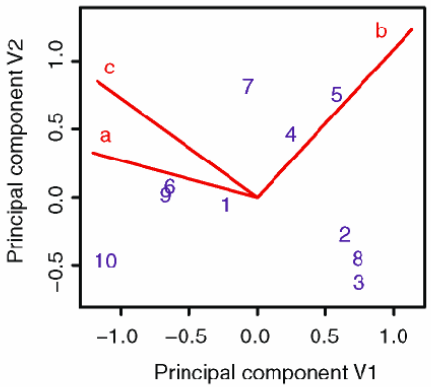
$$A\mathbf{v}_i = s_i\mathbf{u}_i$$

$$A^T\mathbf{u}_i = s_i\mathbf{v}_i$$

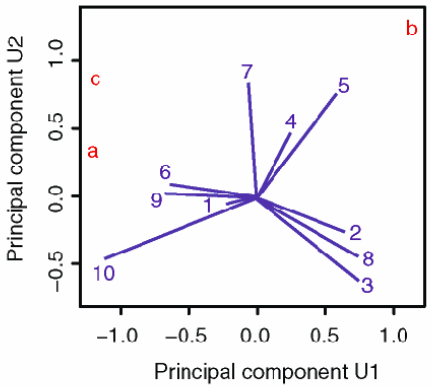
Assuming s=1,

$A\mathbf{v}_i = \mathbf{u}_i$

$A^T\mathbf{u}_i = \mathbf{v}_i$



The same rank-2 approximation of the original data matrix

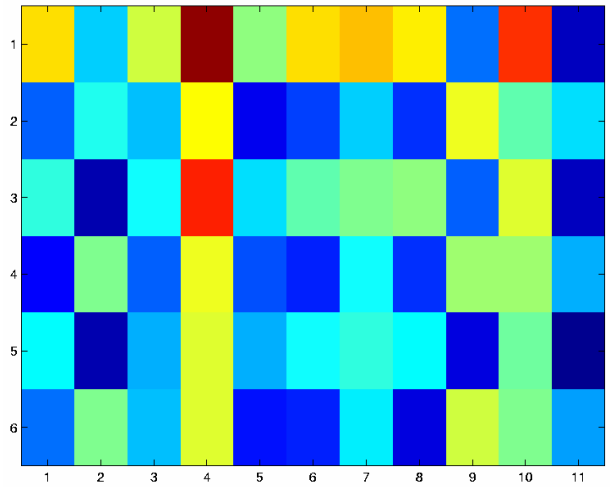


Spectral Biclustering

Biclustering to associate particular genes with certain phenotypes

Matrix of raw data

Genes

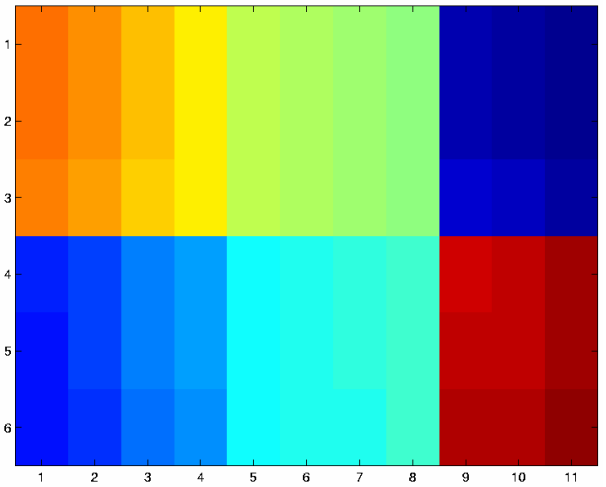


Conditions

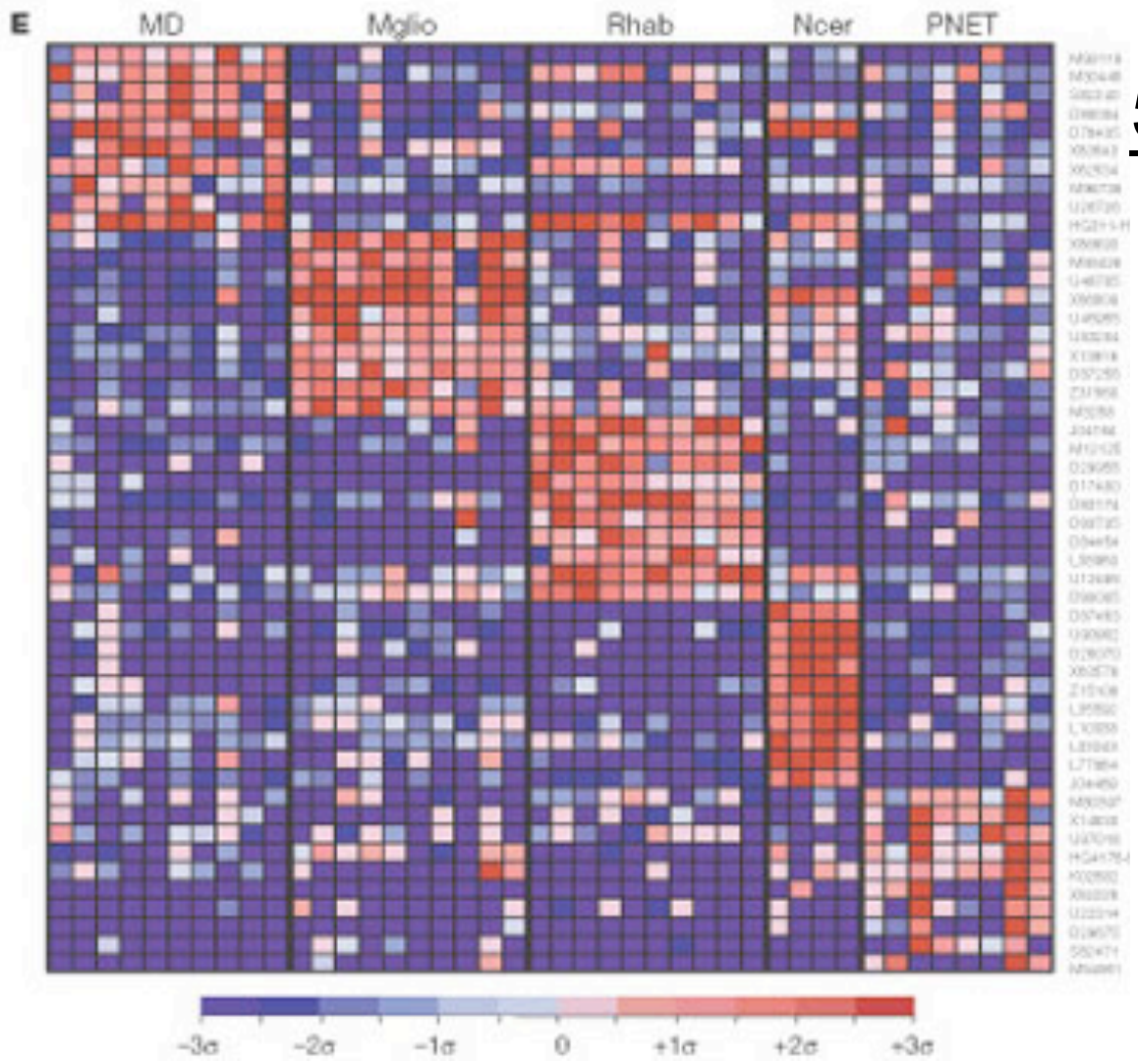
?

Shuffled Matrix
(containing checkerboard "biclusters" of conditions with marker genes)

Reordered Genes
(Sorted according to a classification vector)



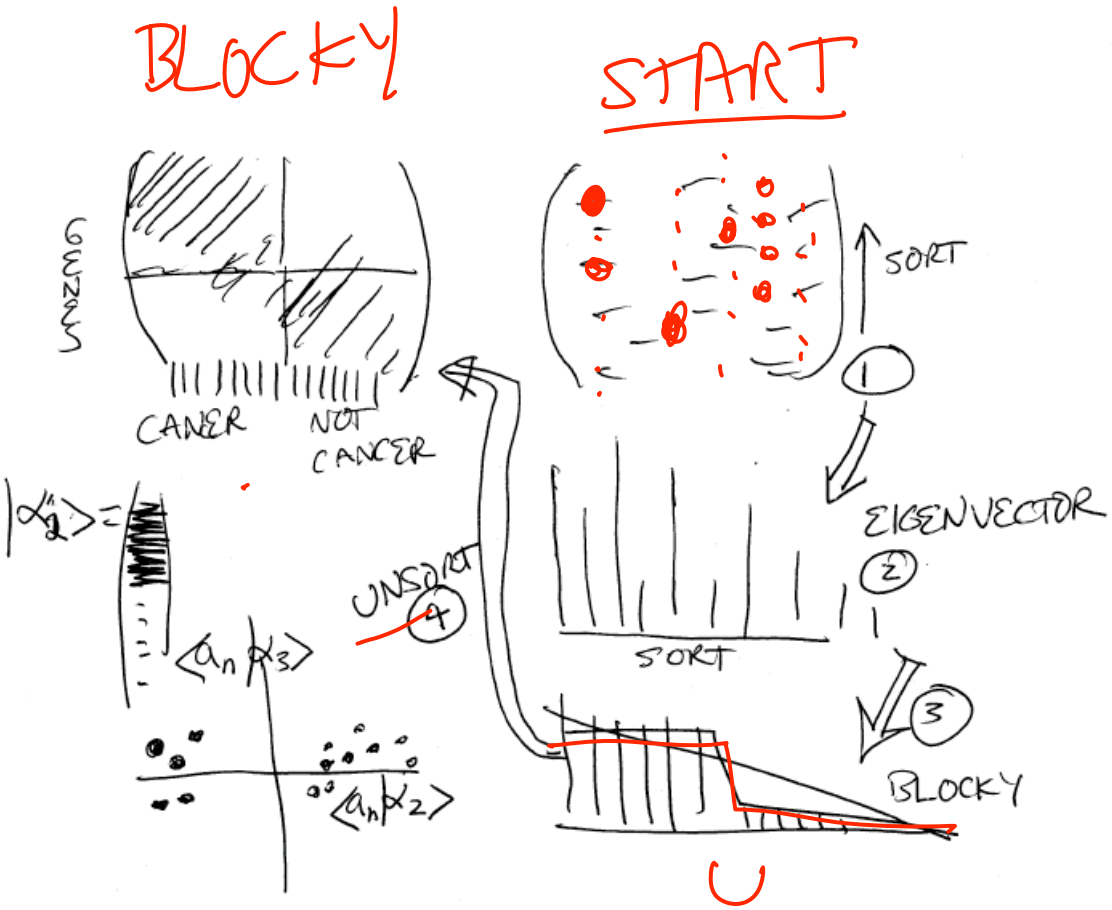
Reordered Conditions
(Sorted according to a classification vector)

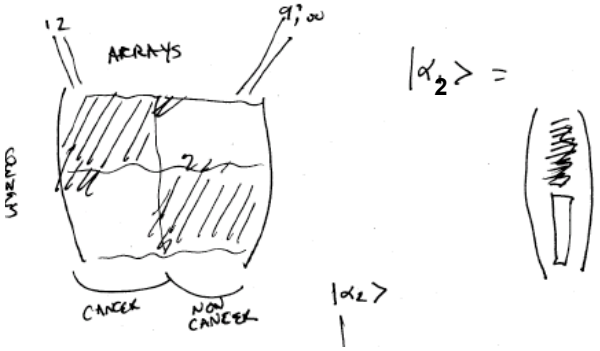


5 types of brain tumors

Pomeroy et. al. , Nature 415 (2002) 436
Prediction of central nervous system embryonal tumor outcome based on gene expression

Intuition on Identification of Blocky Matrices





Gene partition vector

$$Ax = y$$

	tumor 1				tumor 2				tumor 3			
Gene cluster 1	8	8	8	8	7	7	7	7	3	3	3	
	8	8	8	8	7	7	7	7	3	3	3	
	8	8	8	8	7	7	7	7	3	3	3	
Gene cluster 2	6	6	6	6	4	4	4	4	5	5	5	
	6	6	6	6	4	4	4	4	5	5	5	
	6	6	6	6	4	4	4	4	5	5	5	

$$\begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}$$

Tissue partition vector

$$A^T y = x'$$

$$\begin{pmatrix} 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 3 & 3 & 3 & 5 & 5 & 5 \\ 3 & 3 & 3 & 5 & 5 & 5 \\ 3 & 3 & 3 & 5 & 5 & 5 \end{pmatrix} \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix} = \begin{pmatrix} a' \\ a' \\ a' \\ a' \\ b' \\ b' \\ b' \\ b' \\ c' \\ c' \\ c' \end{pmatrix}$$

Biclustering by SVD

$$Ax = y$$

$$\begin{pmatrix} 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 8 & 8 & 8 & 8 & 7 & 7 & 7 & 7 & 3 & 3 & 3 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \\ 6 & 6 & 6 & 6 & 4 & 4 & 4 & 4 & 5 & 5 & 5 \end{pmatrix} \begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}$$

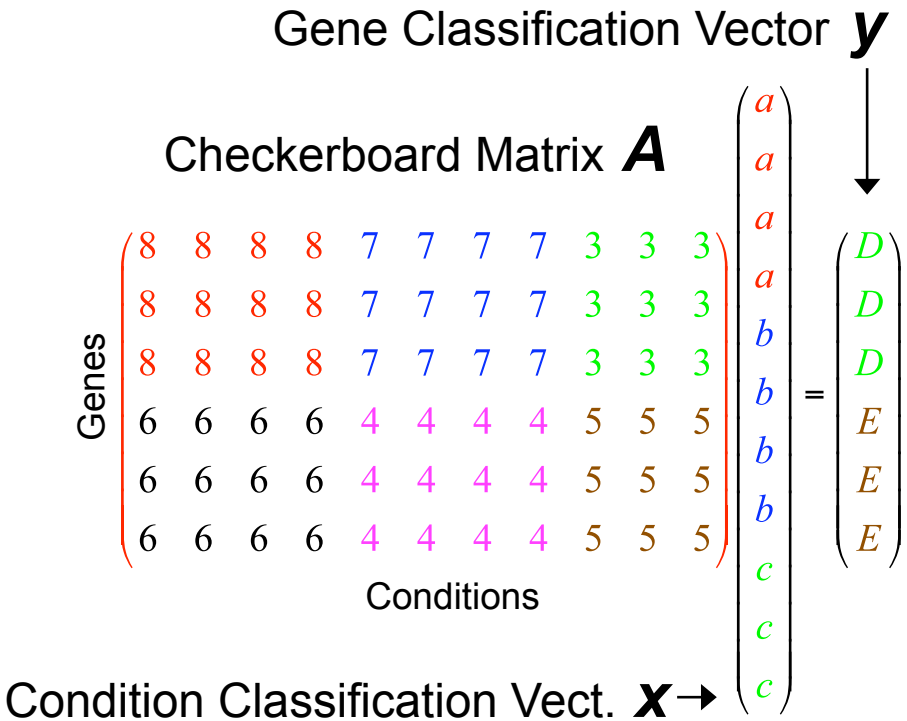
$$A^T y = x'$$

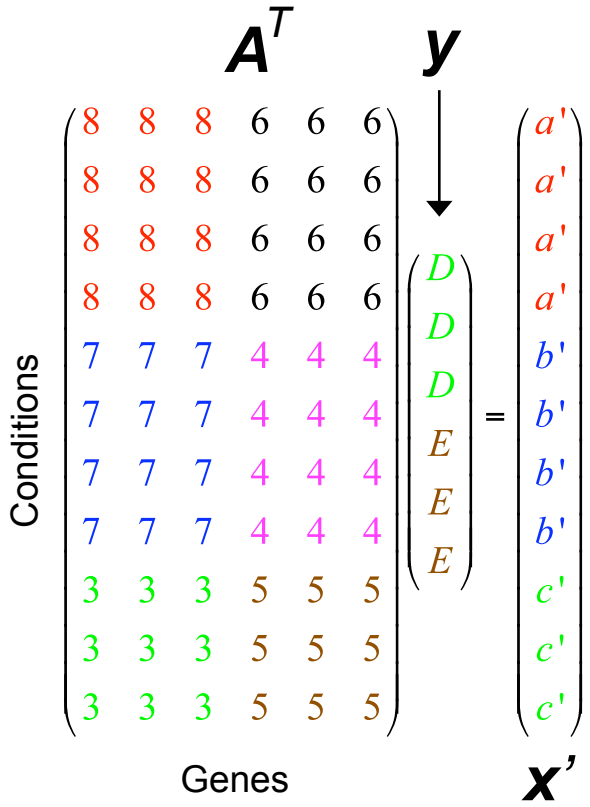
$$\begin{pmatrix} 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 8 & 8 & 8 & 6 & 6 & 6 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 7 & 7 & 7 & 4 & 4 & 4 \\ 3 & 3 & 3 & 5 & 5 & 5 \\ 3 & 3 & 3 & 5 & 5 & 5 \\ 3 & 3 & 3 & 5 & 5 & 5 \end{pmatrix} \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix} = \begin{pmatrix} a' \\ a' \\ a' \\ a' \\ b' \\ b' \\ b' \\ b' \\ c' \\ c' \\ c' \end{pmatrix}$$

$$A^T A x = x'$$

Identify checkerboard matrices by their action on classification vectors: Formulation as "eigenproblem"

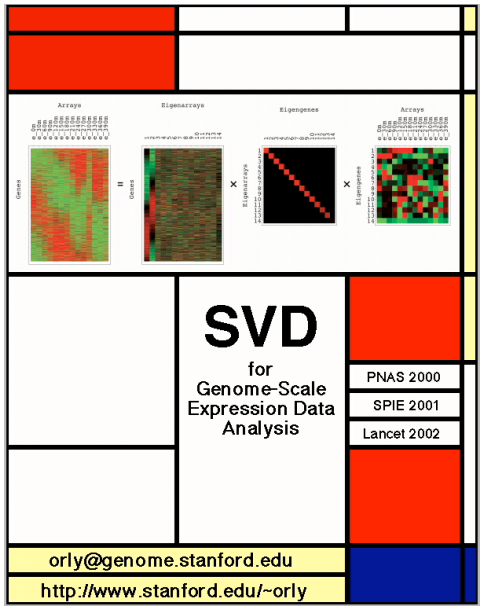
$$A^T A\, x = x'$$
$$A\, A^T y = y'$$





SVD to Solve Eigenproblem

[Botstein]



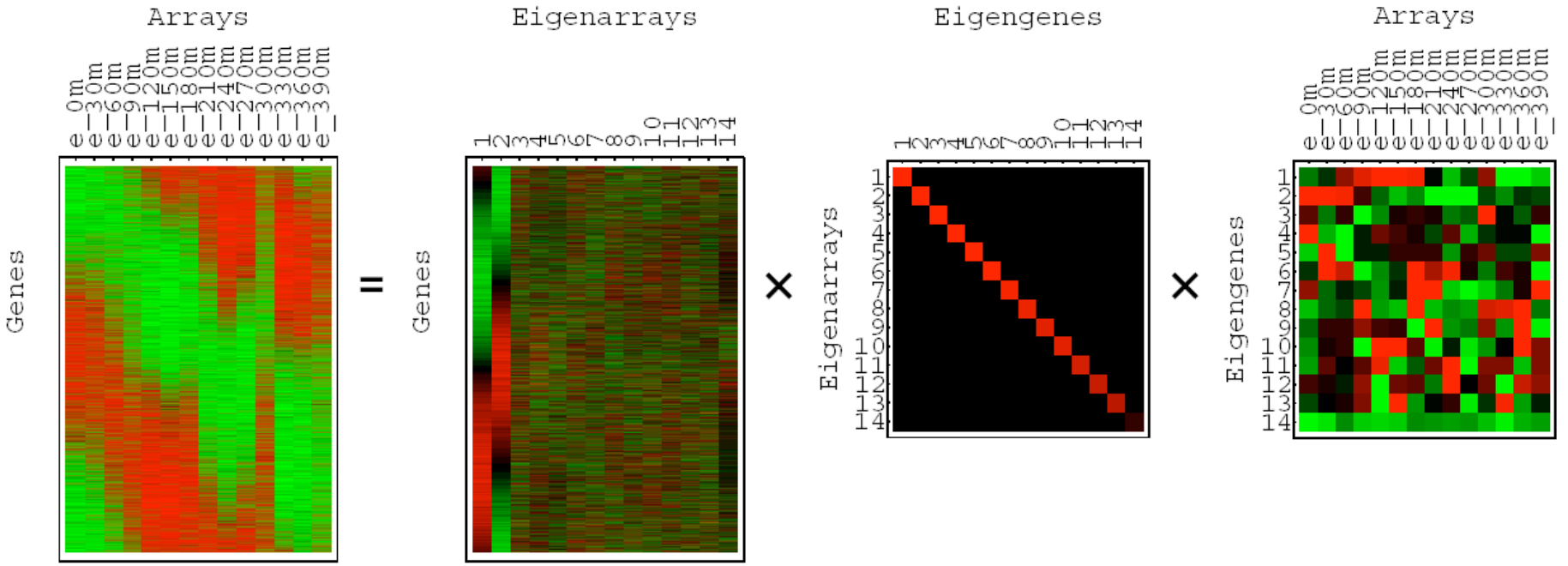
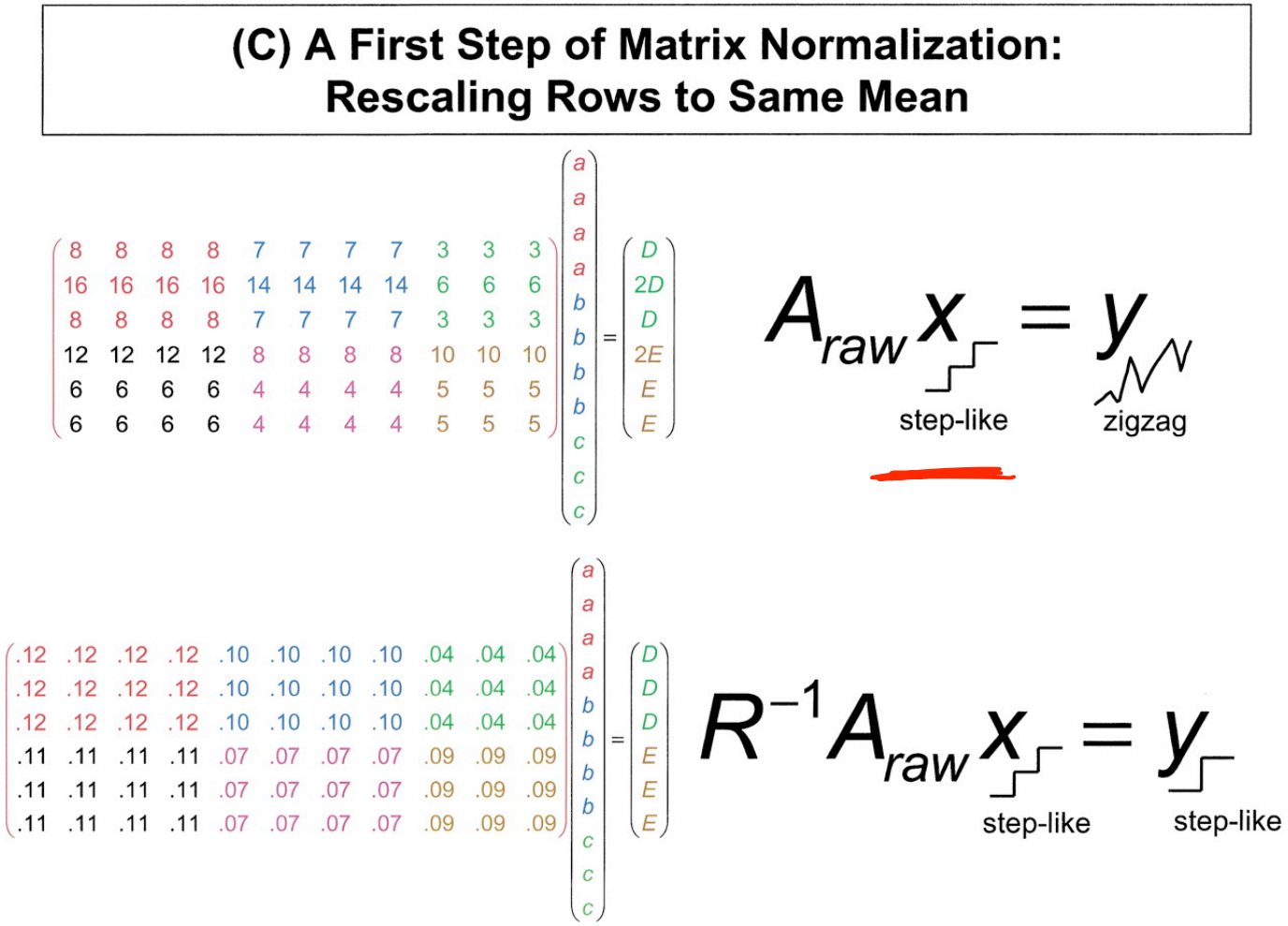


Figure 1. Overview of important parts of the biclustering process



Yuval Kluger et al. Genome Res. 2003; 13: 703-716

Gene partition with noisy data

$$\begin{pmatrix} 8 & 9 & 7 & 8 & 8 & 9 & 4 & 7 & 6 & 1 & 2 \\ 9 & 6 & 8 & 9 & 7 & 5 & 8 & 8 & 1 & 5 & 3 \\ 7 & 9 & 9 & 7 & 6 & 7 & 9 & 6 & 2 & 3 & 4 \\ 4 & 8 & 9 & 3 & 1 & 6 & 6 & 3 & 2 & 7 & 6 \\ 8 & 5 & 2 & 9 & 7 & 3 & 2 & 4 & 8 & 6 & 1 \\ 6 & 5 & 7 & 6 & 4 & 3 & 4 & 5 & 5 & 2 & 8 \end{pmatrix} \begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}$$

Normalization Rescales Rows and Columns to Same Means

$$\begin{pmatrix} .12 & .12 & .12 & .12 & .10 & .10 & .10 & .10 & .04 & .04 & .04 \\ .12 & .12 & .12 & .12 & .10 & .10 & .10 & .10 & .04 & .04 & .04 \\ .12 & .12 & .12 & .12 & .10 & .10 & .10 & .10 & .04 & .04 & .04 \\ .11 & .11 & .11 & .11 & .07 & .07 & .07 & .07 & .09 & .09 & .09 \\ .11 & .11 & .11 & .11 & .07 & .07 & .07 & .07 & .09 & .09 & .09 \\ .11 & .11 & .11 & .11 & .07 & .07 & .07 & .07 & .09 & .09 & .09 \end{pmatrix} \begin{pmatrix} a \\ a \\ a \\ a \\ b \\ b \\ b \\ b \\ c \\ c \\ c \end{pmatrix} = \begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}$$

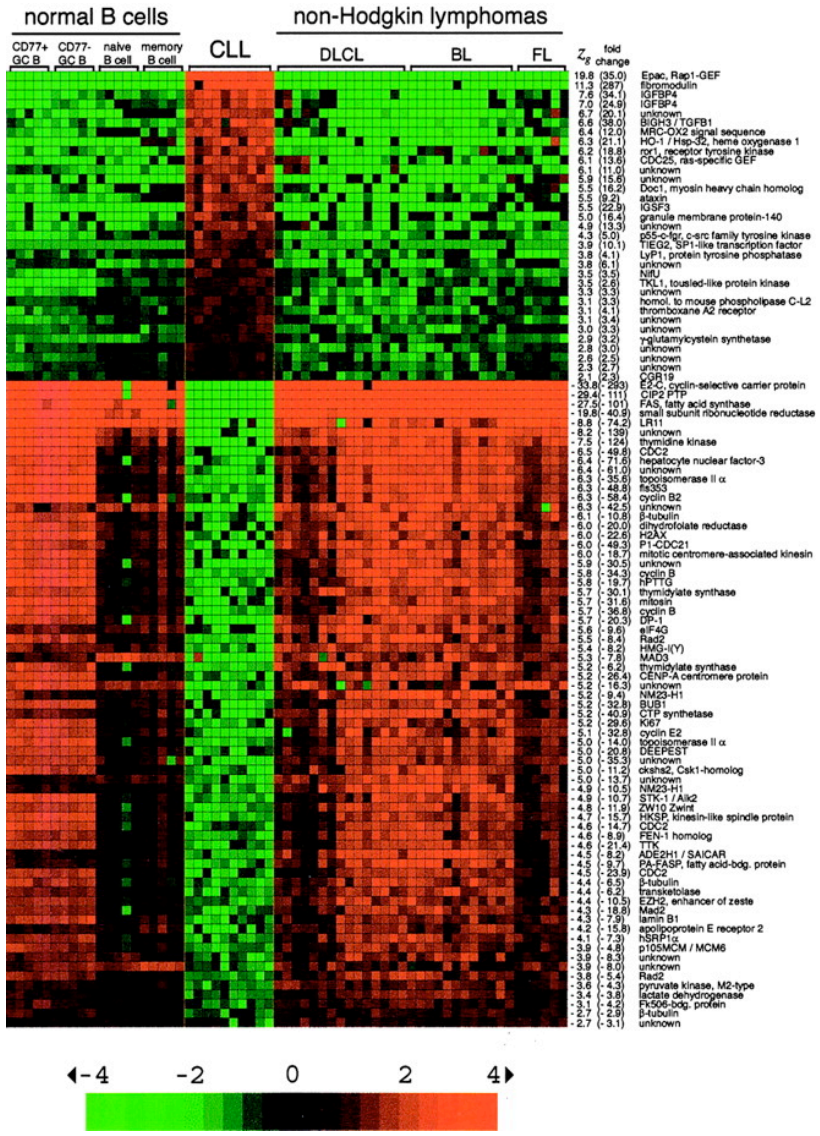
$$R^{-1}Ax = y$$

Rescale columns

$$C^{-1} A^T y = x'$$

$$\begin{pmatrix}
.14 & .28 & .14 & .214 & .107 & .107 \\
.14 & .28 & .14 & .214 & .107 & .107 \\
.14 & .28 & .14 & .214 & .107 & .107 \\
.14 & .28 & .14 & .214 & .107 & .107 \\
.16 & .32 & .16 & .18 & .09 & .09 \\
.16 & .32 & .16 & .18 & .09 & .09 \\
.16 & .32 & .16 & .18 & .09 & .09 \\
.16 & .32 & .16 & .18 & .09 & .09 \\
.09 & .18 & .09 & .312 & .156 & .156 \\
.09 & .18 & .09 & .312 & .156 & .156 \\
.09 & .18 & .09 & .312 & .156 & .156
\end{pmatrix}
\begin{pmatrix} D \\ D \\ D \\ E \\ E \\ E \end{pmatrix}
=
\begin{pmatrix} a' \\ a' \\ a' \\ a' \\ b' \\ b' \\ b' \\ b' \\ c' \\ c' \\ c' \end{pmatrix}$$

Representative Cancer Data set



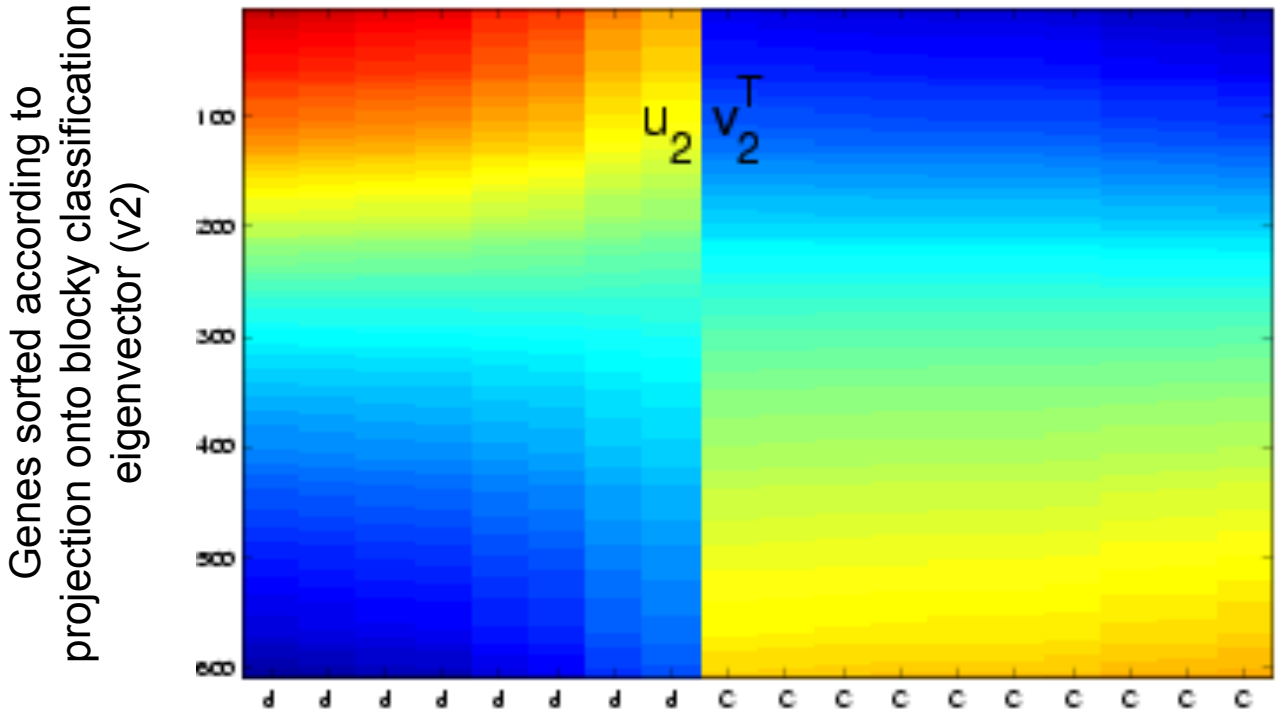
- Lymphoma Data from Dalla-Favera et al. at Columbia
- Informatics from Stolovitzky & Califano at IBM
- Supervised learning some identified characteristic genes associated with different types of lymphoma

Results on Representative Cancer Data set

A

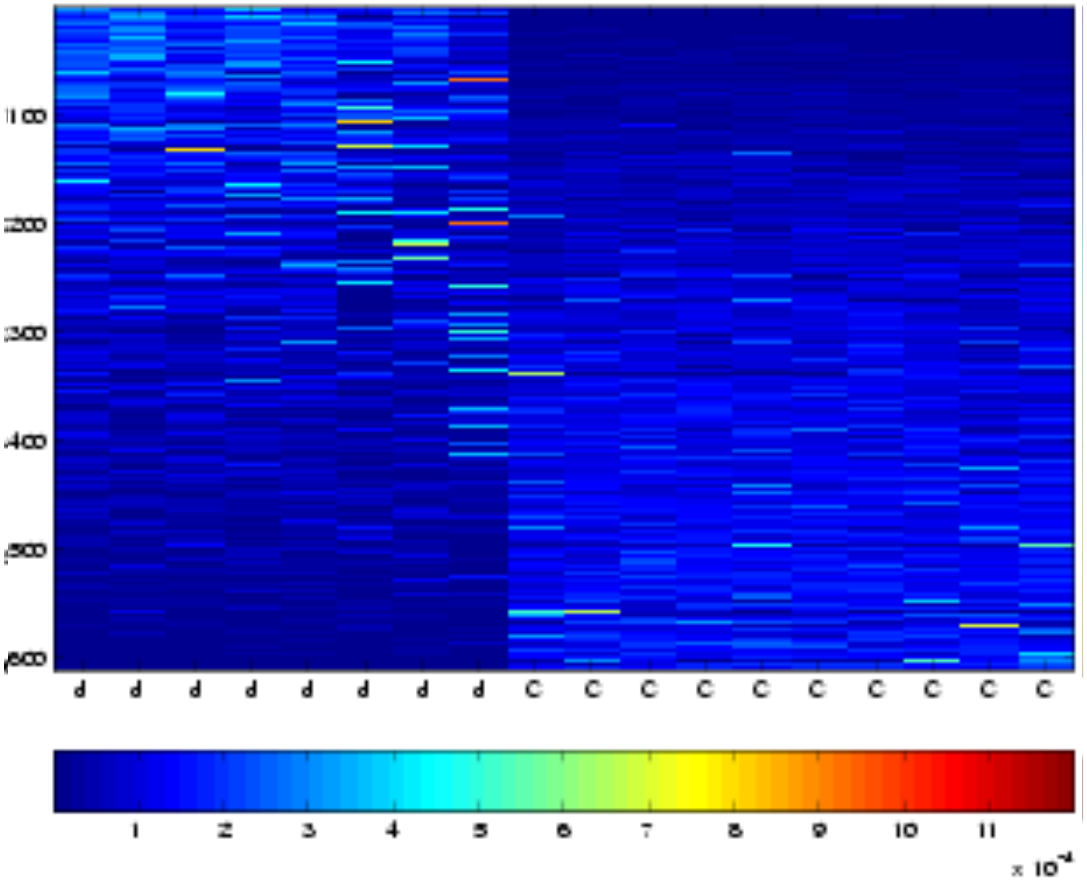
RANK 2

Matrix values represent outer products of two blocky classification eigenvectors

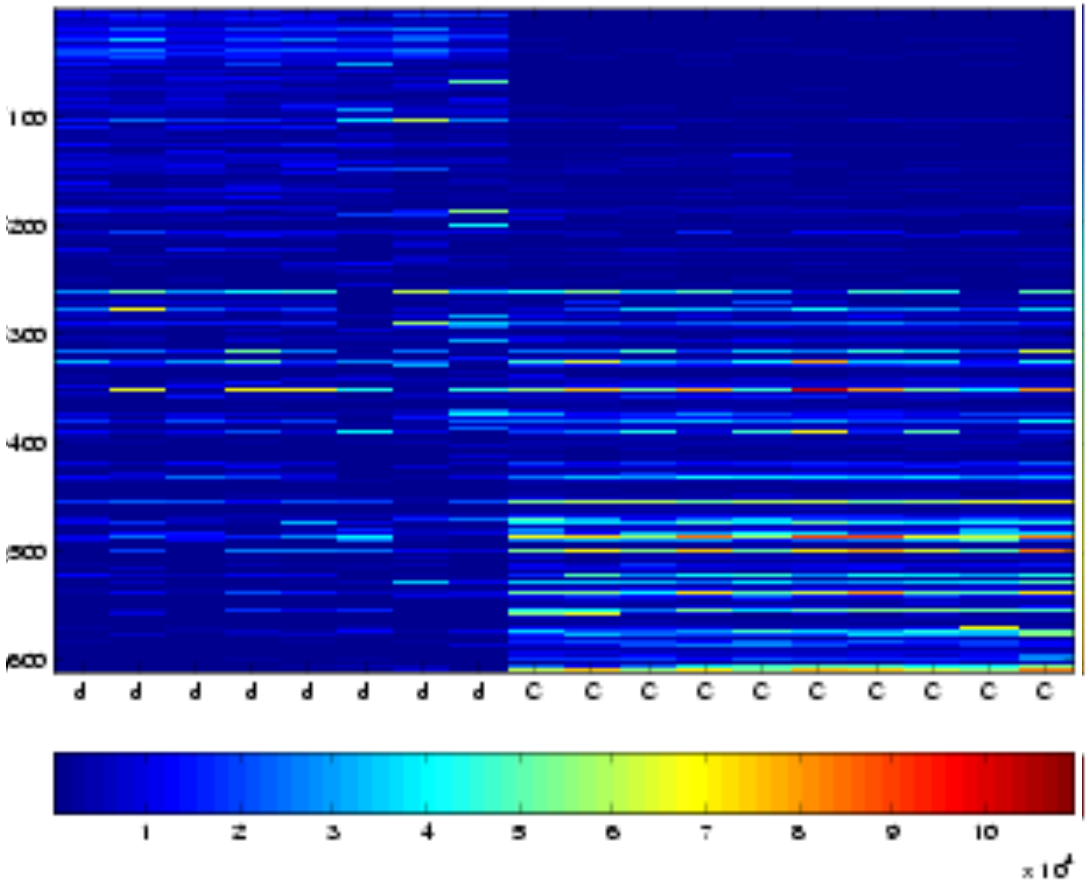


Patients (samples) sorted according to projection onto blocky classification eigenvector (u2)

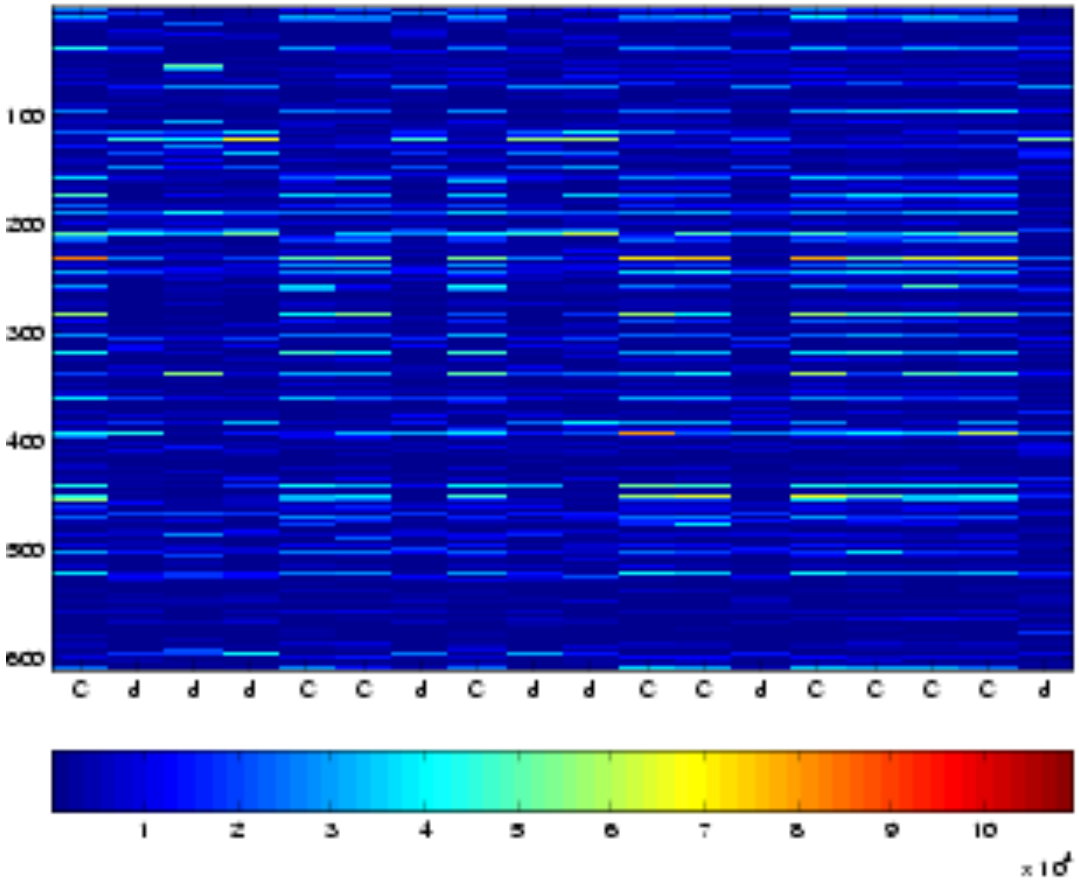
Actual Data with Normalization and Sorting



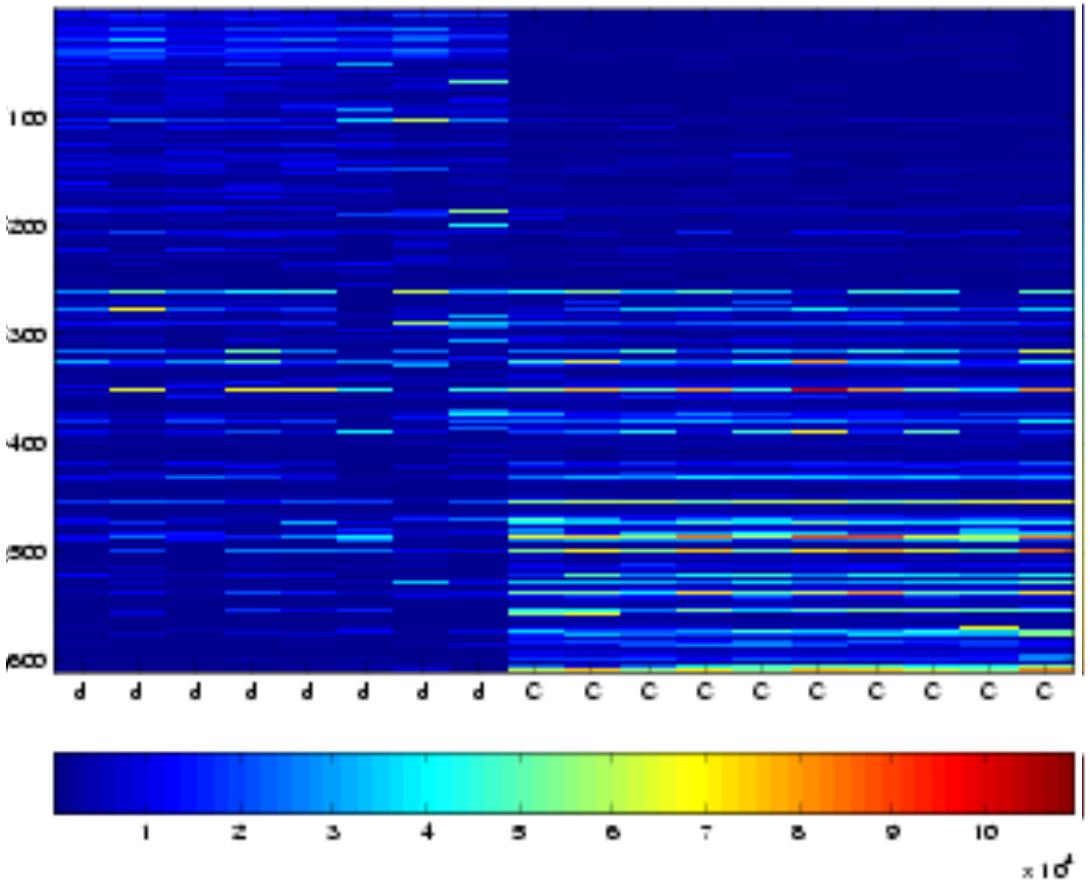
Actual Data just with Sorting
(no normalization)



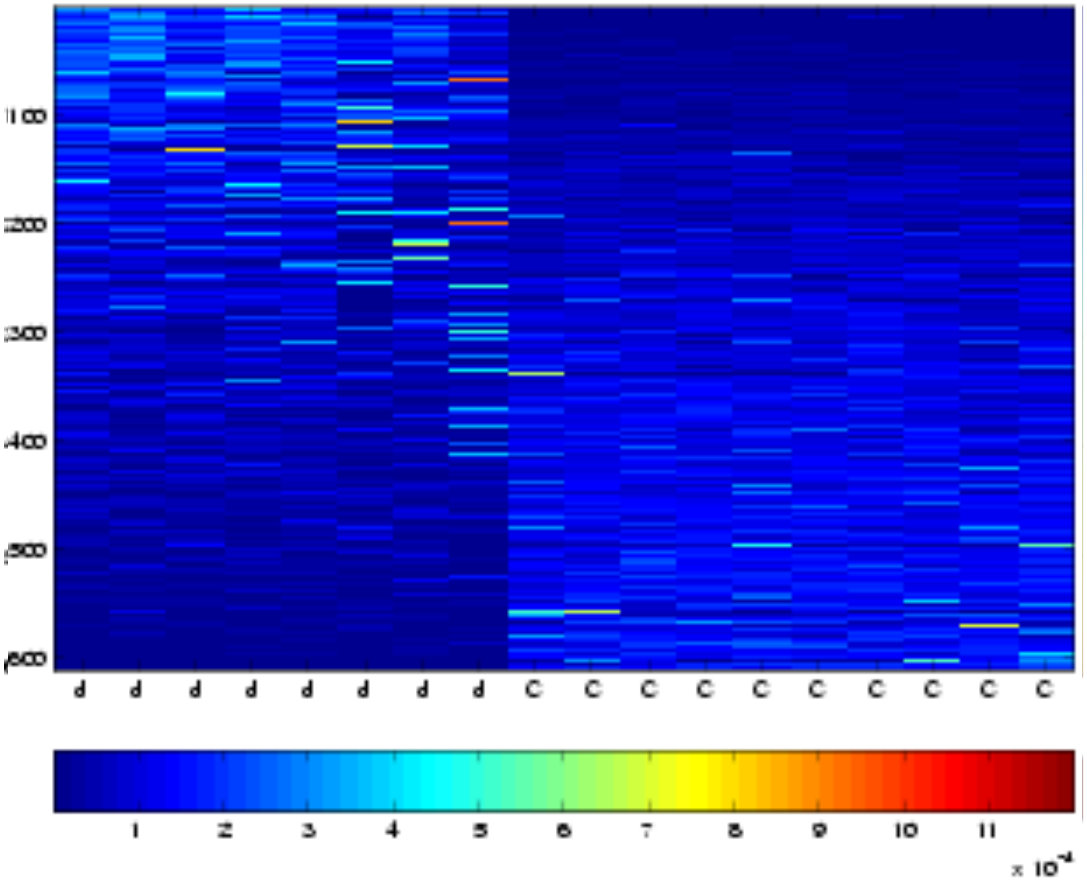
Actual Data
(no normalization or sorting)



Actual Data just with Sorting
(no normalization)

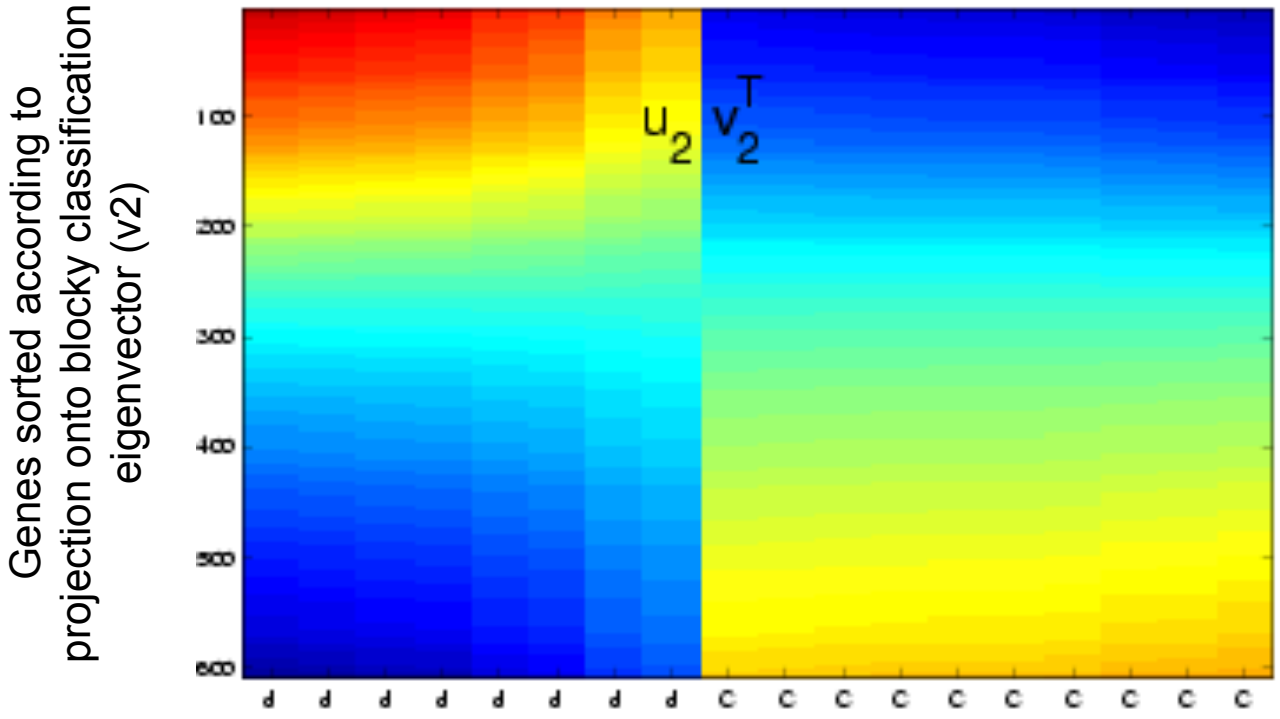


Actual Data with Normalization and Sorting

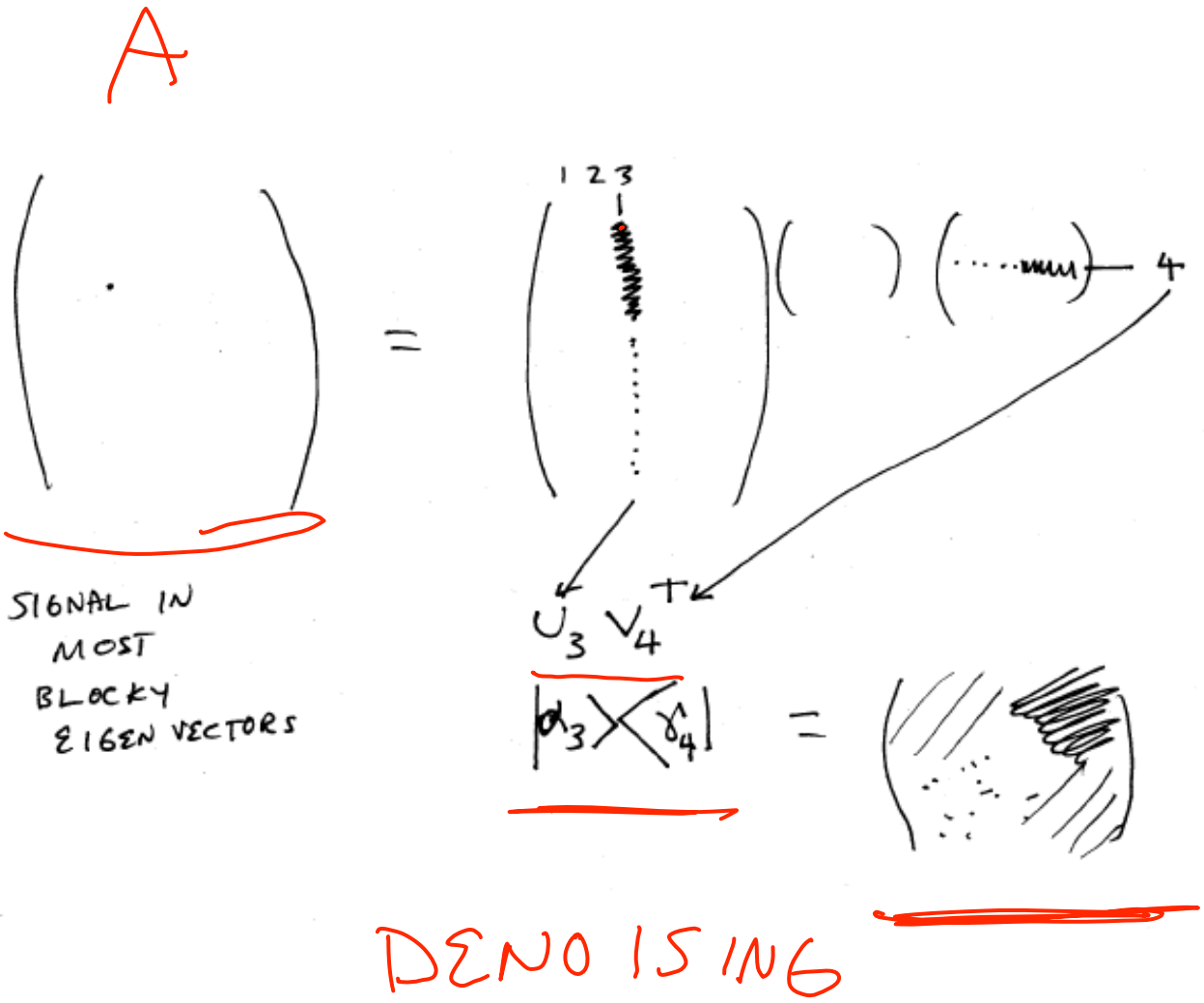


Just signal from top classification eigenvectors

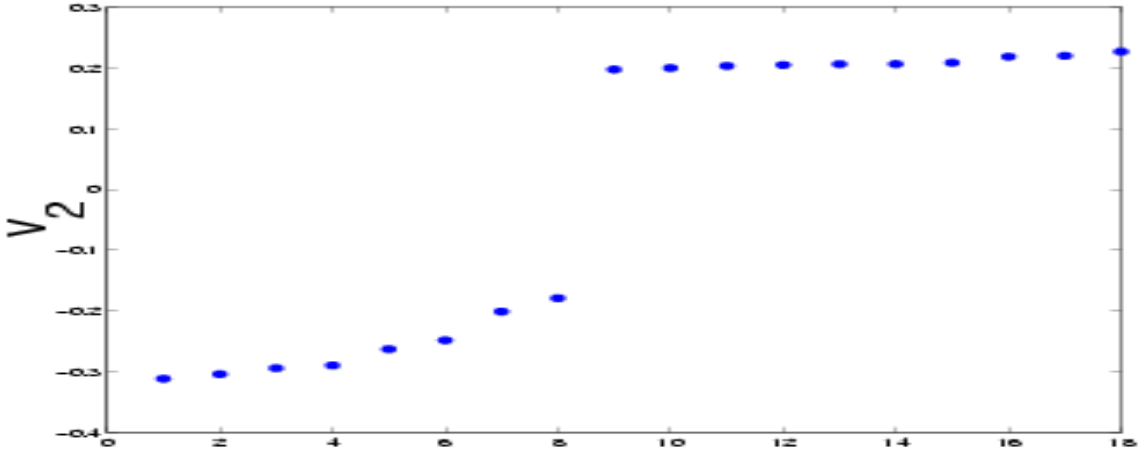
Matrix values represent outer products of two blocky classification eigenvectors

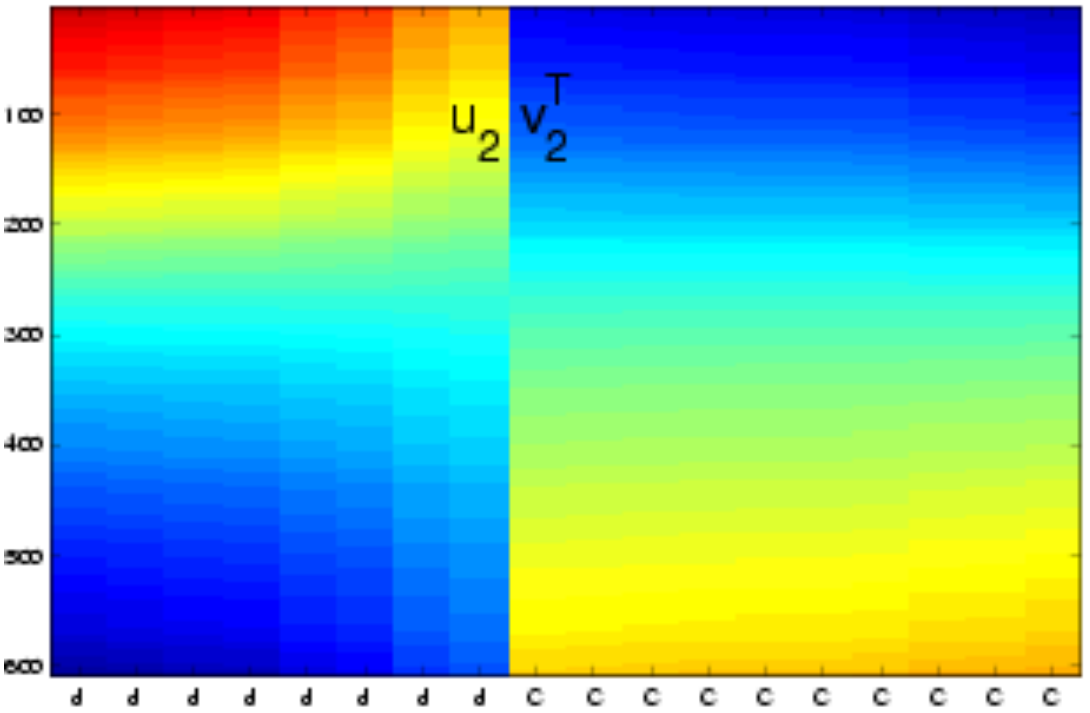


Low Dimension Representation

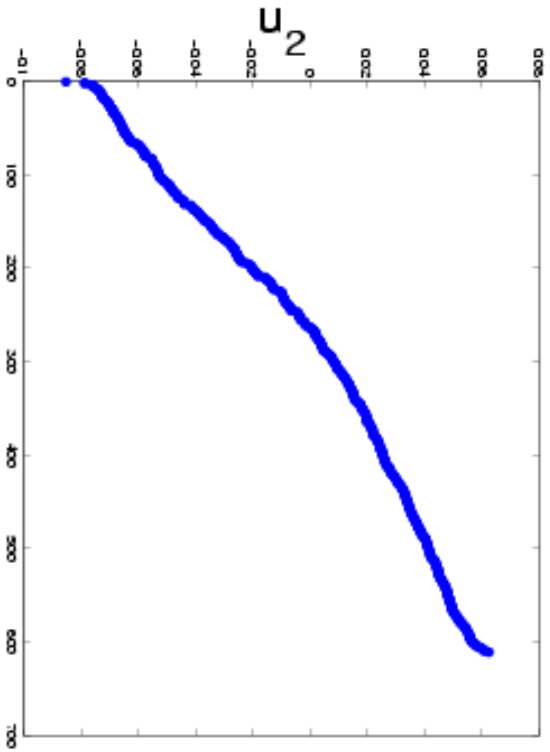


Actual Values of Projections onto Classification Eigenvectors





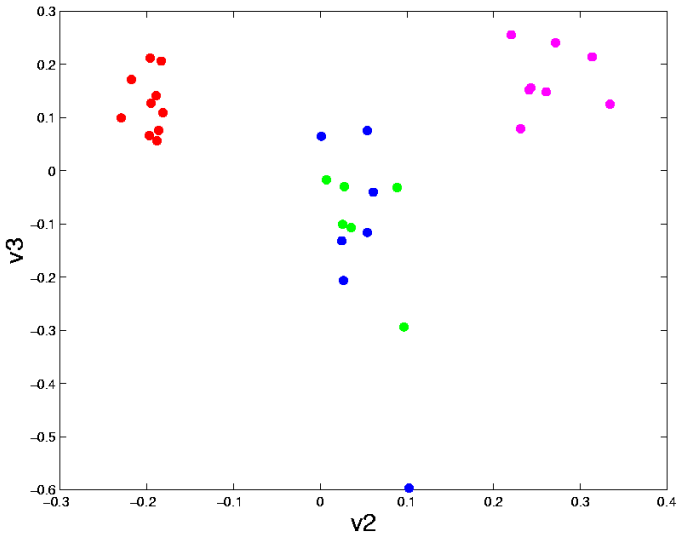
Genes sorted according to projection onto blocky classification eigenvector (v2)



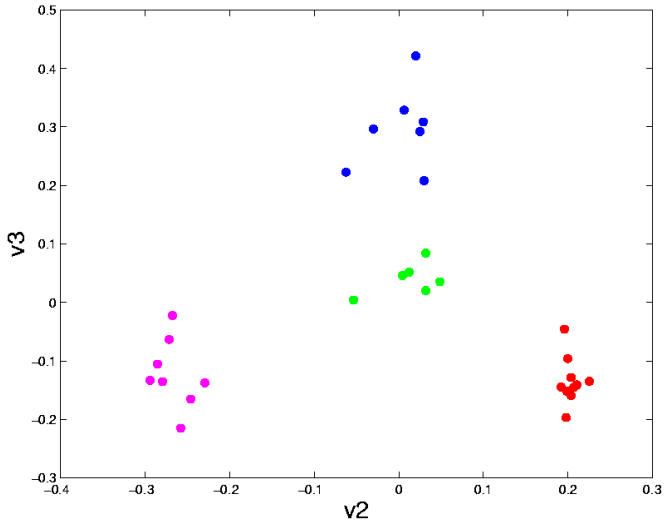
Patients (samples) sorted according to projection onto blocky classification eigenvector (u2)

Classification of Cancers Based on Projection onto two top classification eigenvectors: Better with Normalization

Straight SVD



Normalized ("bistochastization")

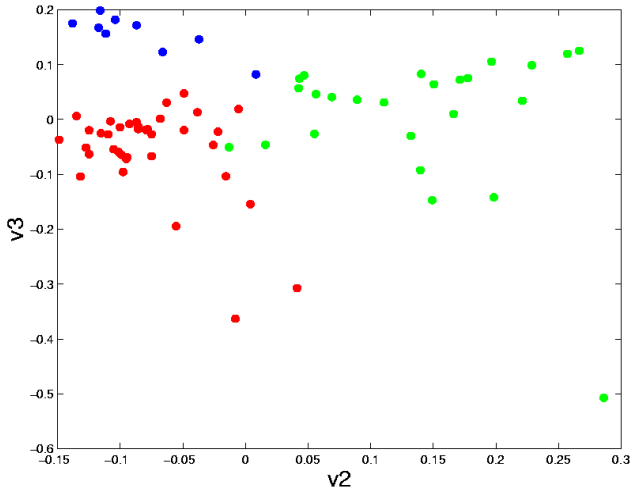


Four types of Cancer in Della Favera dataset

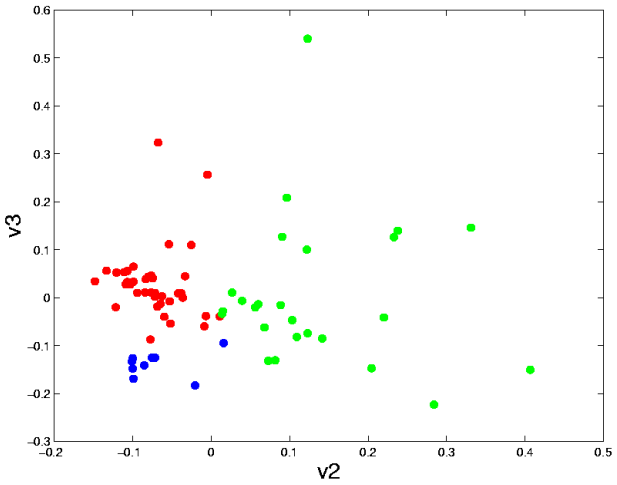
- CLL
- DLCL
- FL
- DLCL

Golub, TR et. al., *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*. Science, 1999 **286**

biclustering

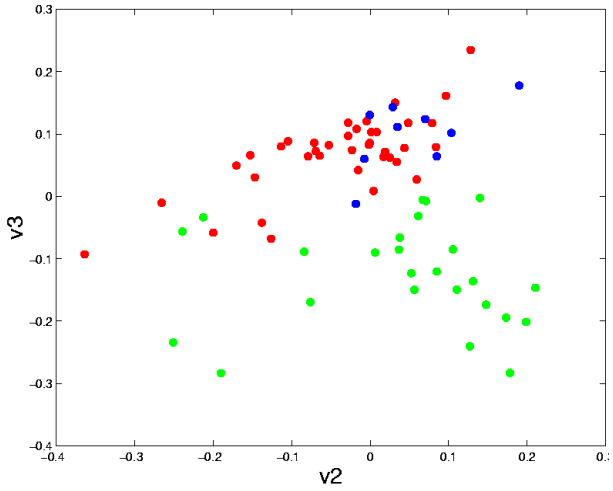


bistochastization

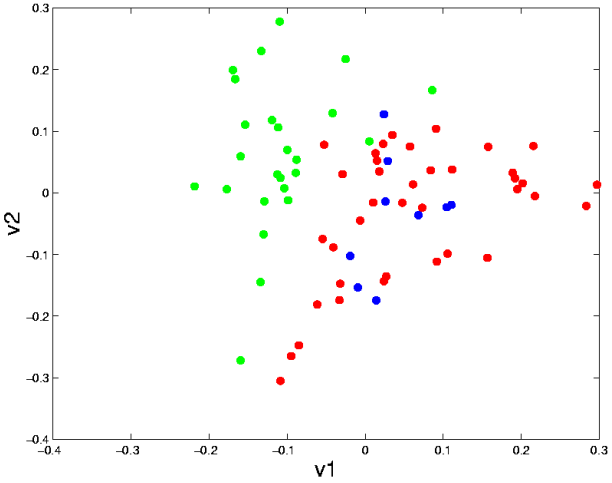


- ALL (B)
- ALL (T)
- AML

SVD



bi-normalization



Normalized cuts