Bioinformatics

Dr. George Garinis

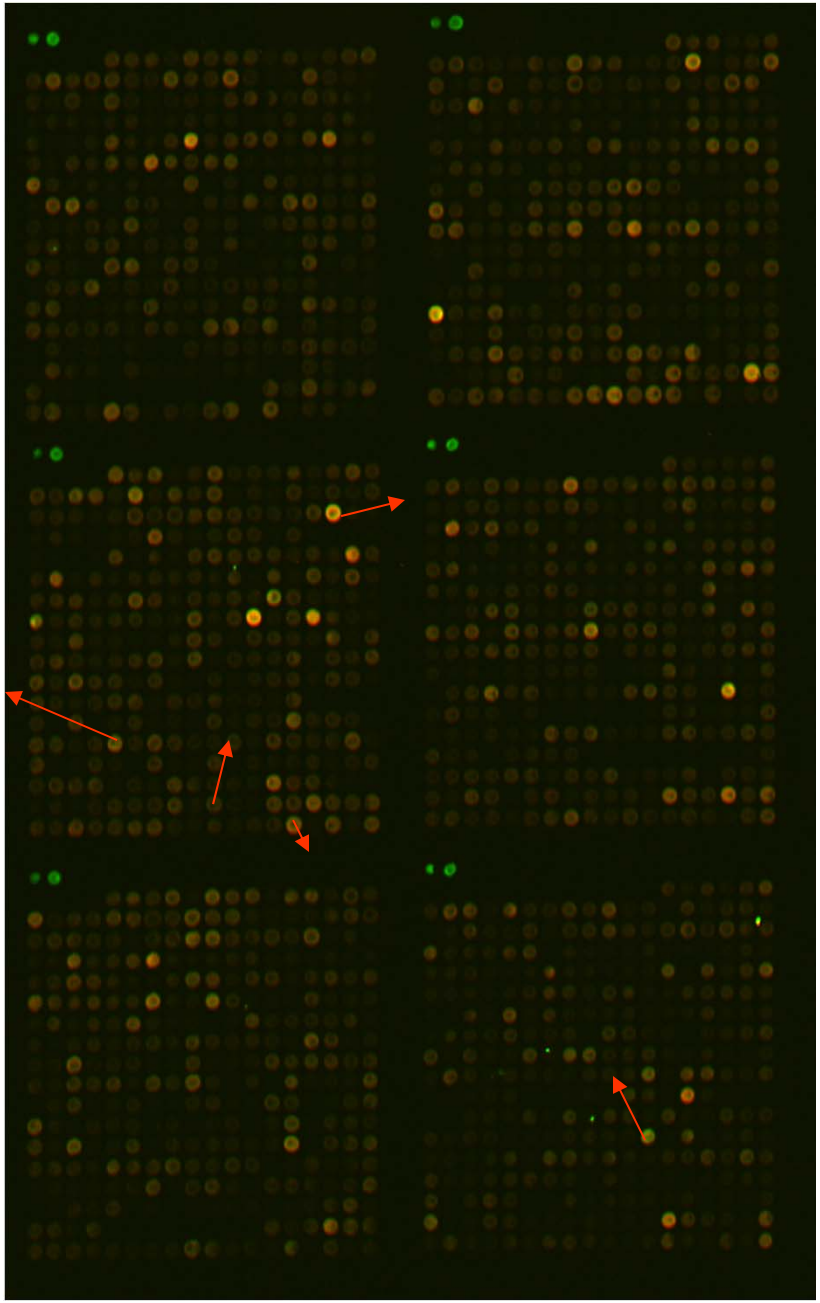
The problem in Genomics

Too many dimensions (data points)

- 15.000 genes X 100 microarray experiments = 1.500.000 dimensions
- But we can only perceive 3 dimensions.

So, we must decrease the dimensions of the data set
in order to perceive the data

Fold changes can have any possible direction in a 3-dimensional space

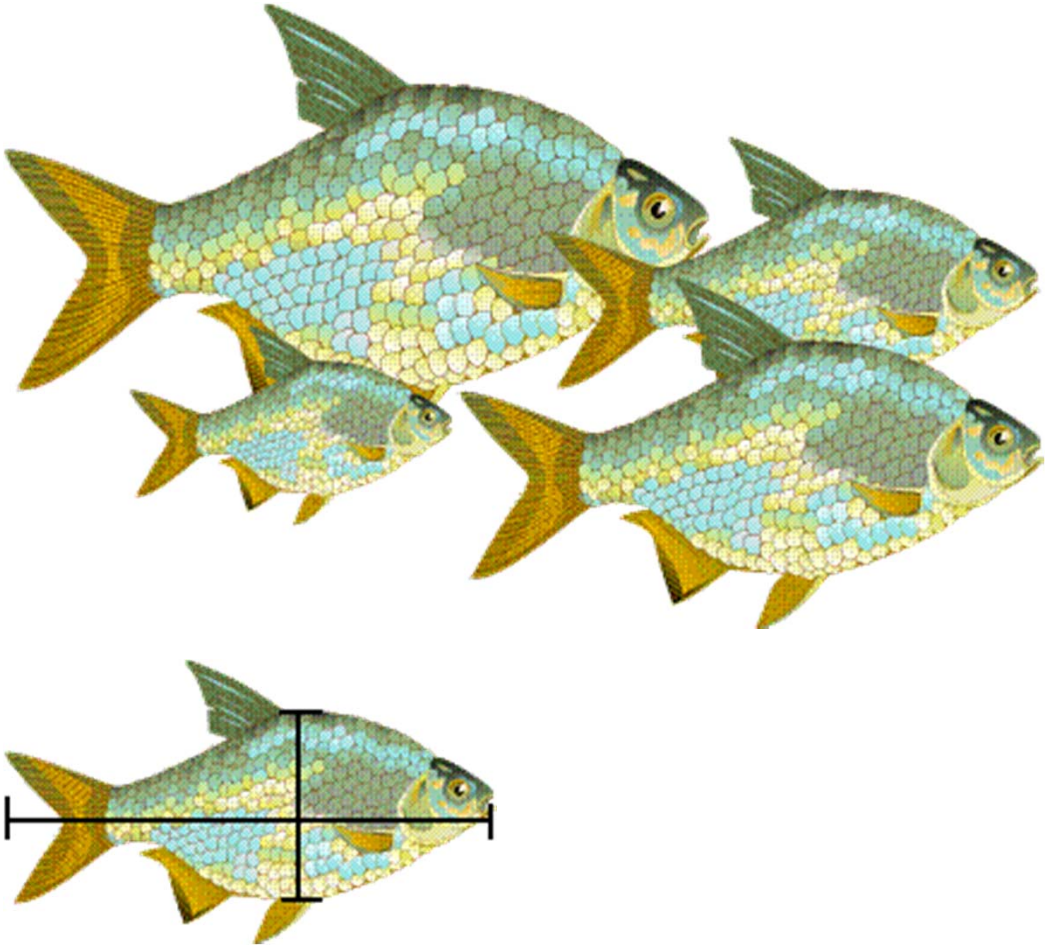


Projecting data onto fewer dimensions may sound like science fiction but you are all familiar with it.

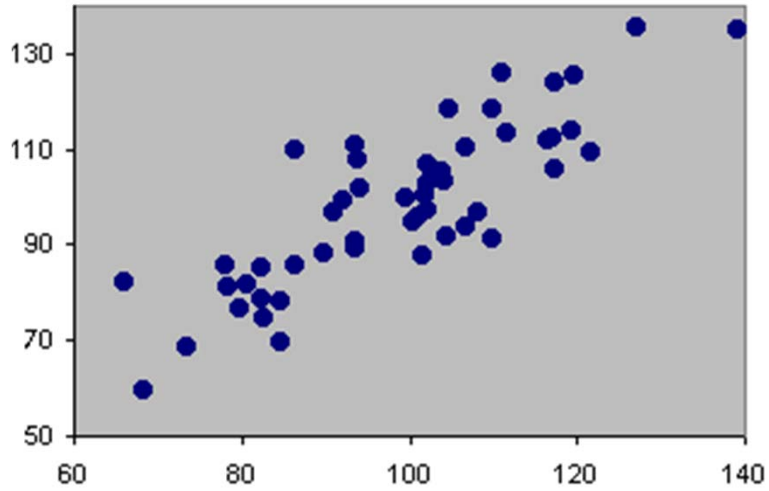
That's an eagle!

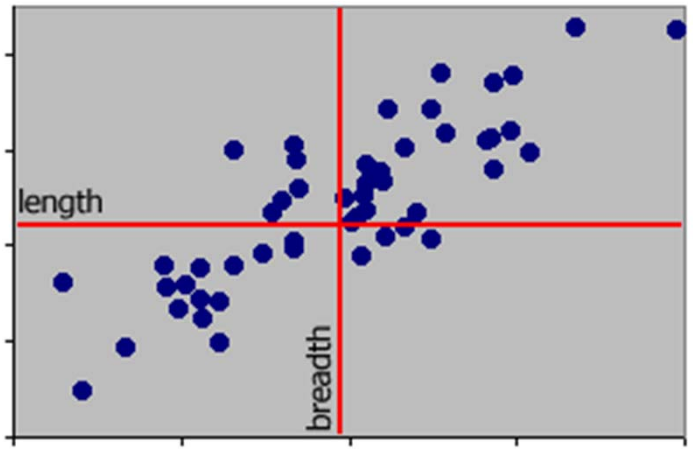
But: how do you know? It is a 2D picture… Not the real 3D bird

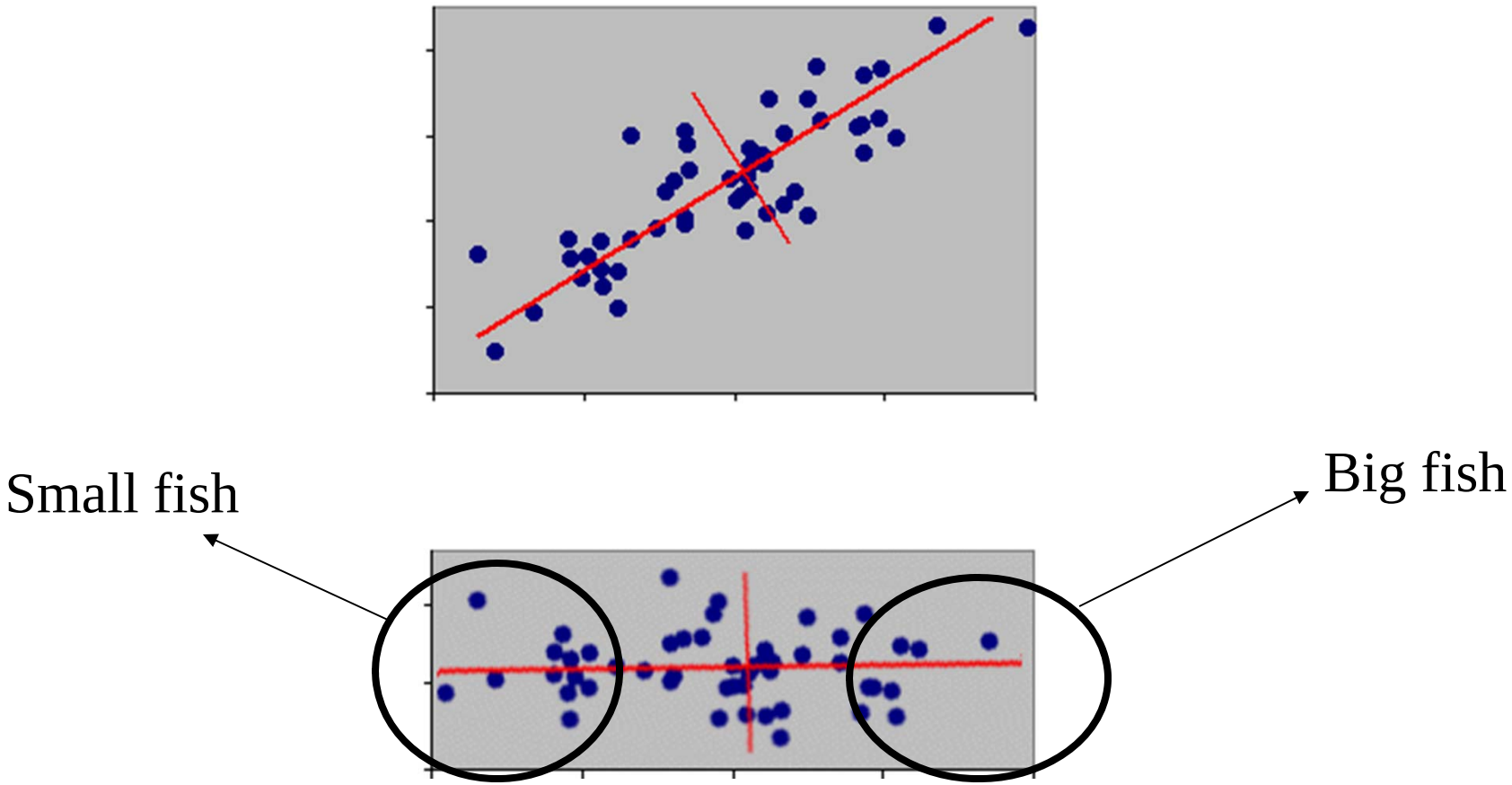
The truth is: **fewer dimensions can still retain much of the original information**



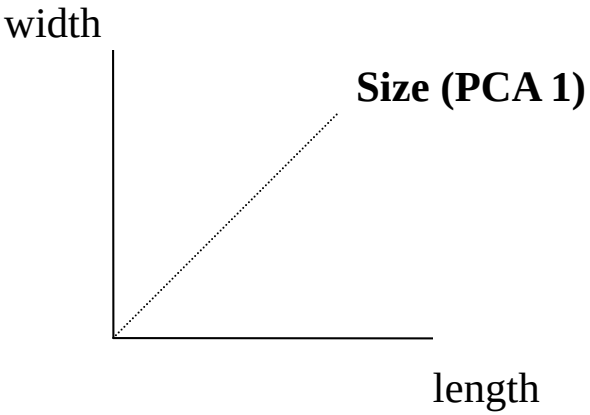
Suppose that 50 fish were measured for their length and width…

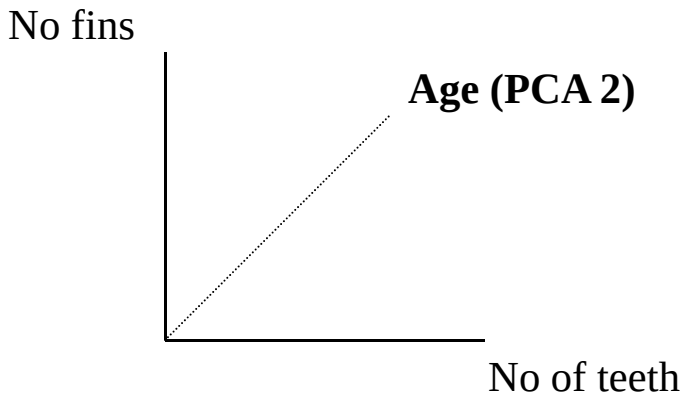

130
110
90
70
50
60
80
100
120
140

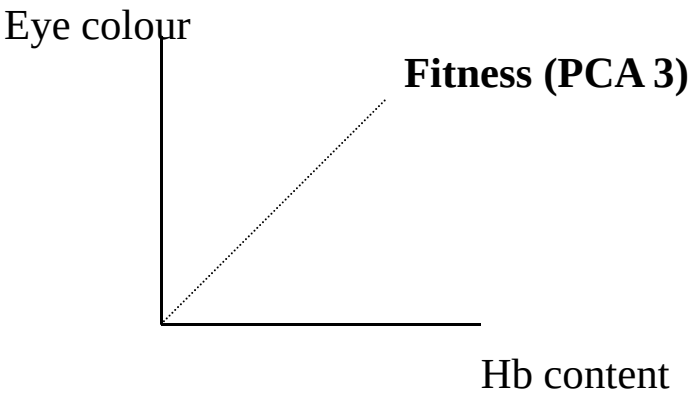

length
breadth

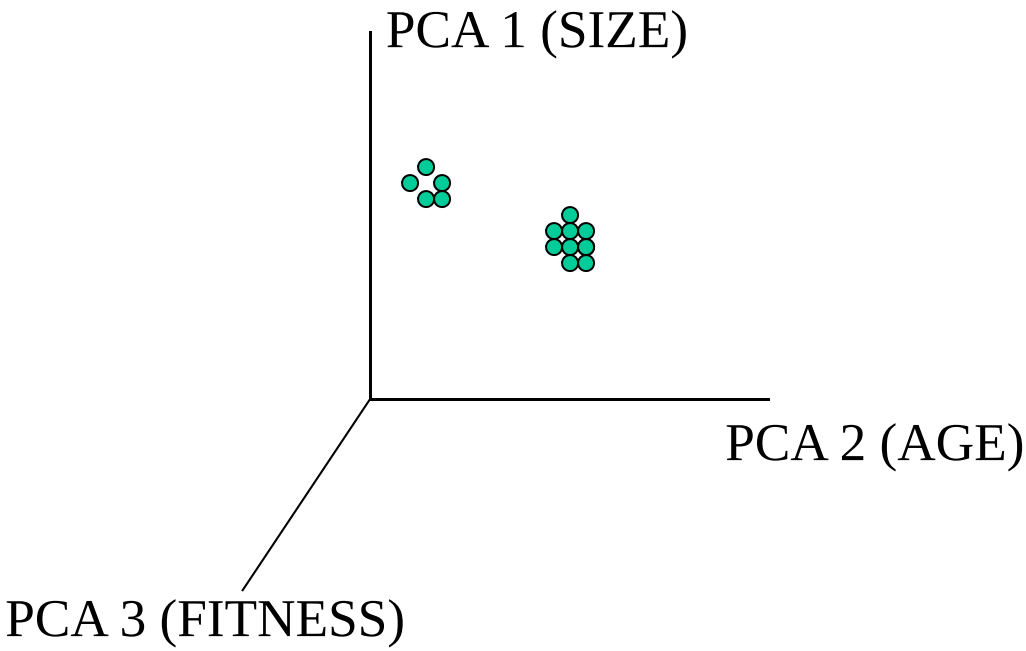


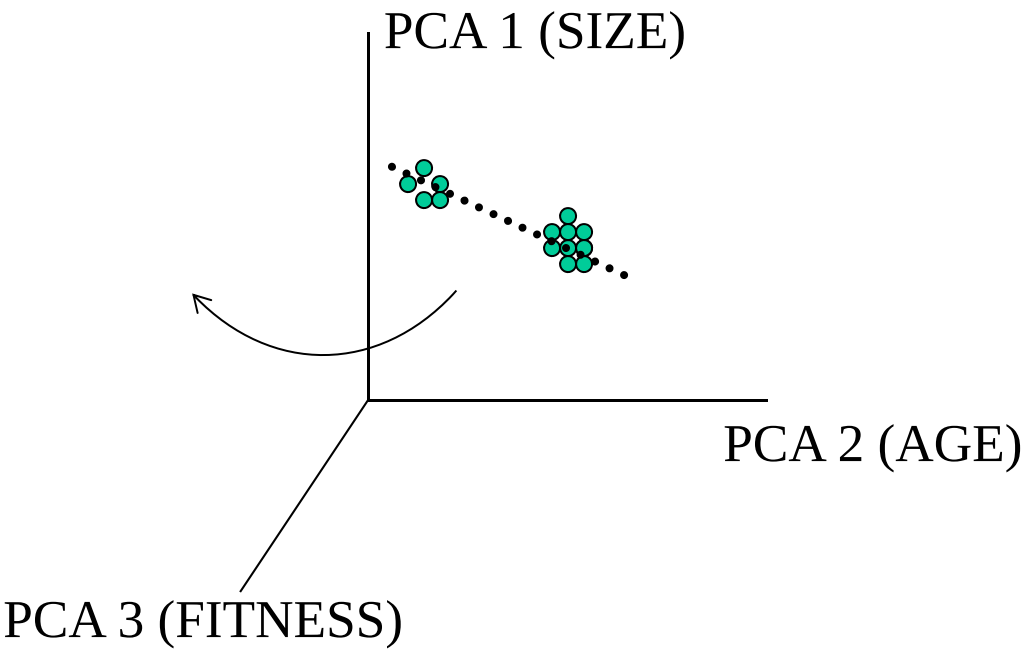
Length? Width? after all, you mean size…?

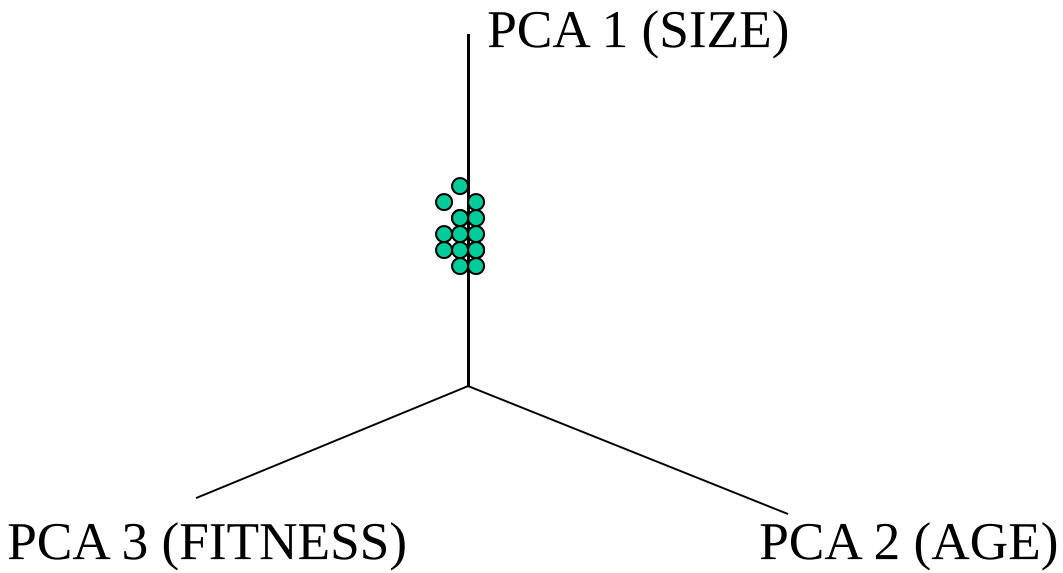

width
Size (PCA 1)
length


No fins
Age (PCA 2)
No of teeth


Eye colour
Fitness (PCA 3)
Hb content

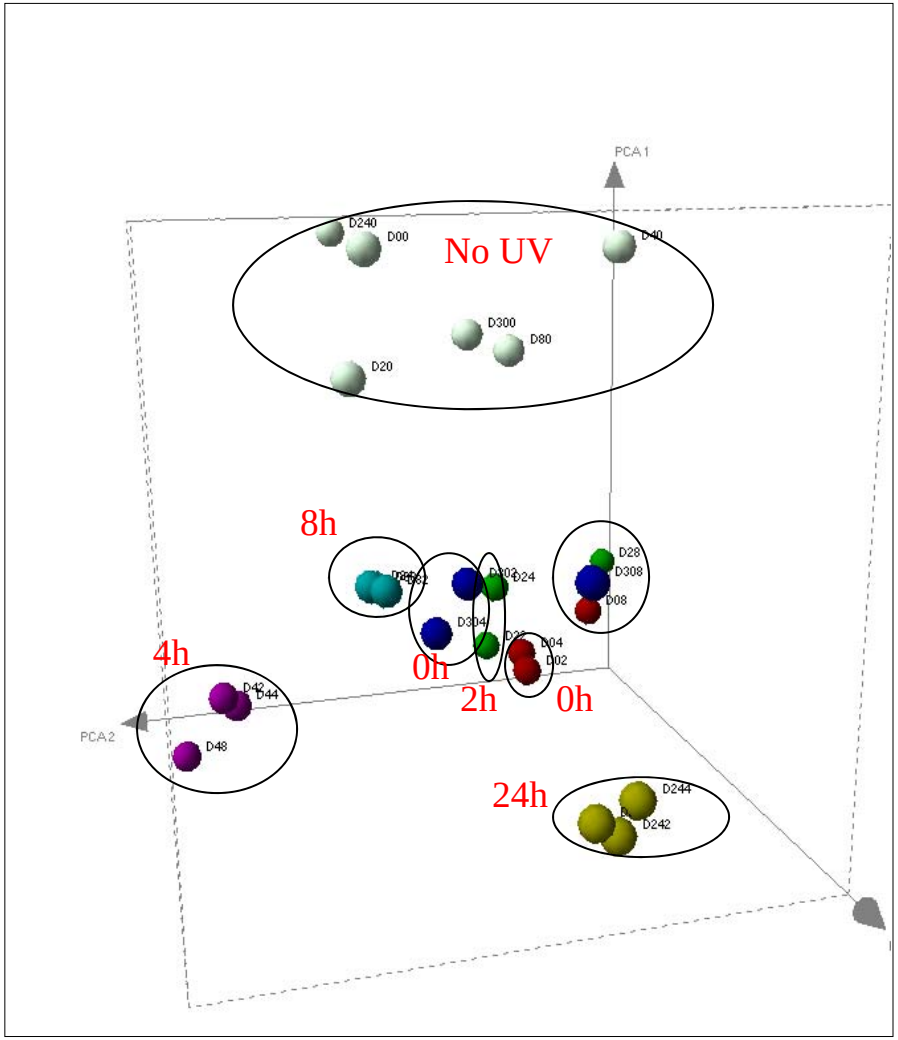

PCA 1 (SIZE)
PCA 2 (AGE)
PCA 3 (FITNESS)


PCA 1 (SIZE)
PCA 2 (AGE)
PCA 3 (FITNESS)


PCA 1 (SIZE)
PCA 3 (FITNESS)
PCA 2 (AGE)

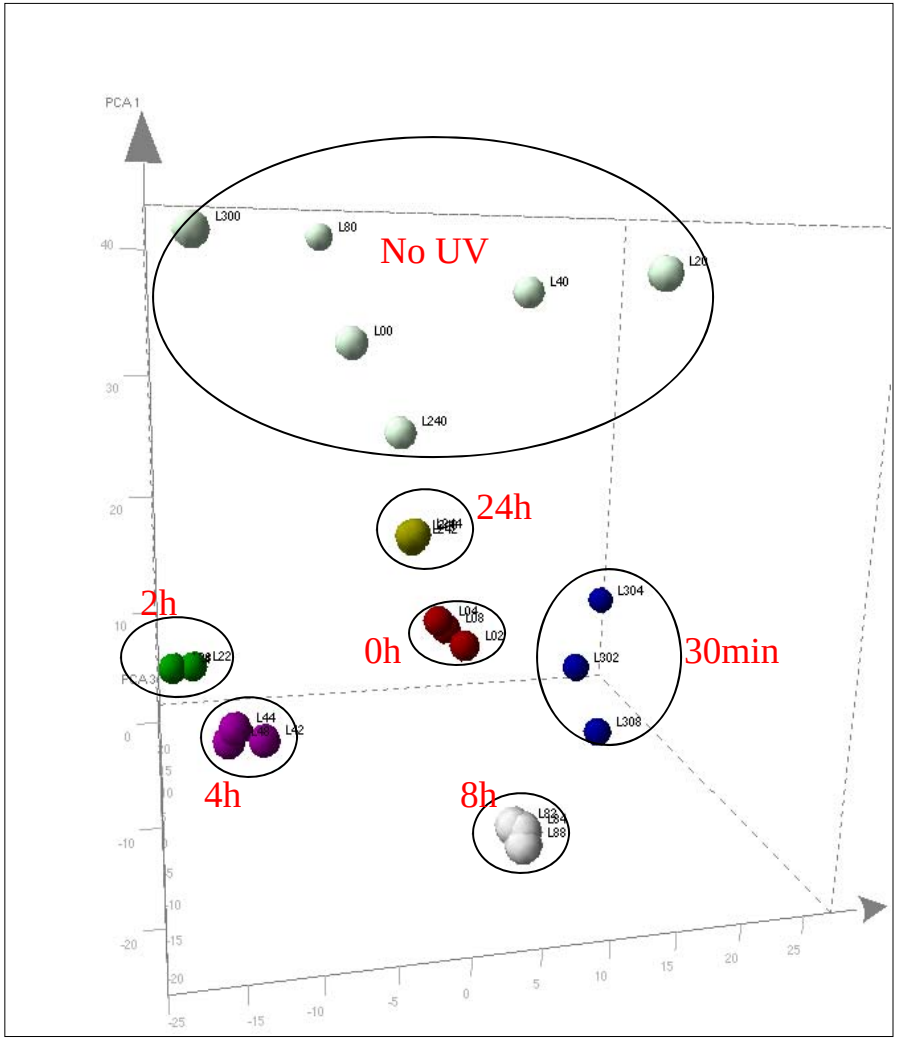
Non-photoreactivated MDFs

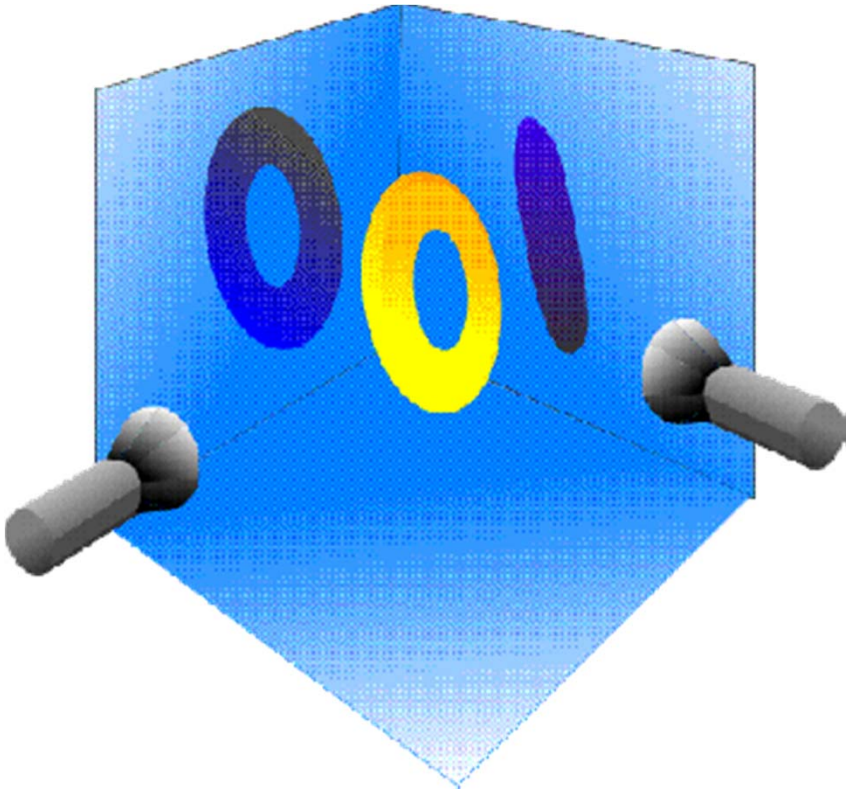
PCA



Photoreactivated MDFs

PCA

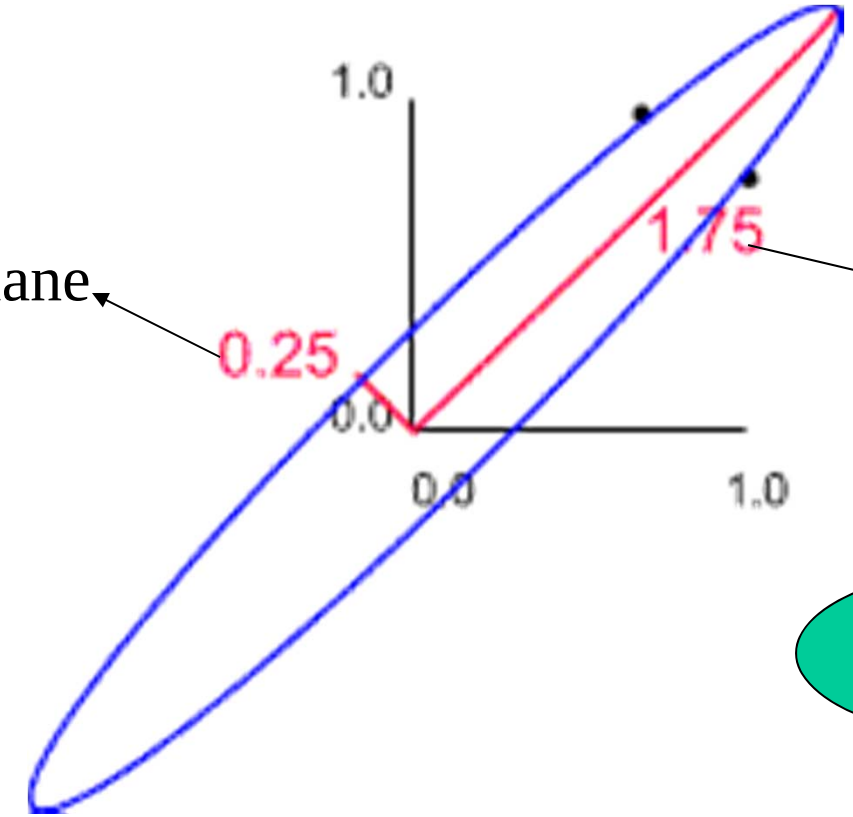




Shine a light onto this doughnut from two different directions.

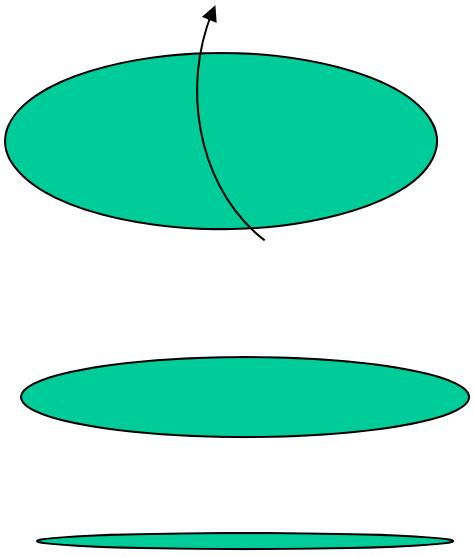
These lights cast shadows onto two 'screens'. The nature of the shadow is dependent on the position of the torch.

1. move the doughnut and keep the torches stationary
2. keep the doughnut stationary and move the torches.



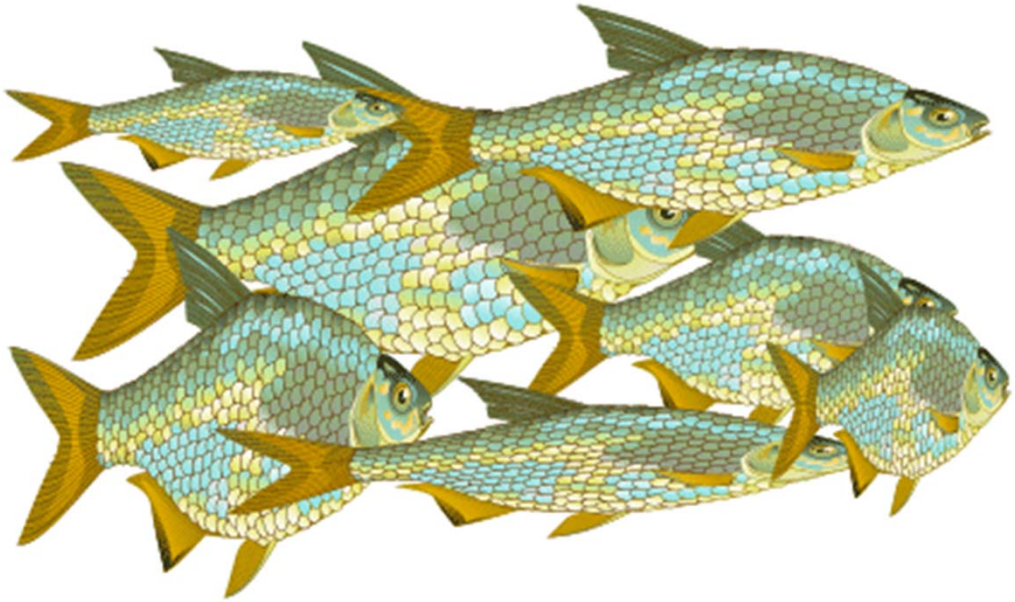
Width of the plane

Length of the plane



1.75X100/2=87.5 preserved variability

0.25X100/2=12.5



Could you represent these fish by a size variable alone? If not, why not?

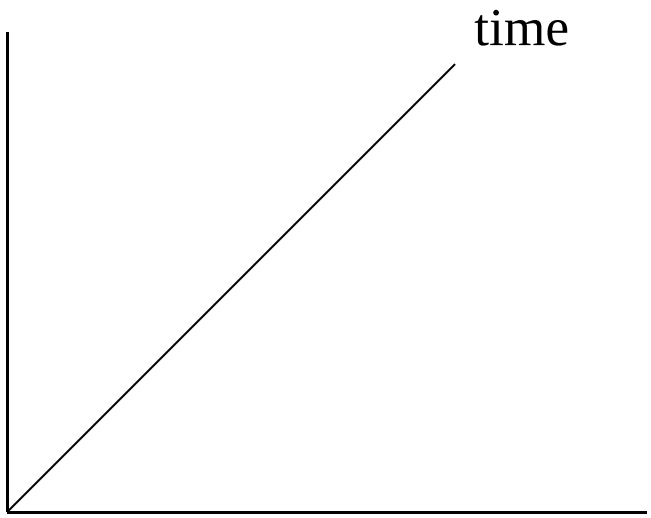
The obvious answer is that they are different sizes *and* shapes

probably because these fish are of a different species…

What is the principal component of AGING process?

variables

Grey hair
Weight
Heart failure incidence
Cancer incidence
Bone structure

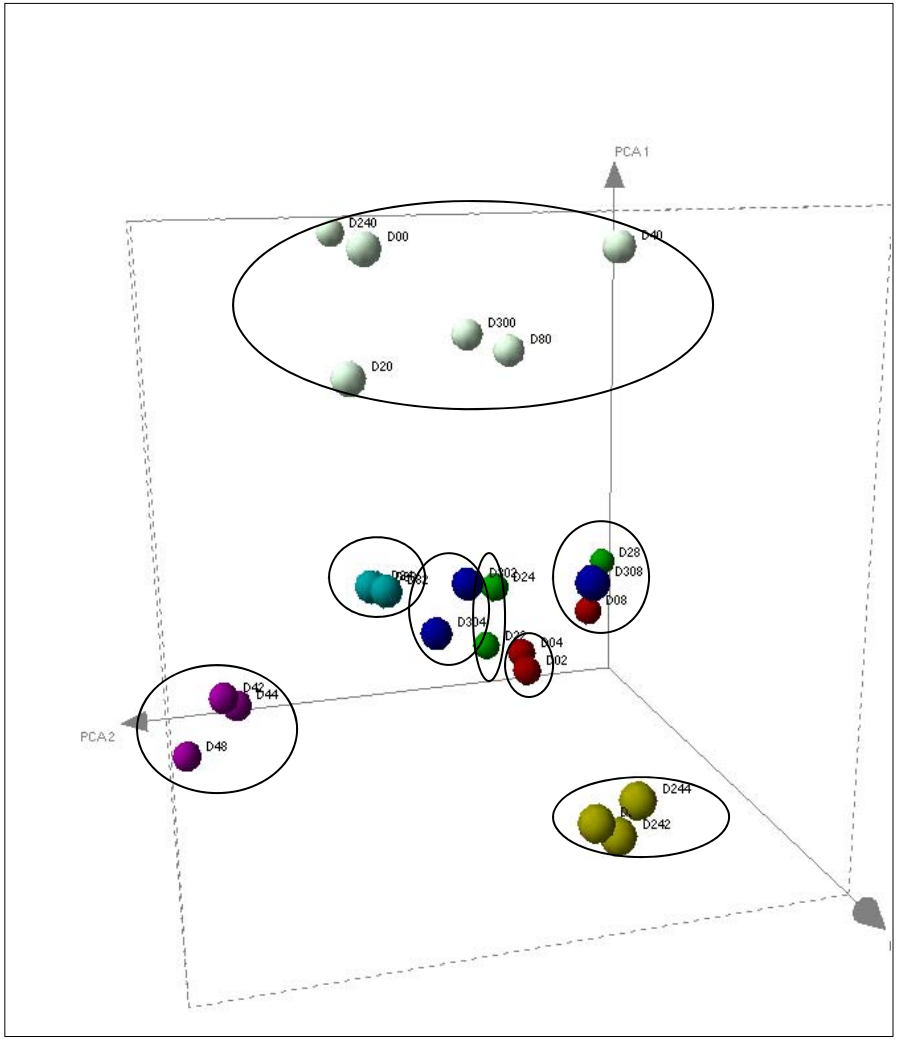


Probably…time is one of the principle components

Questions for aging:
What are the variables that project in the PC of time?
And which are the other 2 principle components?

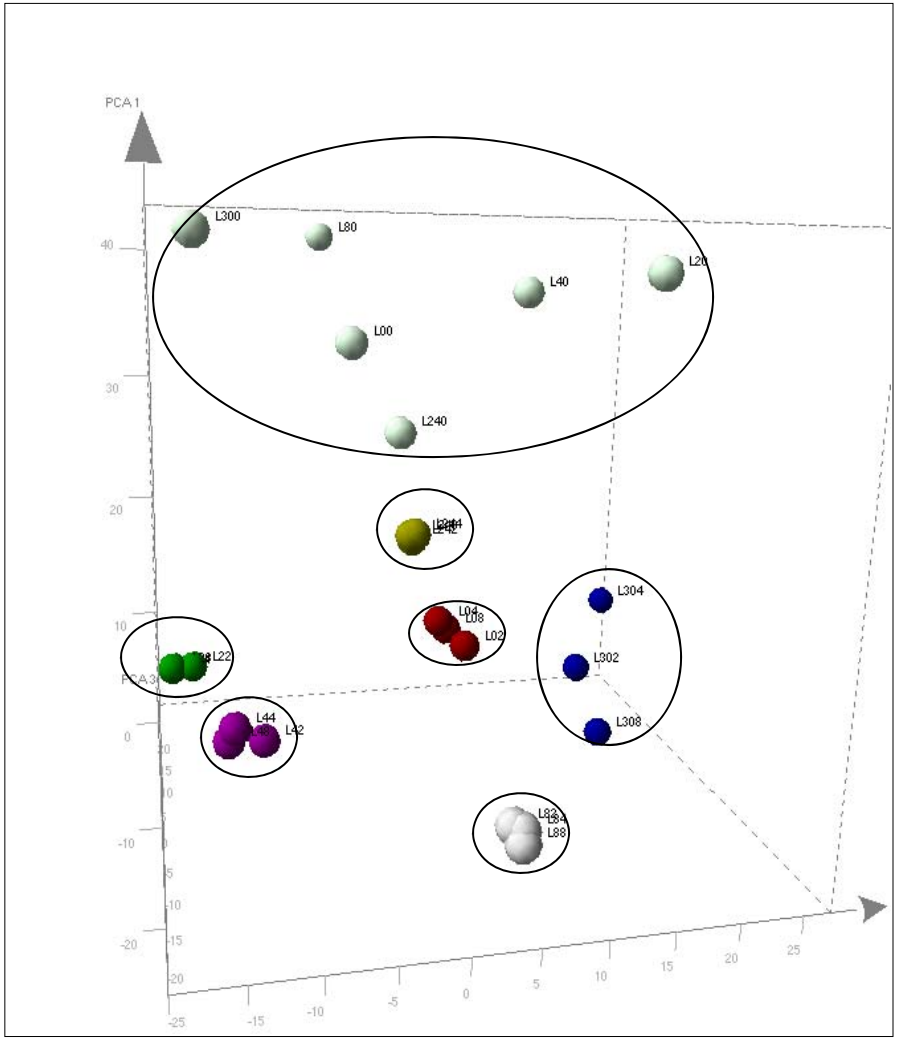
Non-photoreactivated MDFs

PCA



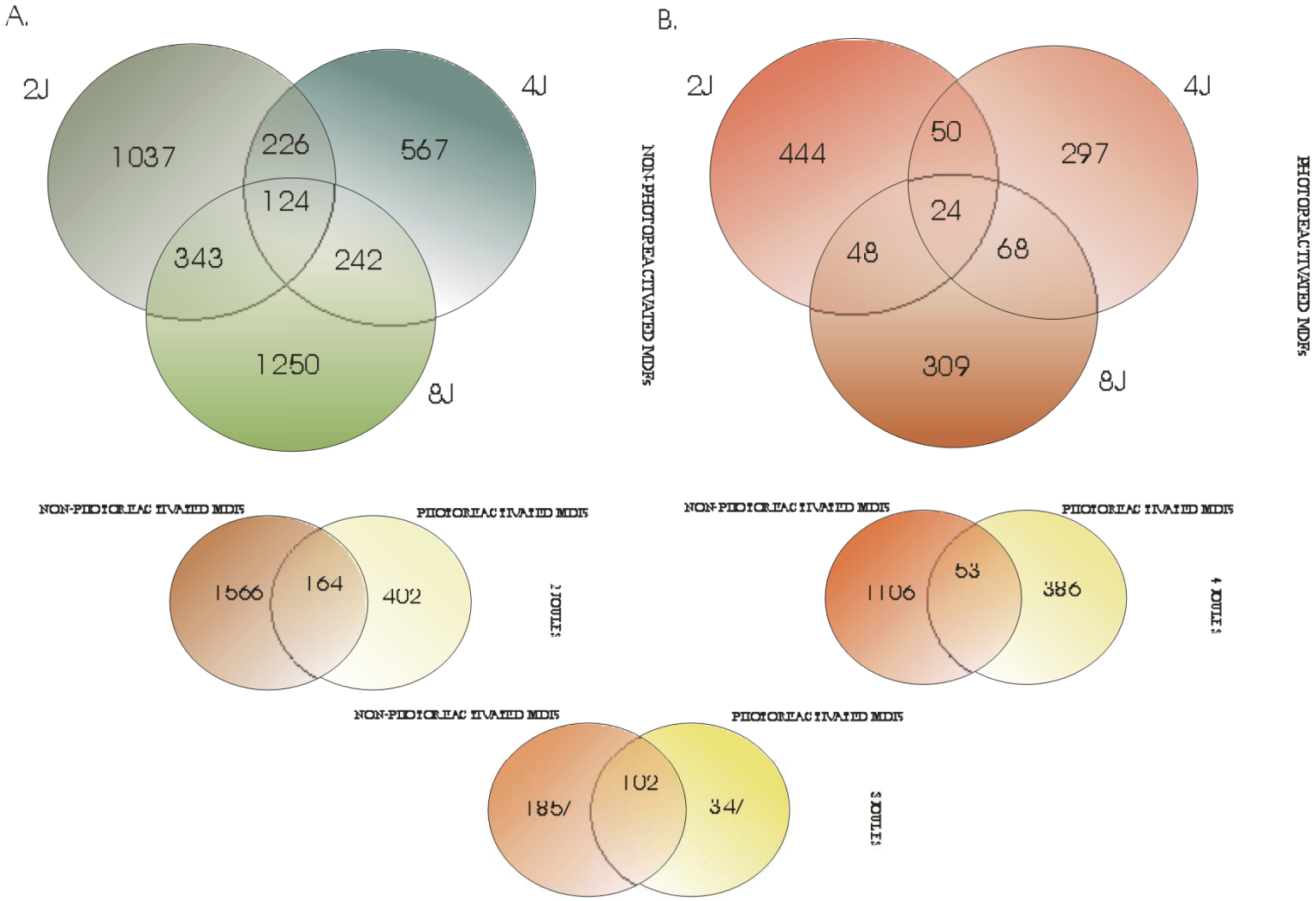
Photoreactivated MDFs

PCA



0-24h

Time-, dose- and lesion-specific significant genes



Other ways to decrease dimensions of our data sets

is

GROUPING THE VARIABLES

Various ways to group (cluster)

Hierarchical clusters

-Single linkage
-Average linkage
-Complete linkage

Partitioning clusters

-K-cluster
-Artificial neural networks

SINGLE LINKAGE

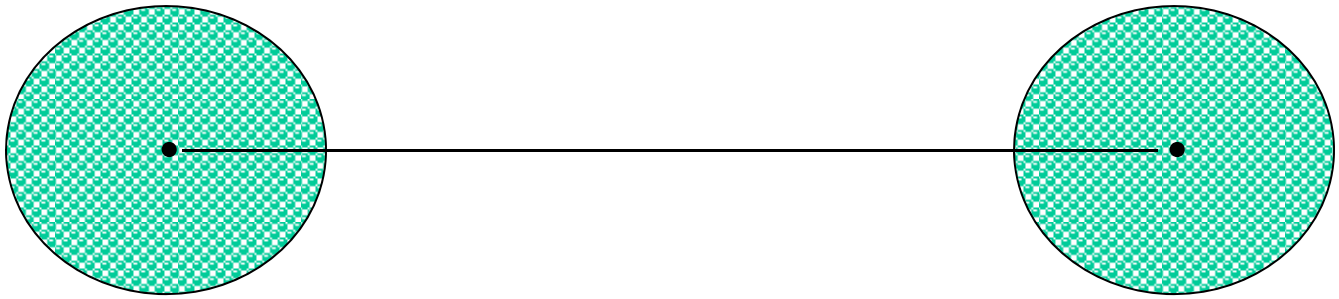
In *single-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

Complete linkage

In *complete-linkage* clustering we consider the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Average linkage

In *average-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.



Hierarchical clusters

Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of hierarchical clustering is this:

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

After merging BOS with NY:

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	**0**	**223**	**1308**	**802**	**2815**	**2934**	**2786**	**1771**
DC	**223**	**0**	**1075**	**671**	**2684**	**2799**	**2631**	**1616**
MIA	**1308**	**1075**	**0**	**1329**	**3273**	**3053**	**2687**	**2037**
CHI	**802**	**671**	**1329**	**0**	**2013**	**2142**	**2054**	**996**
SEA	**2815**	**2684**	**3273**	**2013**	**0**	**808**	**1131**	**1307**
SF	**2934**	**2799**	**3053**	**2142**	**808**	**0**	**379**	**1235**
LA	**2786**	**2631**	**2687**	**2054**	**1131**	**379**	**0**	**1059**
DEN	**1771**	**1616**	**2037**	**996**	**1307**	**1235**	**1059**	**0**

Then we compute the distance from this new cluster to all other clusters, to get a new distance matrix:

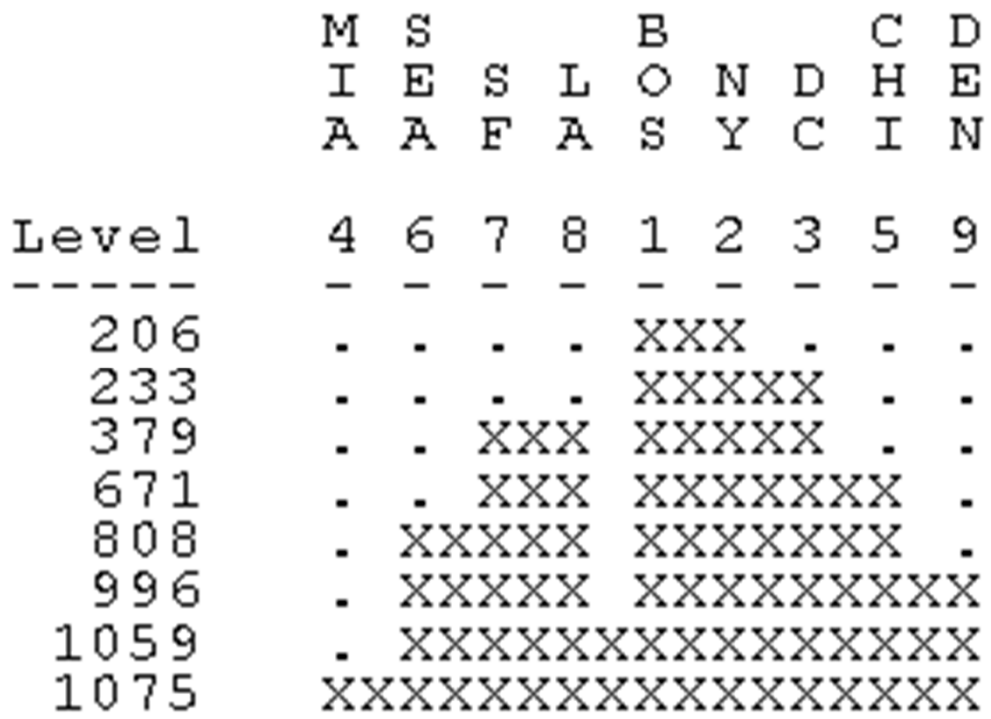
After merging DC with BOS-NY:

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	**379**	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

Now, the nearest pair of objects is SF and LA, at distance 379. These are merged into a single cluster called "SF/LA". Then we compute the distance from this new cluster to all other objects, to get a new distance matrix:

	BOS/ NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0

The whole process is then summarized:

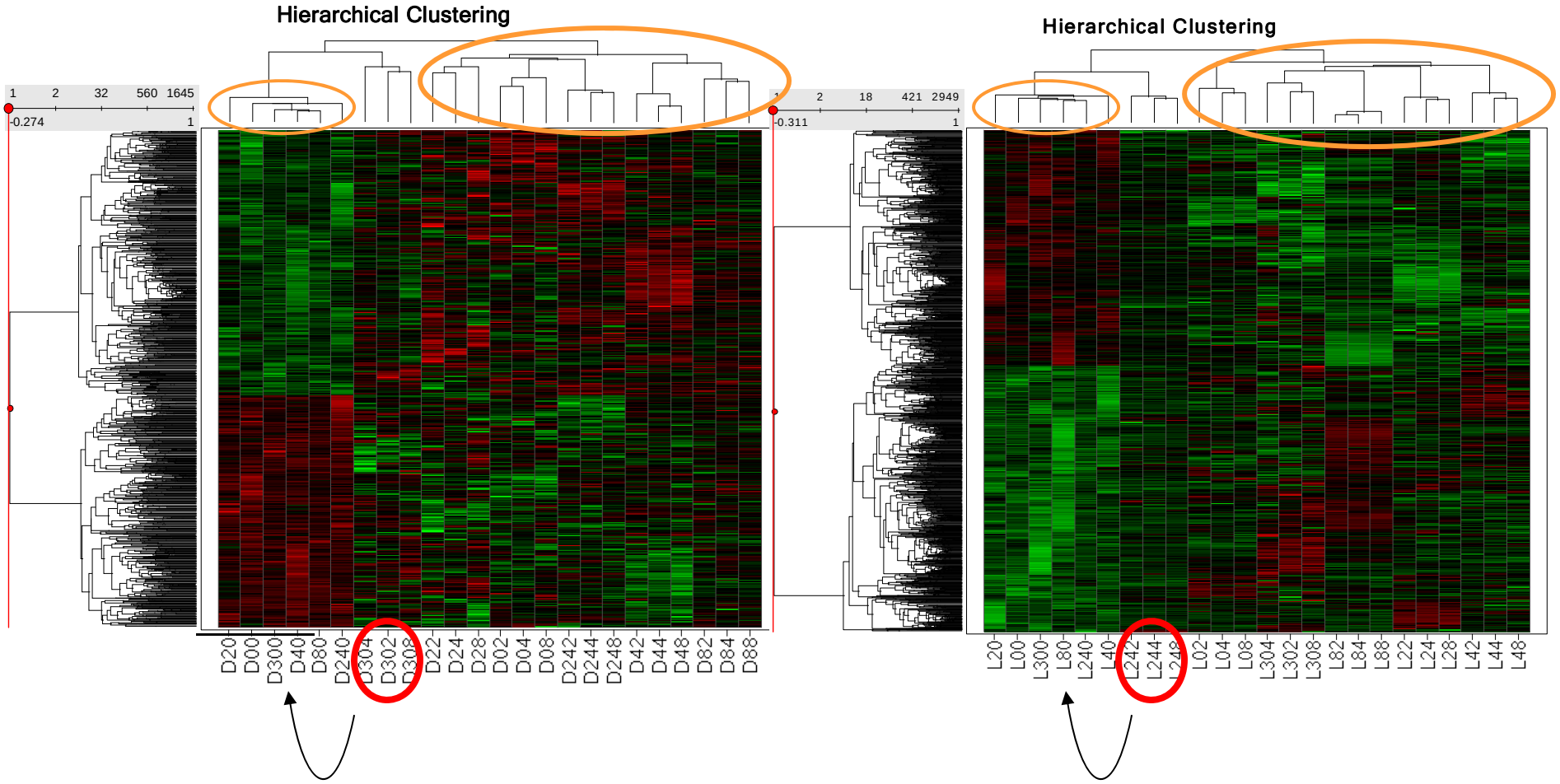


Level	4	6	7	8	1	2	3	5	9
	M	S			B			C	D
	I	E	S	L	O	N	D	H	E
	A	A	F	A	S	Y	C	I	N
206	.	.	.	.	XXX	.	.	.	
233	.	.	.	.	XXXXX	.	.		
379	.	.	XXX		XXXXX	.	.		
671	.	.	XXX		XXXXXXX	.			
808	.	XXXXX			XXXXXXX	.			
996	.	XXXXX			XXXXXXXXX				
1059	.	XXXXXXXXXXXXXXX							
1075	XXXXXXXXXXXXXXXXX								

In the diagram, the columns are associated with the items and the rows are associated with levels (stages) of clustering. An 'X' is placed between two columns in a given row if the corresponding items are merged at that stage in the clustering.

Non photoreactivated MDFs

Photoreactivated MDFs



Partitioning clusters

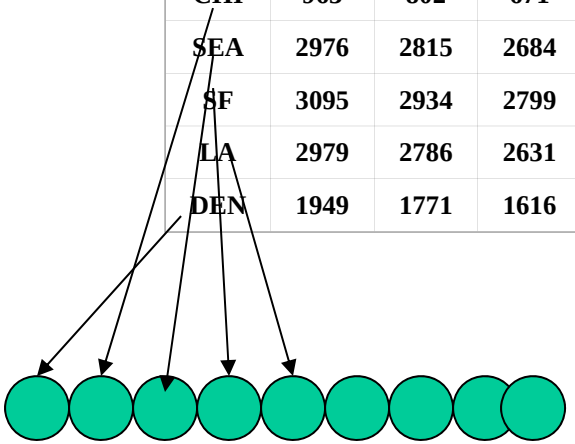
The difference between partitioning clusters from hierarchical clusters is:

-Here there is no summary of the clustering process

-We define the number of K different clusters that will have the greatest possible distinction

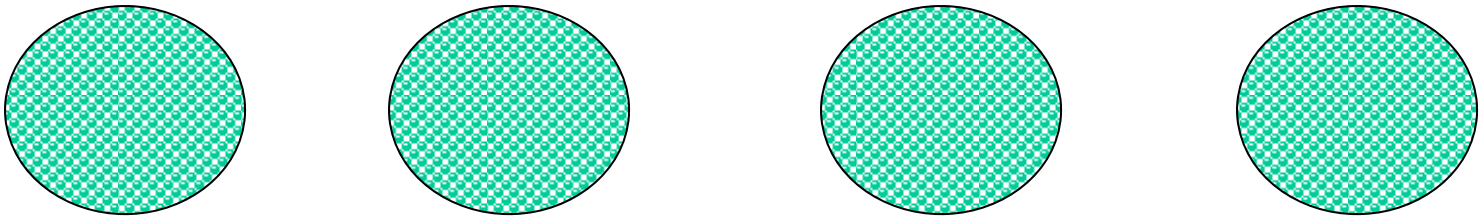
K-cluster

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0



…..K clusters where K= number of variables

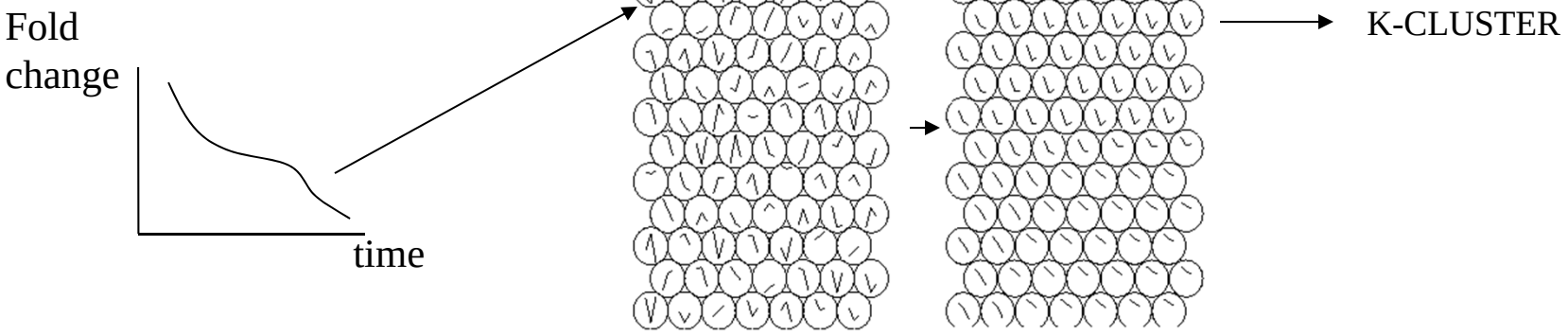
K1 K2 K3 K4



Self organizing maps

Self organizing maps

Random Ordered



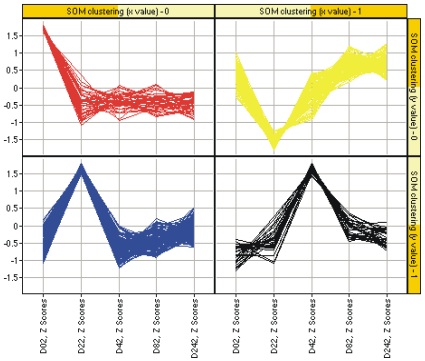
SOMS group K-clusters by similarity of expression profiles

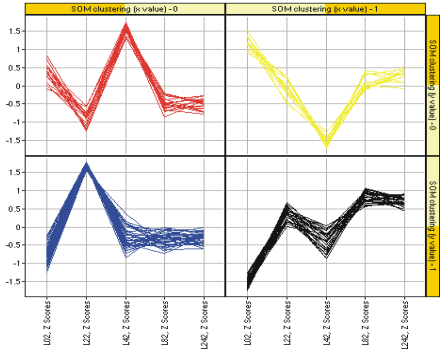
Self organizing maps of time-, dose- and lesion-specific transcriptional profiles

Non-photoreactivated MDFs

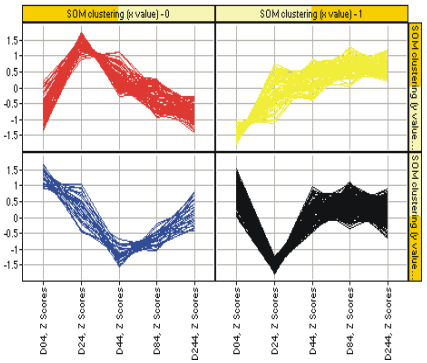
Photoreactivated MDFs

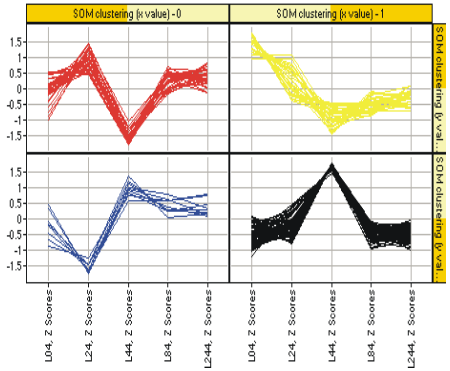
2 joules



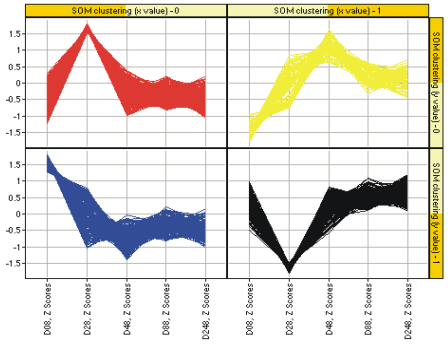


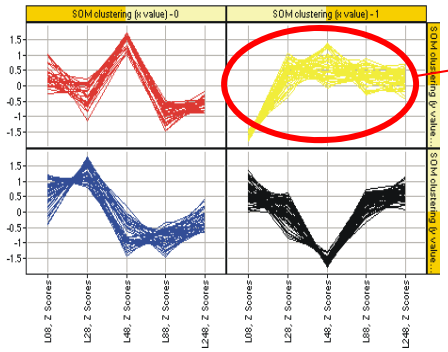
4 joules





8 joules





chromatin assembly-disassembly	0 hours	2 hours	4 hours	8 hours	p-value	expression profile
Mcmd2	-2.160530721	1.386917	1.039724	3.634409	0.006	1
H1f0	1.110589902	3.579108	1.81788	2.997321	0.003	1
Set	1.322087632	-8.71536	-1.13188	-2.20683	0.01	1
nucleosome assembly						
Mcmd2	-2.160530721	1.386917	1.039724	3.634409	0.004	2
H1f0	1.110589902	3.579108	1.81788	2.997321	0.003	3
Set	1.322087632	-8.71536	-1.13188	-2.20683	0.02	2
DNA repair						
[illegible]	1.0593817	5.824171	1.003755	1.817957	0.0005	2
Ubl1	-1.191205374	1.415204	-1.20876	3.029572	0.04	3
Xpc	-1.001964215	2.293345	-3.7576	1.938758	0.003	3
Rad50	-1.968605952	2.593673	-1.04176	2.758281	0.0002	2
Adprtl2	3.065760791	-1.06462	2.058573	-3.08862	0.05	4
DNA replication						
Mcmd2	-2.160530721	1.386917	1.039724	3.634409	0.05	4
Rpa2	2.357193644	-2.36607	-1.1462	-1.98283	0.05	4
Set	1.322087632	-8.71536	-1.13188	-2.20683	0.04	4
Pold3-pending	1.070745501	-1.70183	-1.1358	-1.58597	0.02	4
Top2b	-1.372042893	1.293502	-1.53442	3.032067	0.001	4
Rrm2	1.779874308	-2.1595	1.753115	-5.80485	0.002	4
DNA dependent DNA replication						
Mcmd2	-2.160530721	1.386917	1.039724	3.634409	0.003	3
Top2b	-1.372042893	1.293502	-1.53442	3.032067	0.0004	5

Total:1.500 out of 15.000 = 10%

DNA repair:

2J: 6 out of 69= 8.7%

8J: 20 out of 69= 28.9%

Biological process?