

Explorative approaches

microarray.no

Kjell Petersen
J-Express Analysis course
Bergen March 2011

Overview

- Non-statistical approaches
 - Clustering
 - Projection
- Get a feel with the data
- Find patterns and trends
- Can very well define a subset of special interest
 - A subset of genes can be followed up statistically

Explorative analysis

Find meaningful groups of genes or samples;

Look for patterns; try to get a feeling of the data;

Do samples from similar tissues group together?

Do patients or controls group together?

Are there other trends in the data?

Co-regulated genes?

Visualise main structures in data;

Outliers?

Explorative analysis

Unsupervised;

No biological information on samples or genes
Calculated purely on measurement values

Possible both ways; across:

Samples – which are more similar?

Genes – which group naturally together?

Projection; PCA, CA,...

Clustering; Hierarchical, K-means, SOM,...

Distance measures

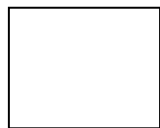
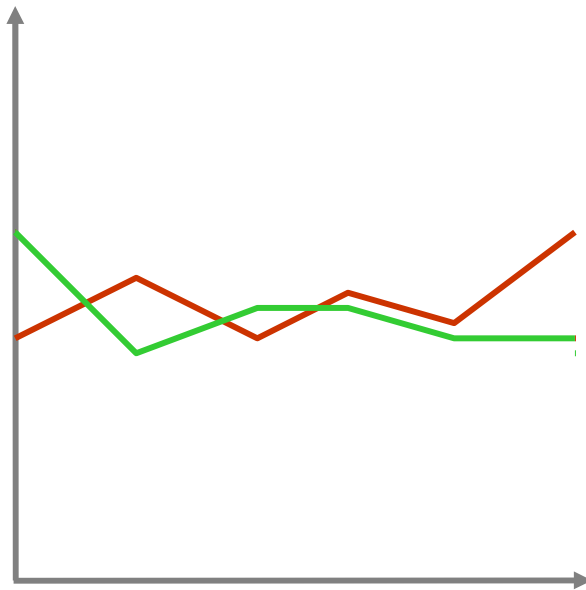
Rearrange genes with regard to some similarity or dissimilarity measure;

In **projection** algorithms, two relatively similar objects should be placed in the vicinity of each other in the projected space.

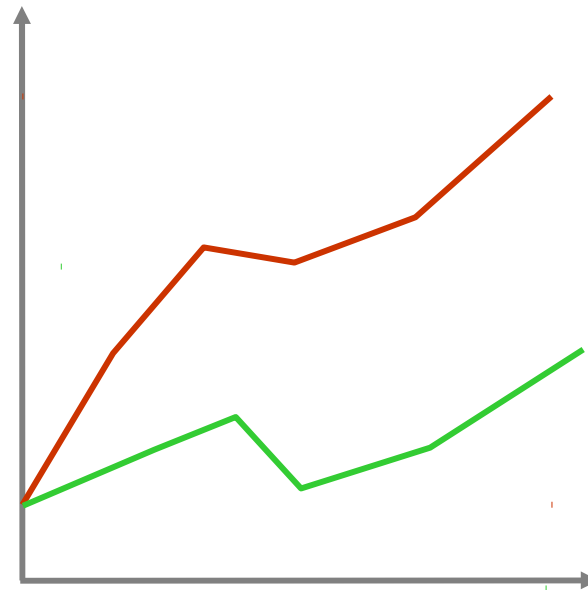
In **clustering** algorithms, two relatively similar objects should be placed in the same cluster.

Distance measures

Which profiles are more similar?



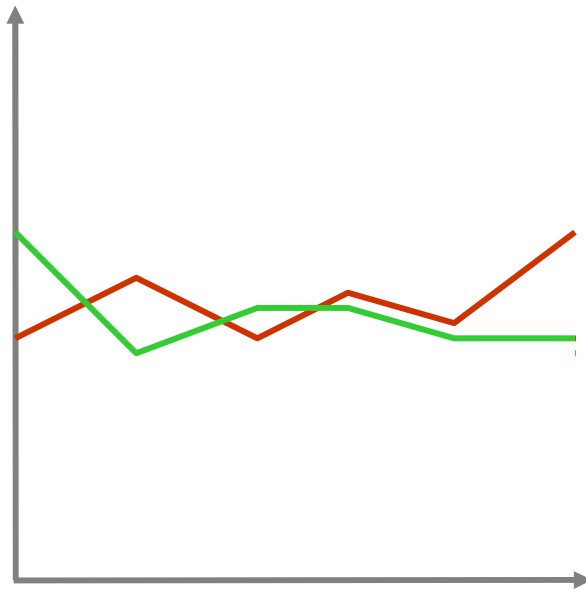
?



?

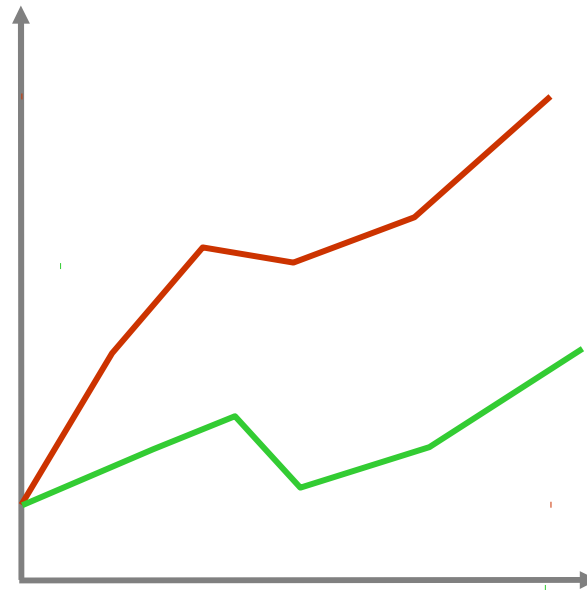
Distance measures

Which profiles are more similar?



X

If Euclidean

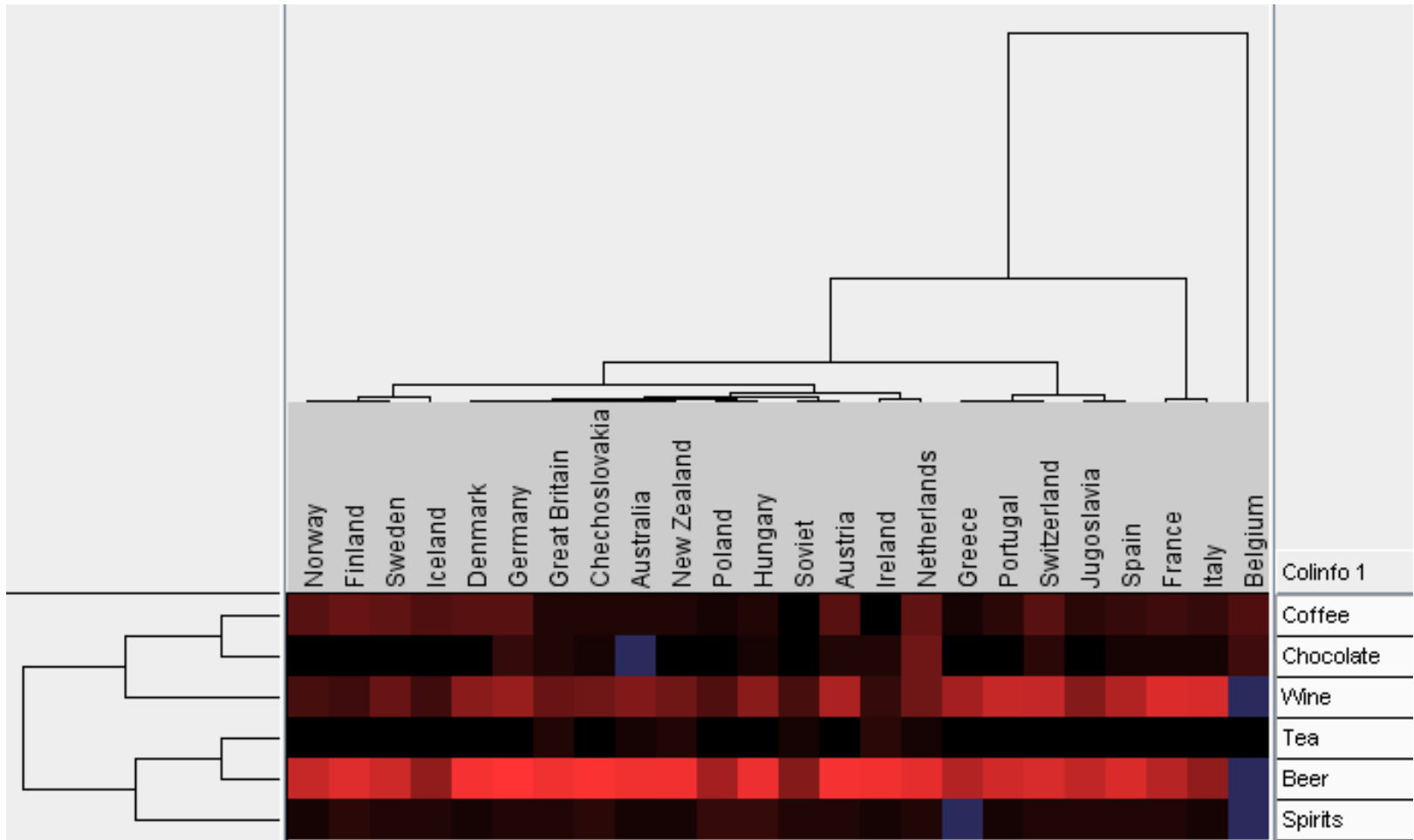


X

If Correlation

microarray.no

Clustering



microarray.no

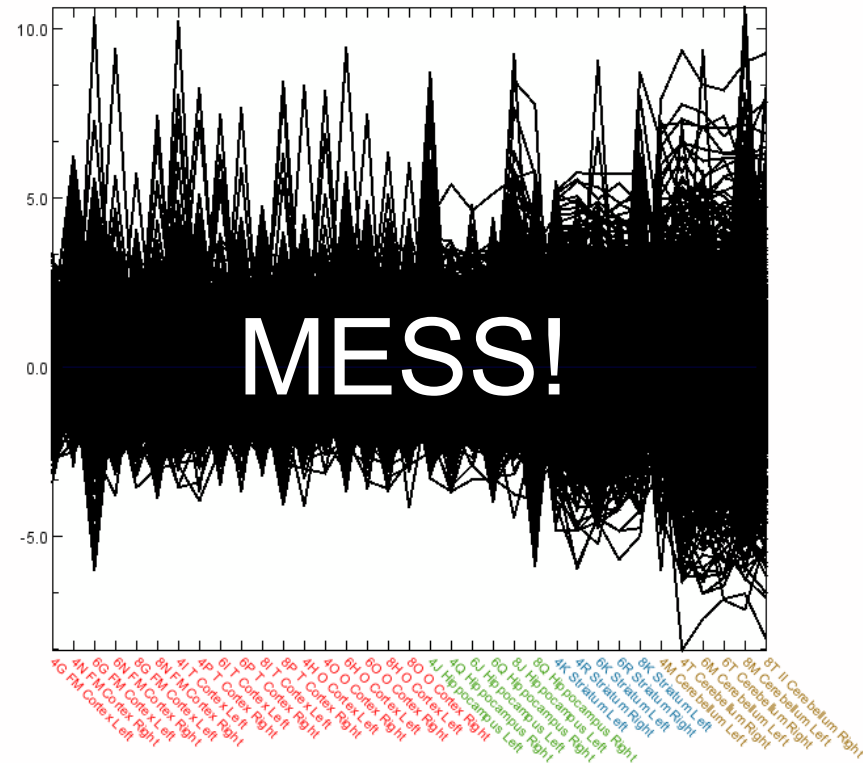
Clustering of genes

Datasets can be large, often 15000 genes and 50-100 samples.

Clustering creates smaller groups of genes that can be viewed and analysed separately.

These groups consist of genes that behave similarly across samples.

This expression similarity is often caused by co-regulation or other interesting biological processes.

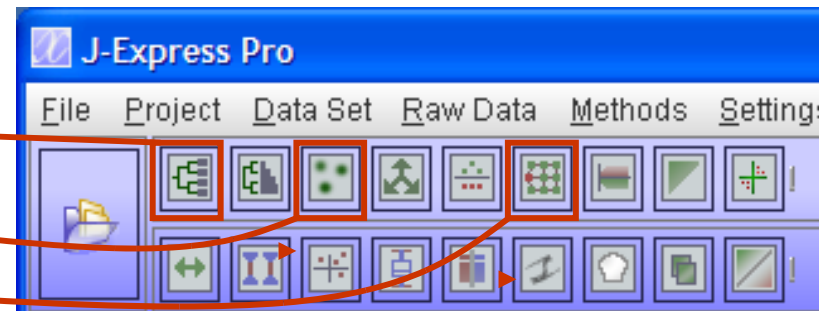


Examples:

Hierarchical clustering

k-means clustering

Self-organizing maps



Clustering algorithms

Hierarchical clustering

Build cluster by merging closest samples

Re-compute distances

→ Tree of clusters

Presents “all” relations – lets you decide the cut-offs between clusters

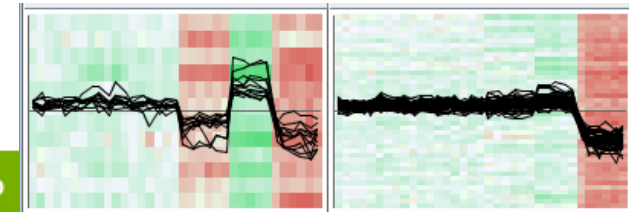
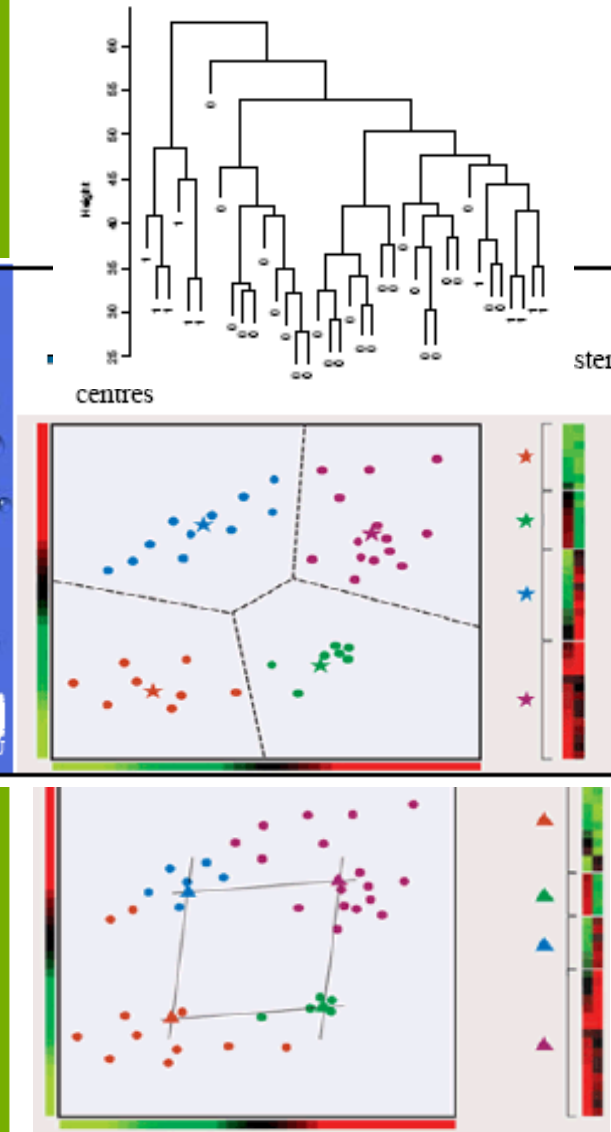
K-means

Optimise location of k location centres

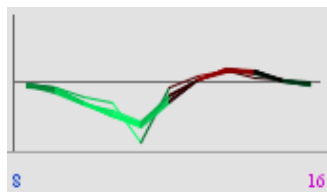
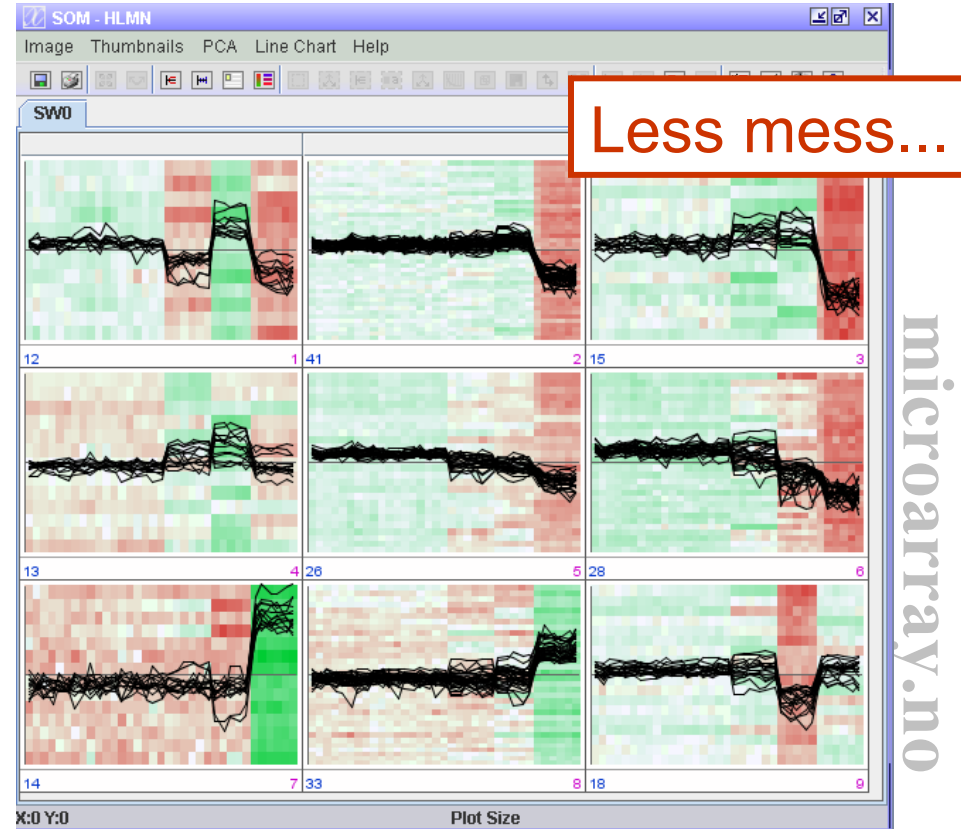
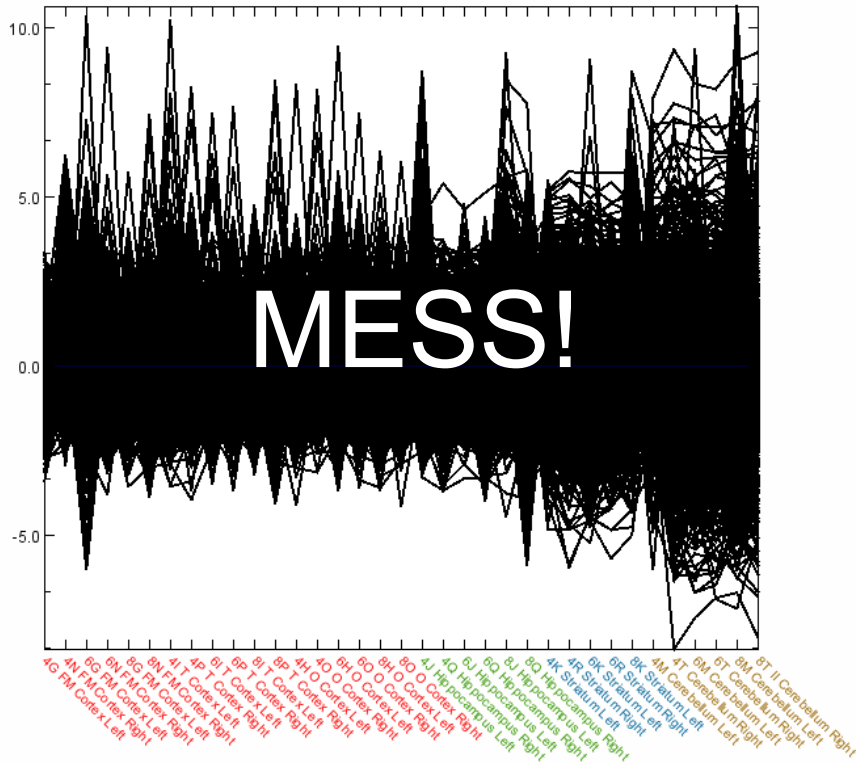
Needs a predefined number of clusters to look for

Self-organising map (SOM)

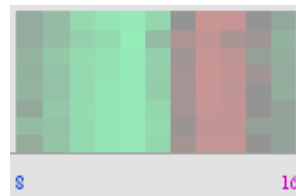
Assign samples to a node in network



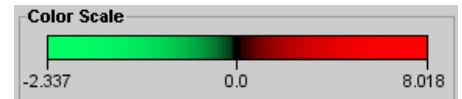
Clustering of genes



+



=



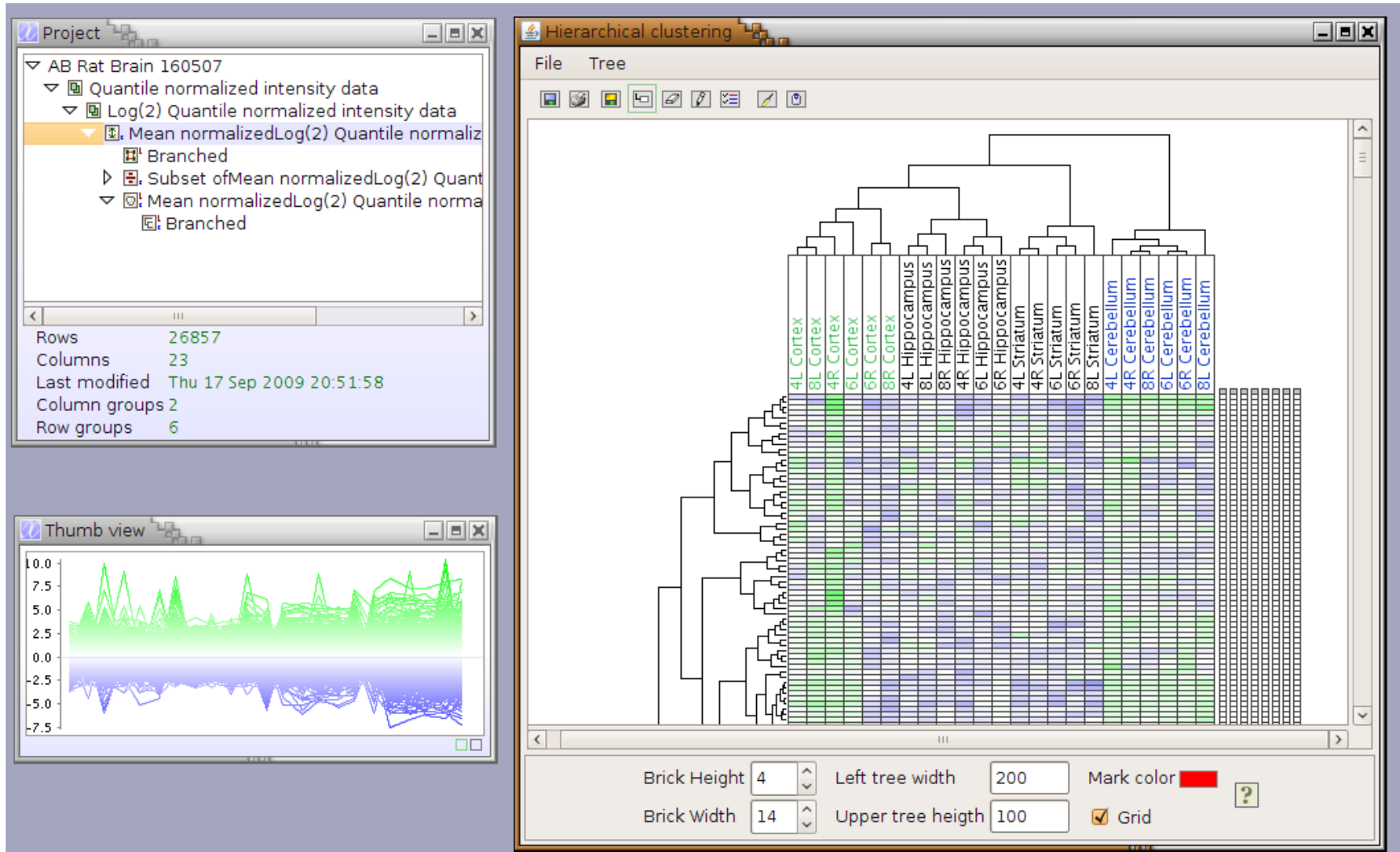
Line chart:
1 line per gene

Heat diagram
1 row per gene

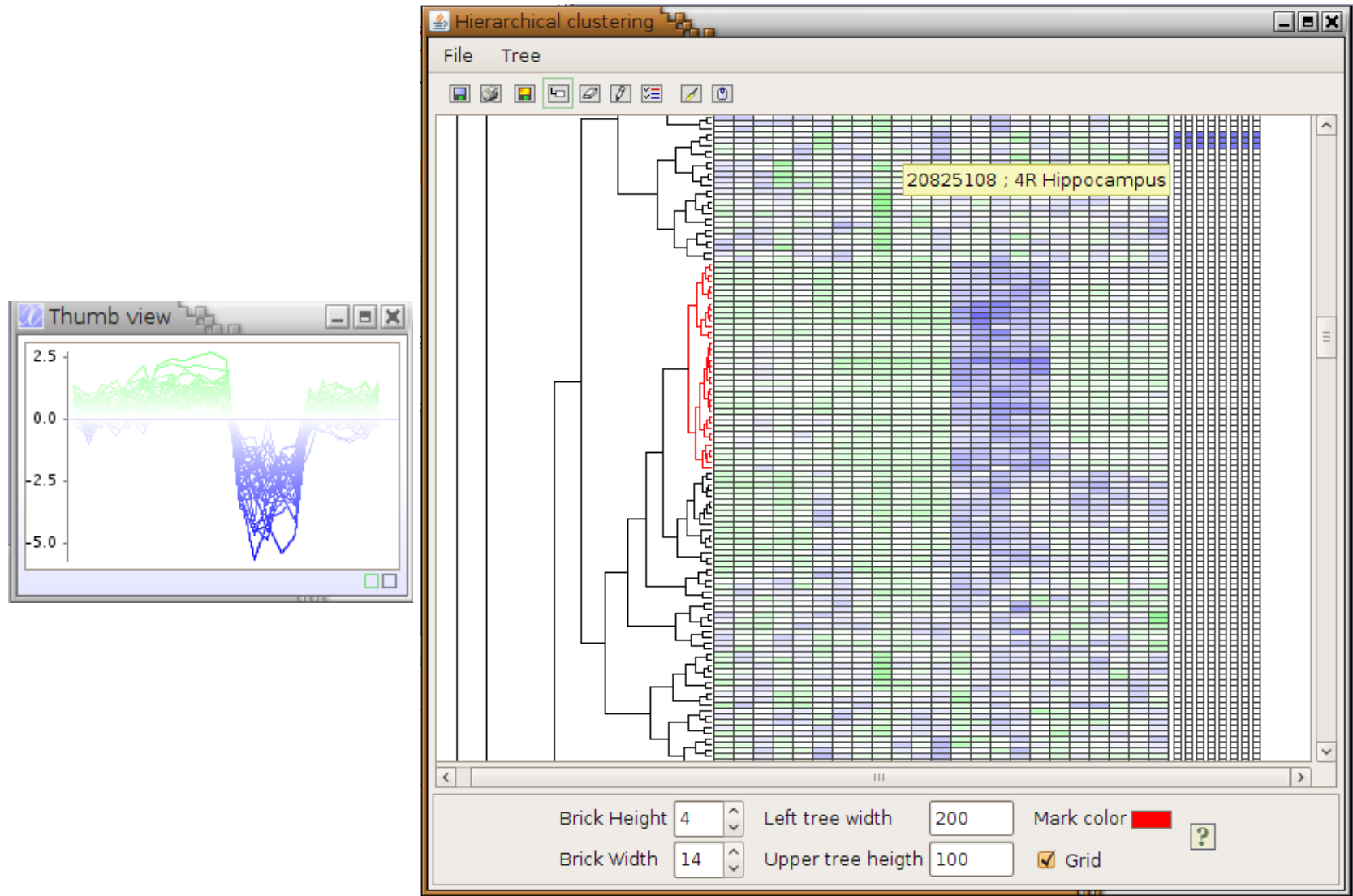
Number of
members

Cluster number

Hierarchical clustering

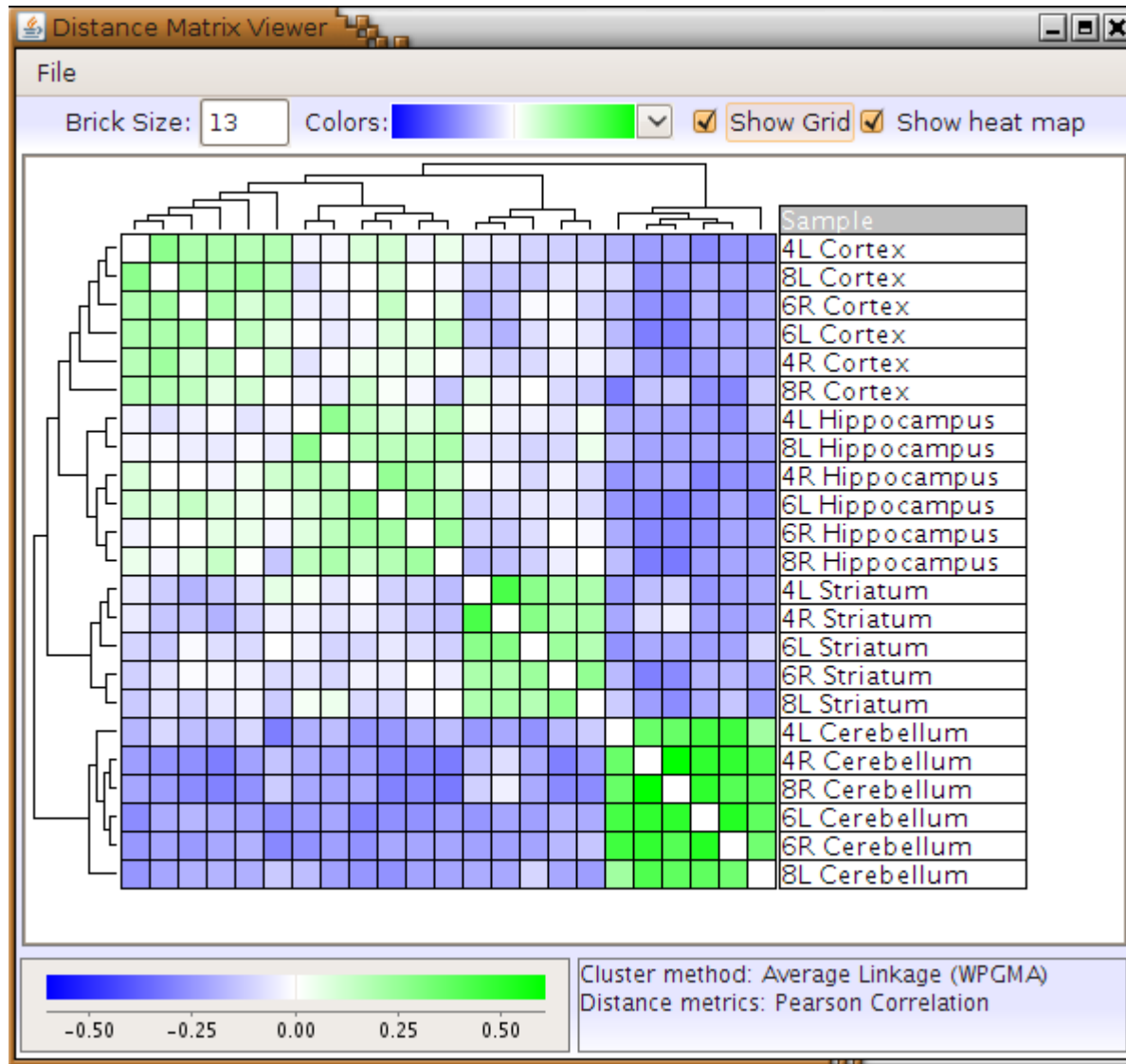


Cluster of genes

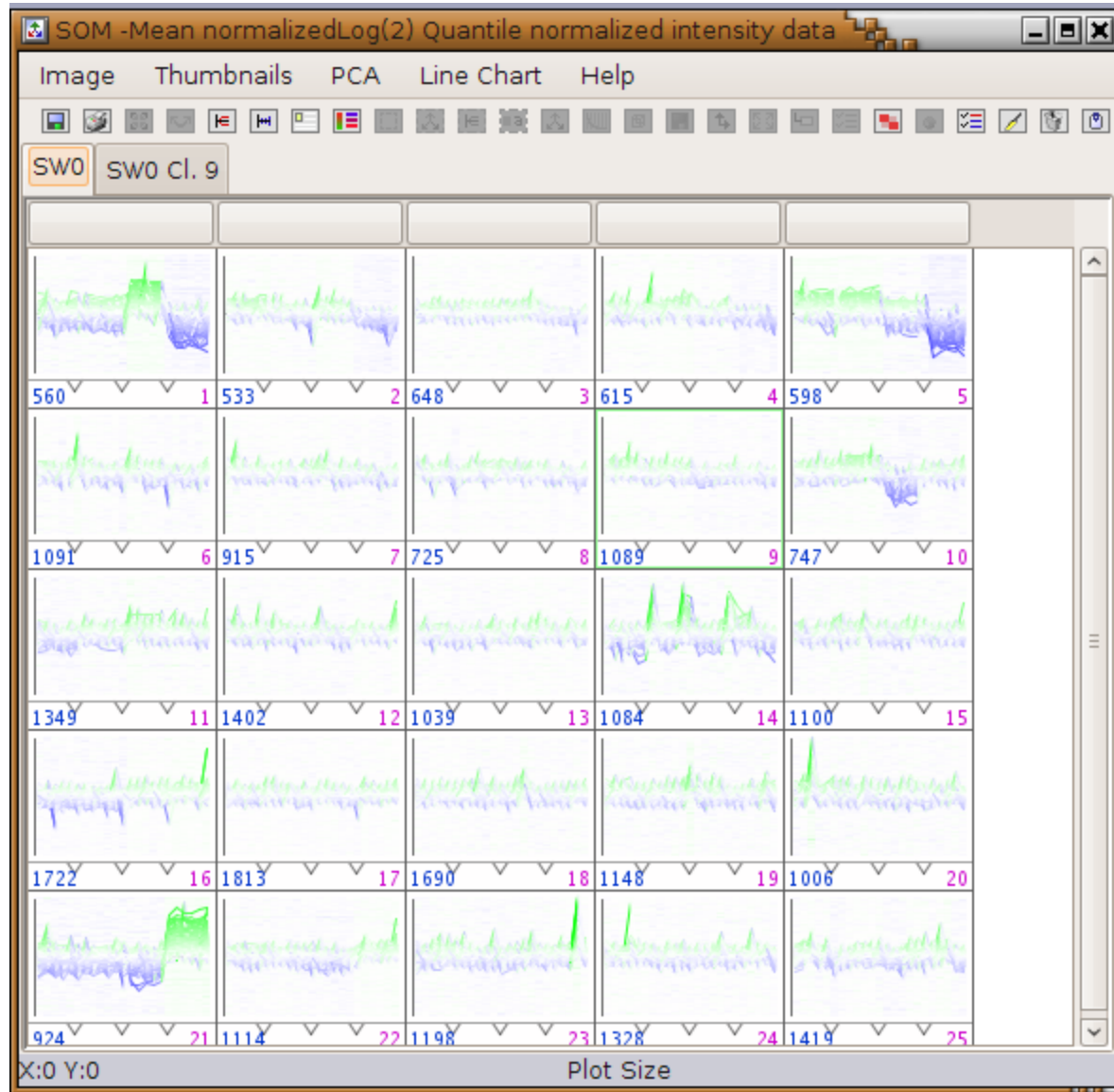


microarray.no

Hierarchical Clustering of Samples



Self Organizing Maps (SOM)



microarray.no



Projection

Dimension reduction

Find variables that are responsible for the largest variation in a dataset;

Is this because of treatment / tissue origin / developmental stage etc...?

Or due to some previously unknown, important factor?

(Or due to technical artefacts?)

Projection

Explanation (VERY simplified):

Gene expression table is decomposed into structure and noise

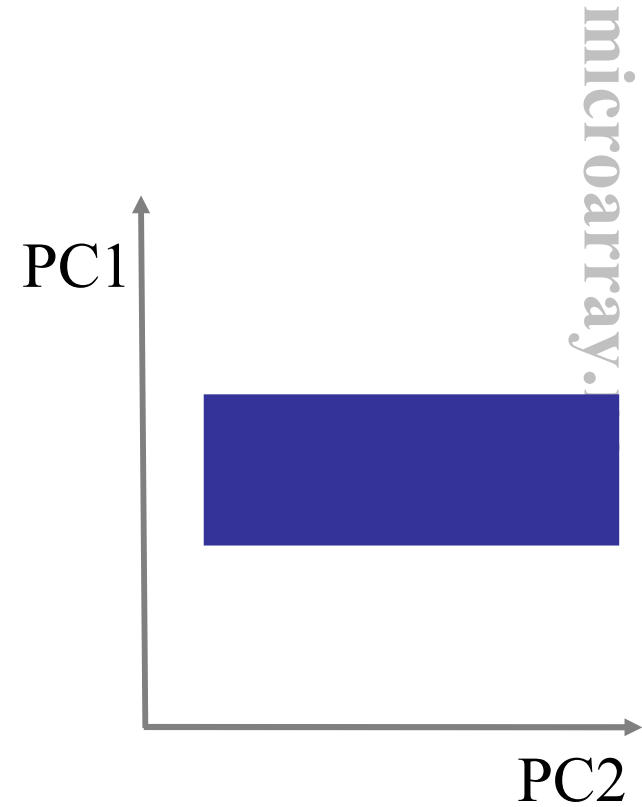
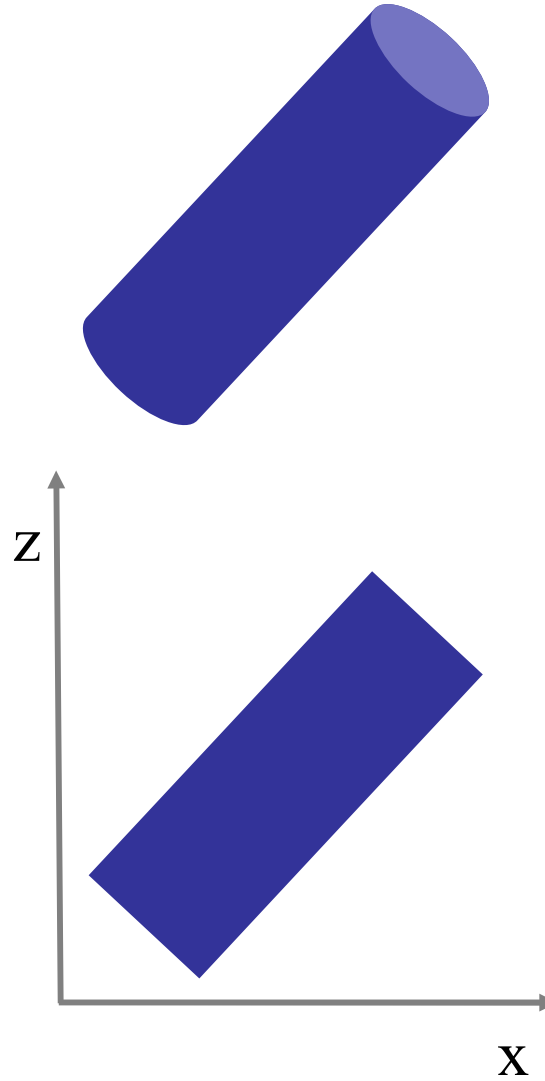
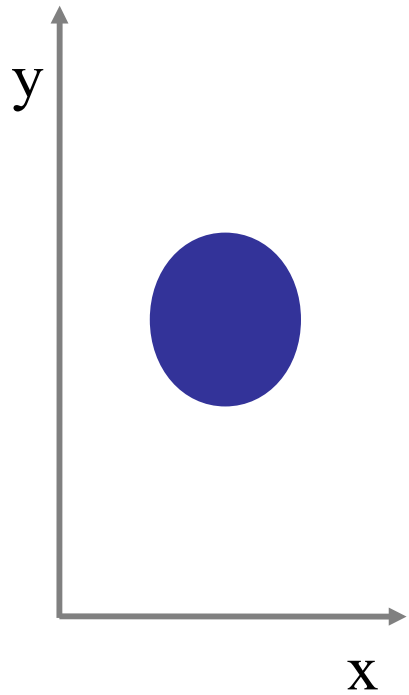
A new coordinate system is generated for the data

Data are represented as points in a multidimensional space

Find the axis that span the most variation

Project the points onto these

Projection



Projection

		Kaffe	Te	Kakao	Brennevi	Vin	Øl
		1	2	3	4	5	6
Norge	1	9.8000	0.2100	0.6100	1.1000	6.4000	52.0000
Danmark	2	10.4000	0.3900	0.5400	1.4000	20.7000	123.2000
Finland	3	12.4500	0.1700	3.0000e-02	3.1000	5.4000	79.0000
Island	4	8.2700	0.2300	0.0000	2.2000	5.2000	23.7000
Sverige	5	10.7100	0.3200	0.1600	1.8000	12.3000	57.4000
Belgia_o	6	7.8700	0.1700	5.5400	m	m	m
Frankrik	7	5.4900	0.2000	1.1800	2.5000	73.8000	40.5000
Hellas	8	1.4200	5.0000e-02	0.3500	m	30.8000	39.1000
Irland	9	0.5500	3.1400	2.7600	1.4000	3.9000	114.0000
Italia	10	4.6700	8.0000e-02	0.9700	1.0000	67.0000	22.7000
Jugoslav	11	3.1000	0.1100	0.5900	1.6000	20.3000	46.8000
Nederlan	12	10.9700	0.8200	15.3500	2.0000	14.7000	87.0000
Polen	13	1.4000	0.5400	0.5600	4.3000	7.6000	30.8000
Portugal	14	3.0800	3.0000e-02	2.0000e-02	0.8000	51.5000	60.7000
Sovjetun	15	0.3000	0.9800	0.5600	1.9000	6.6000	19.3000
Spania	16	4.2500	3.0000e-02	1.1100	2.8000	38.3000	70.7000
Sveits	17	9.4000	0.2500	3.0600	1.9000	49.6000	69.3000
Storbrit	18	2.0600	2.6200	2.7800	1.8000	11.5000	110.5000
Tsjekkos	19	2.2000	0.1300	1.2100	3.3000	13.6000	132.9000
Tyskland	20	8.9700	0.2200	3.9400	2.0000	26.0000	143.0000
Ungarn	21	2.2700	7.0000e-02	0.8500	4.6000	21.5000	103.9000
Østerrik	22	10.2200	0.1600	1.8000	1.5000	34.8000	119.5000
Australi	23	2.5200	1.1600	m	1.3000	18.3000	111.0000
New_Zeal	24	1.9200	1.4600	3.0000e-02	1.4000	14.6000	114.2000

microarray.no

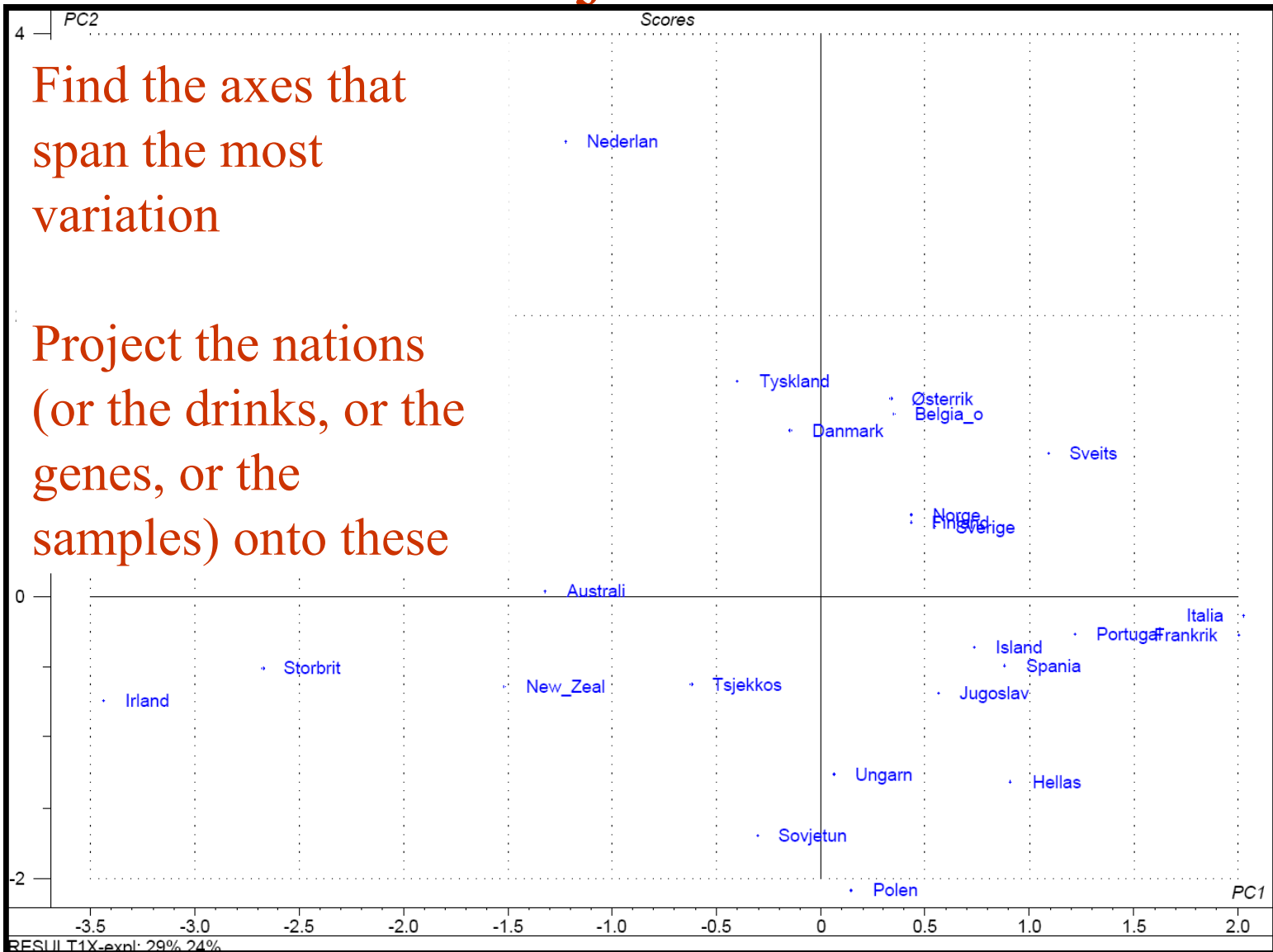
Borrowed
from Endre
Anderssen,
NTNU



Projection

Find the axes that span the most variation

Project the nations (or the drinks, or the genes, or the samples) onto these



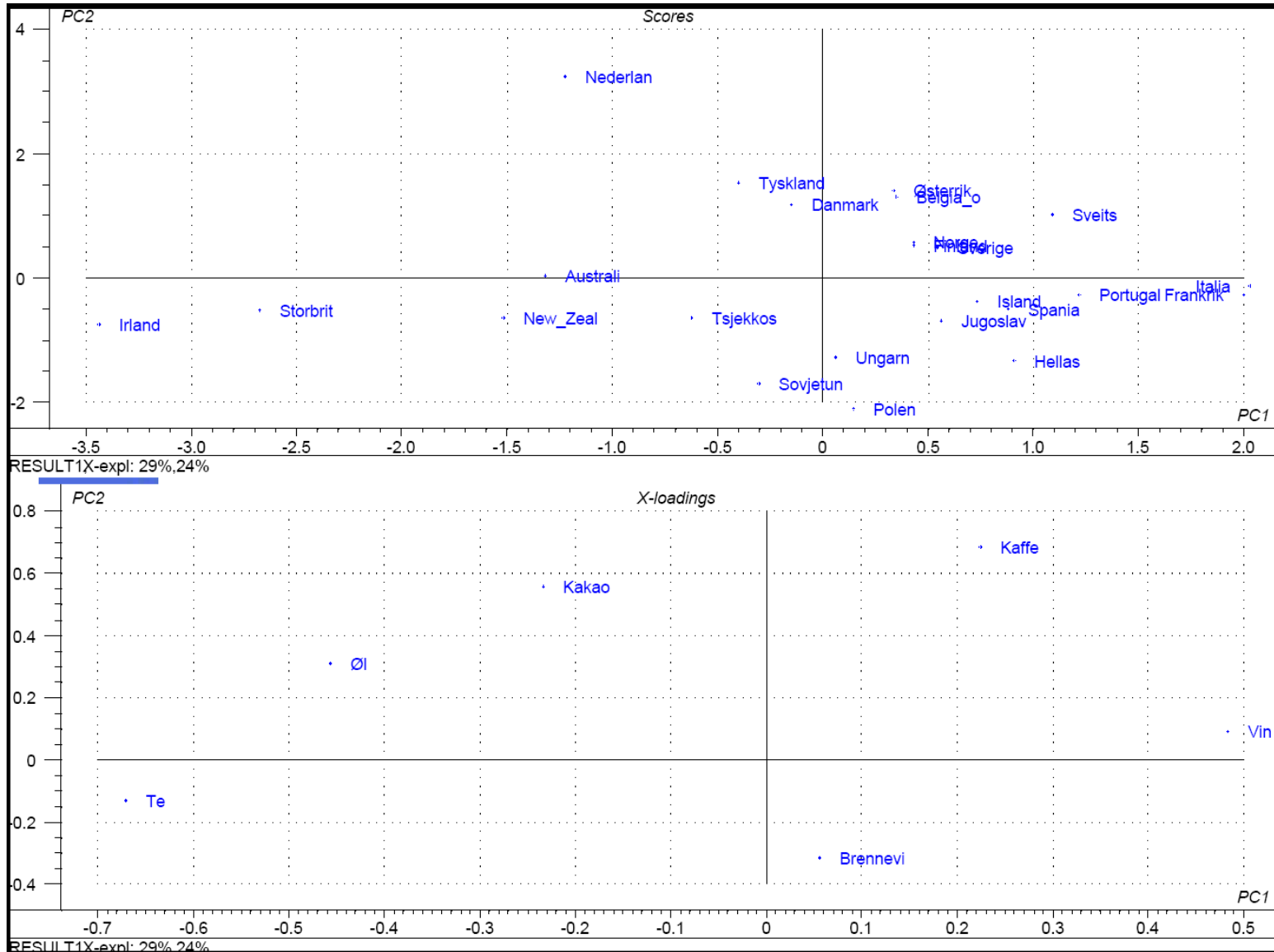
Projection

Find the axes that span the most variation

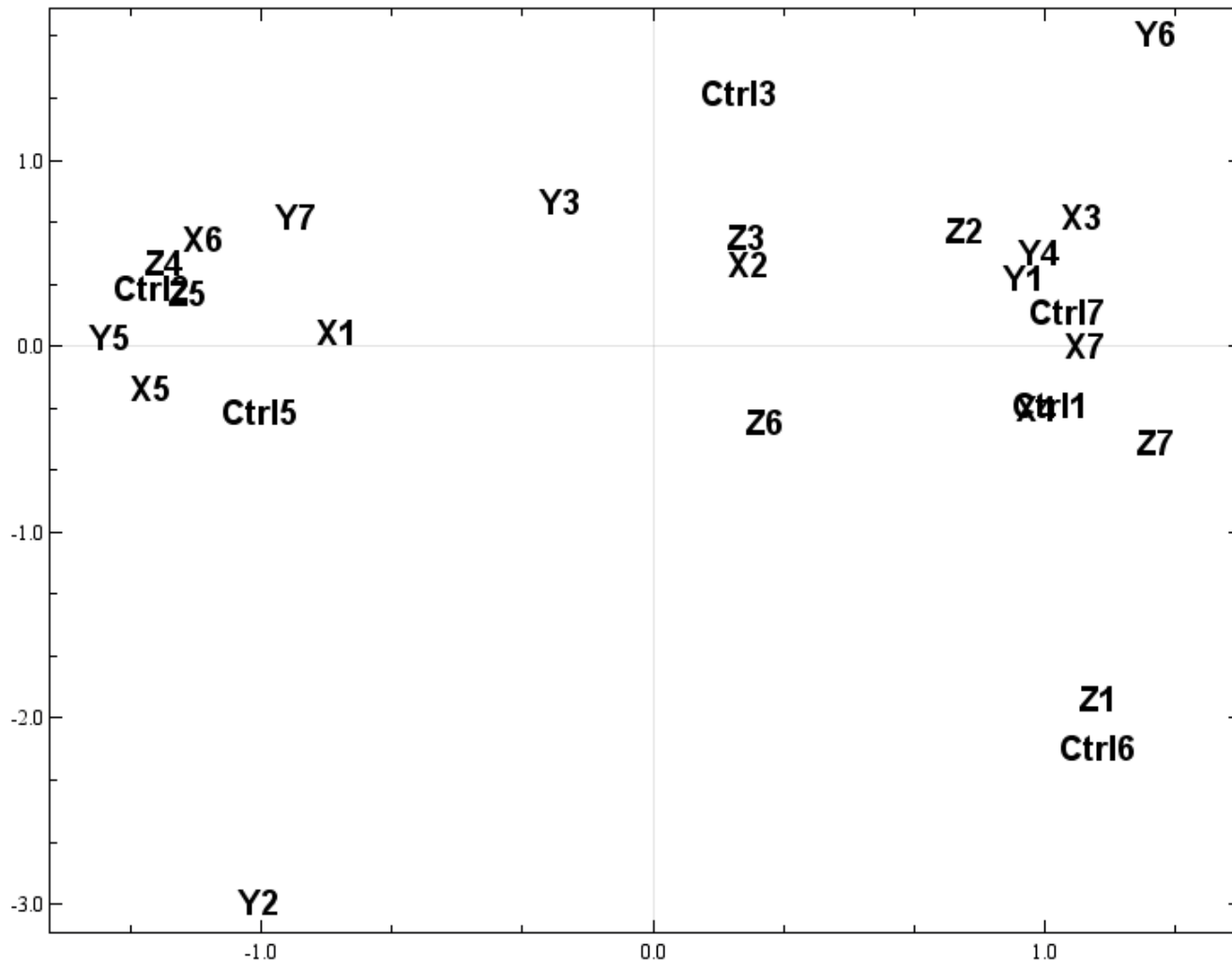
Project the drinks (or the nations, or the genes, or the samples) onto these



Projection

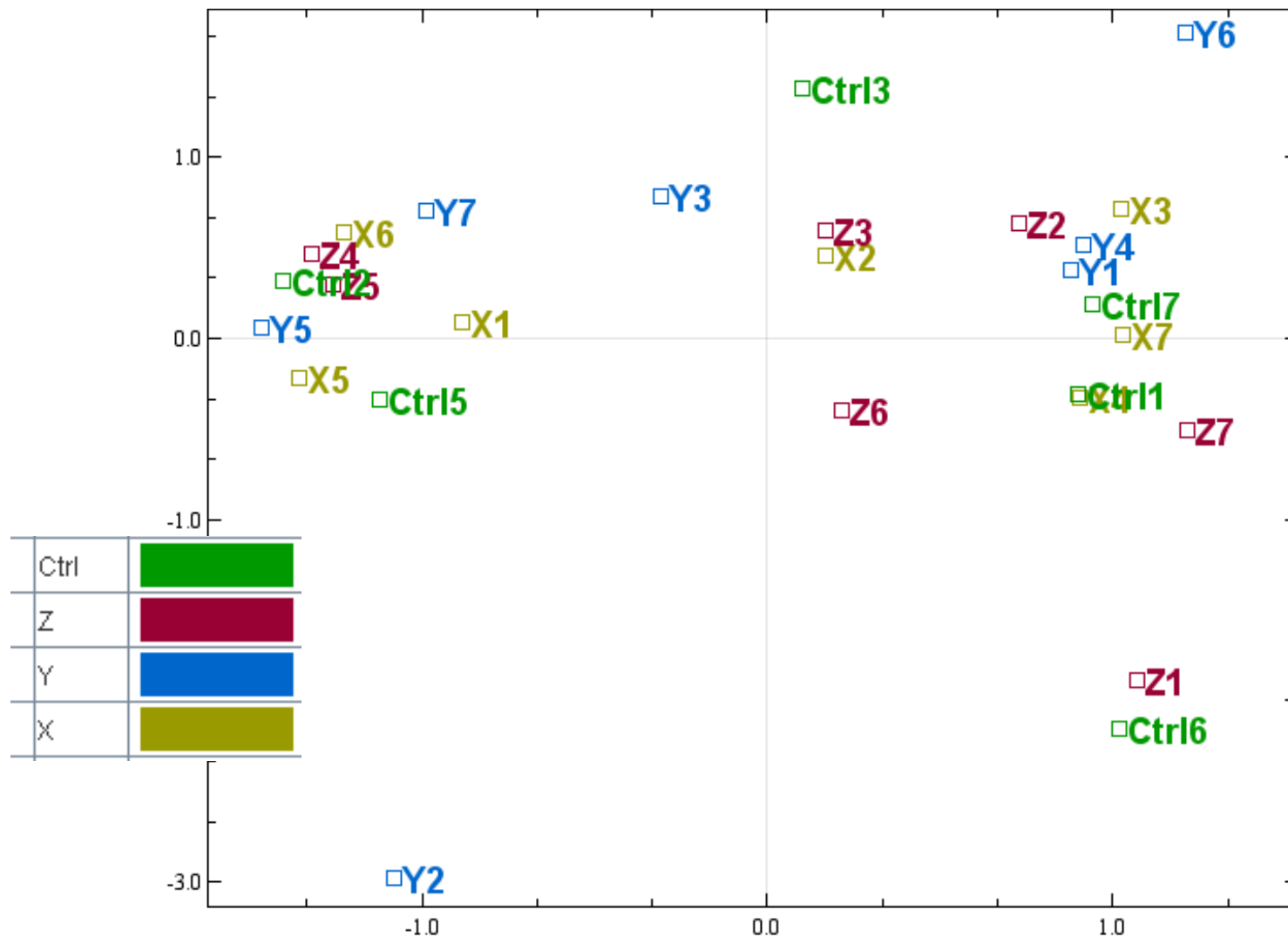


Projection as quality control - 1



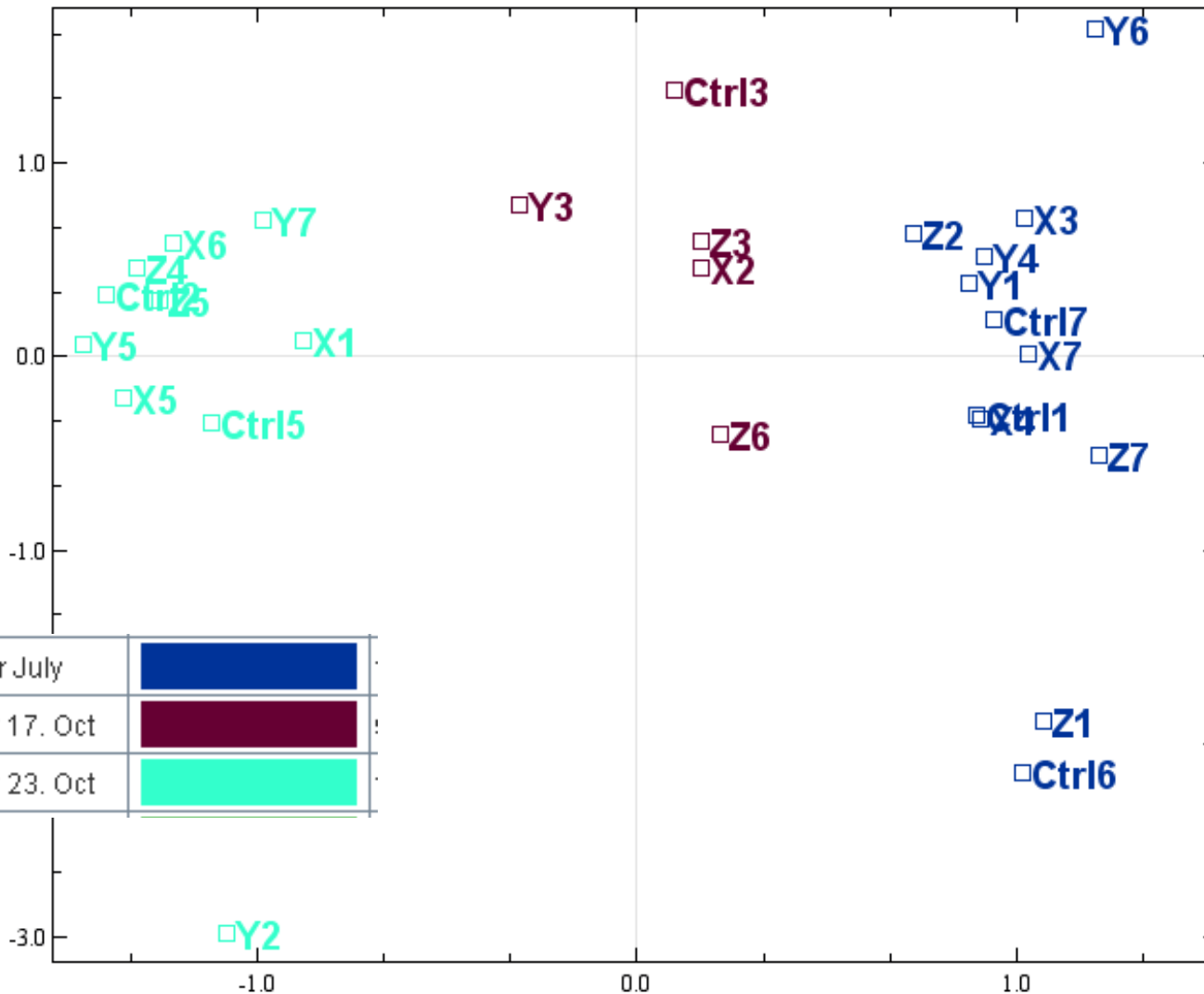
microarray.no

Projection as quality control - 2






microarray.no

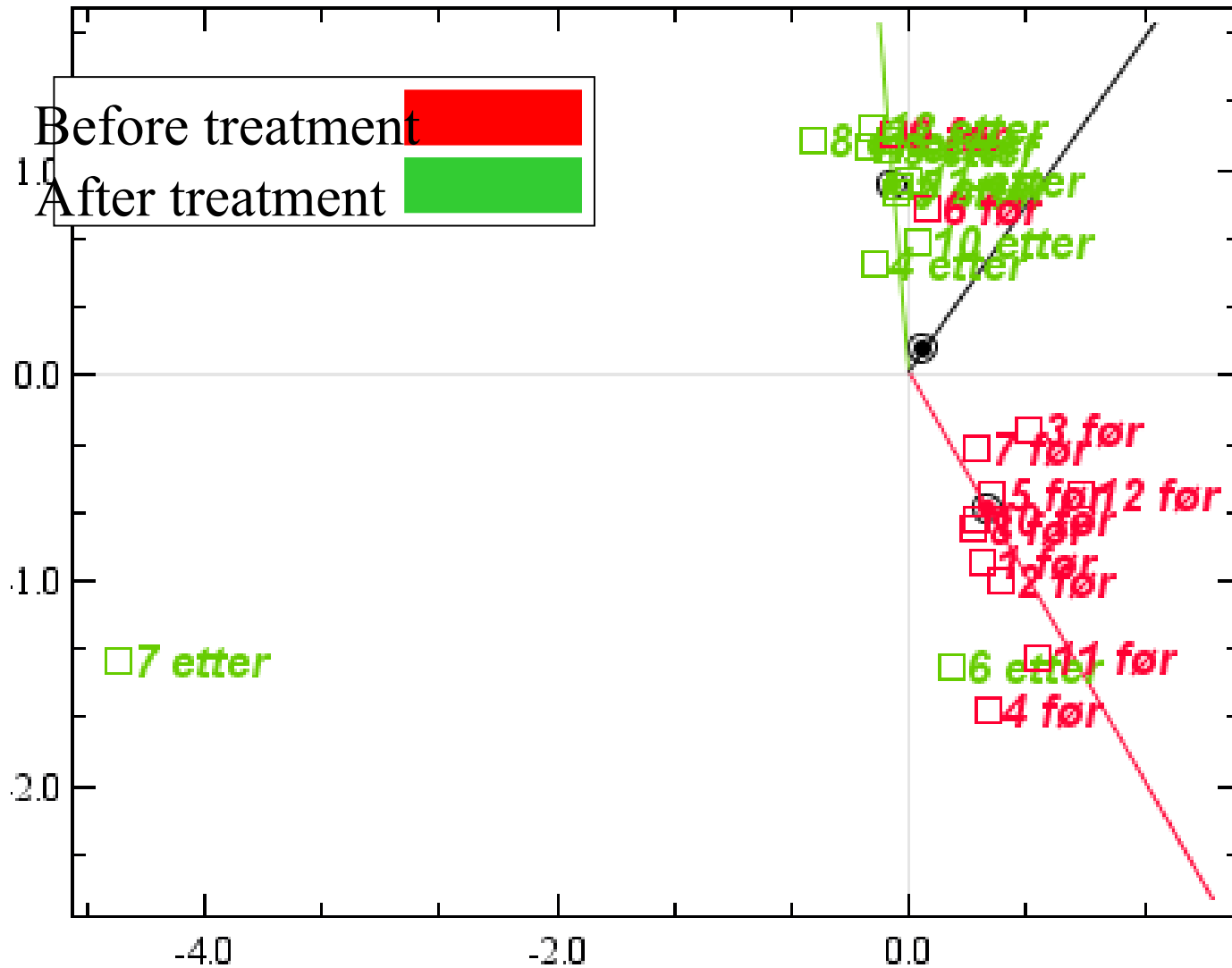
Projection as quality control - 3



microarray.no

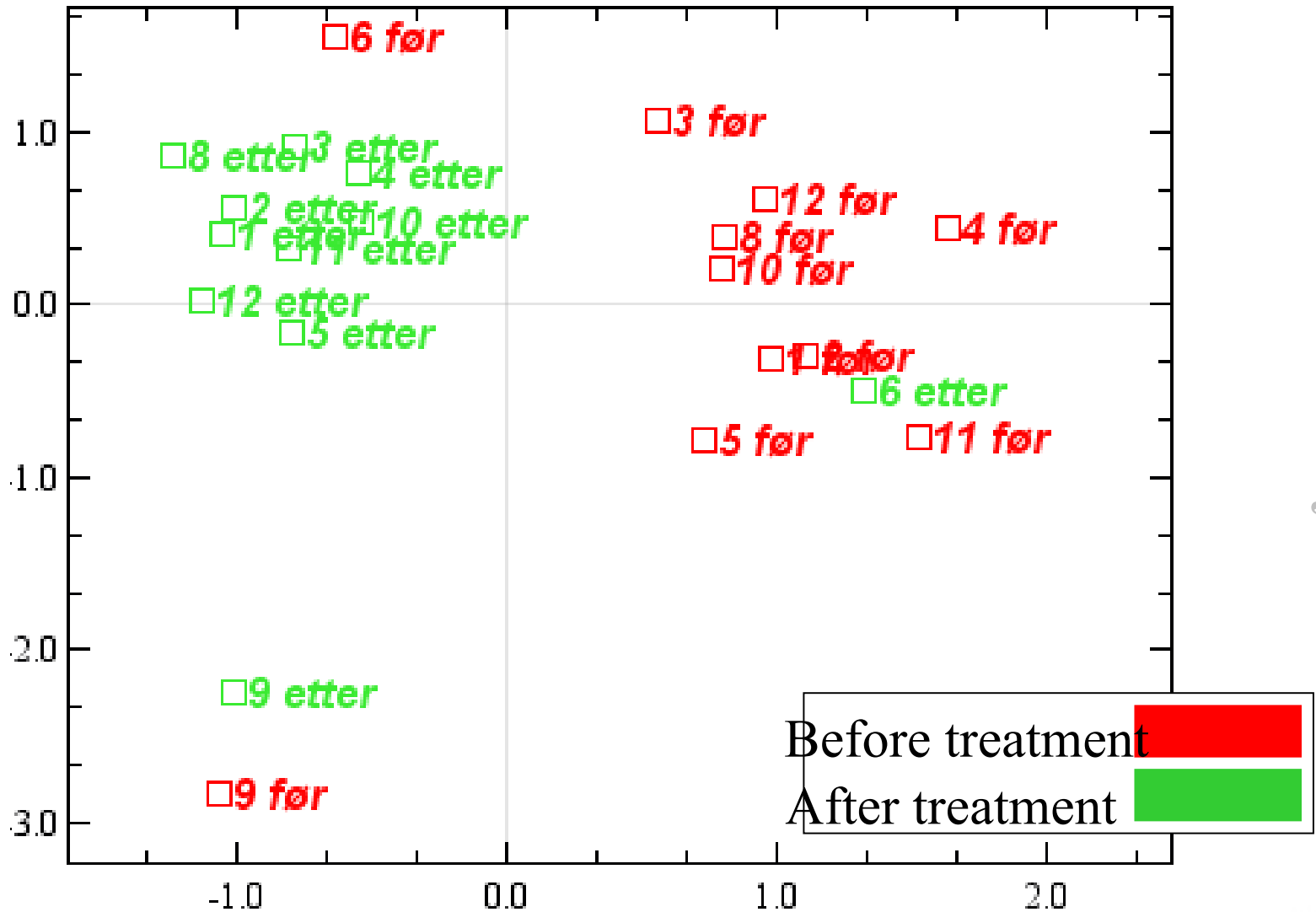
Labelled July - Hybr July	
Labelled Oct - Hybr 17. Oct	
Labelled Oct - Hybr 23. Oct	

Projection as quality control - 4

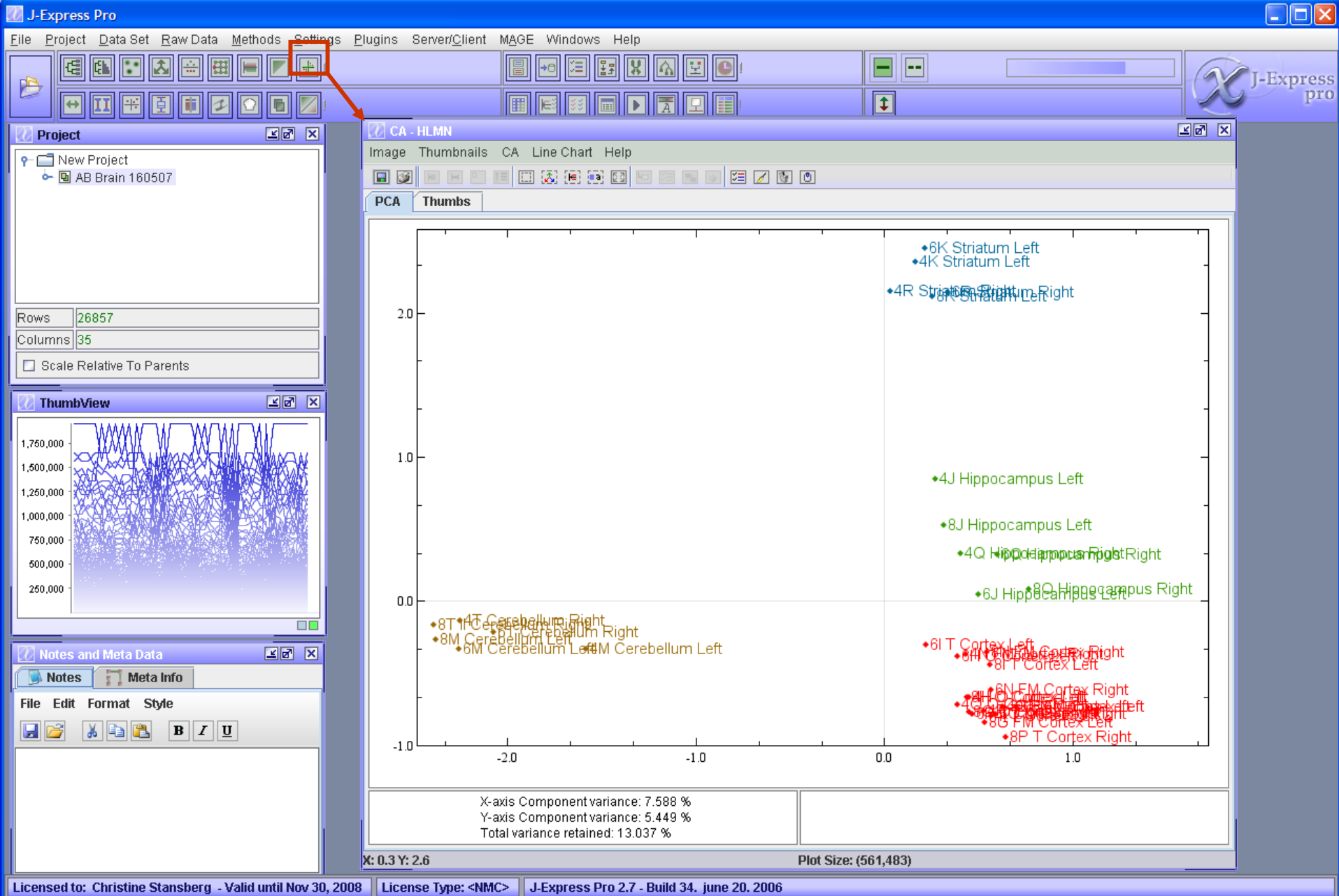


microarray.no

Projection as quality control - 5



Overview of expression data



Projection

Motivation:

Large datasets with a lot of similar looking (low expressed) profiles are common → Use projection to see similar and dissimilar groups of profiles.

The screenshot displays the J-Express Pro software interface with several windows open:

- Project Window:** Shows a tree view with 'New Project', 'AB Brain 160507', 'HLMN', and 'HLMN Filtered'. Below it, 'Rows' is 26857 and 'Columns' is 35. A 'Scale Relative To Parents' checkbox is present.
- ThumbView Window:** Displays a line chart with a y-axis from -7.5 to 10.0. The chart shows multiple profiles, with a red vertical bar highlighting a specific region.
- PCA - HLMN Filtered Window:** Shows a heatmap of the filtered data with a red vertical bar.
- PCA - HLMN Window:** Displays a scatter plot of Principal Component 1 (X-axis, -10.0 to 20.0) versus Principal Component 2 (Y-axis, -10.0 to 2.0). The plot is annotated with brain region names and their corresponding PCA coordinates:
 - 6K Striatum Left
 - 4K Striatum Left
 - 4R Striatum Right
 - 4L Striatum Left
 - 4J Hippocampus Left
 - 8J Hippocampus Left
 - 4Q Hippocampus Right
 - 6J Hippocampus Right
 - 8T Cerebellum Right
 - 8M Cerebellum Right
 - 8L Cerebellum Left
 - 8M Cerebellum Left
 - 6I T. Cortex Left
 - 8P T. Cortex Right
 - 8P T. Cortex Left
 - 8L M. Cortex Right
 - 8L M. Cortex Left
 - 8G M. Cortex Left
 - 8P T. Cortex Right

Statistical information for the PCA plot:

- X-axis Component variance: 7.588 %
- Y-axis Component variance: 5.449 %
- Total variance retained: 13.037 %

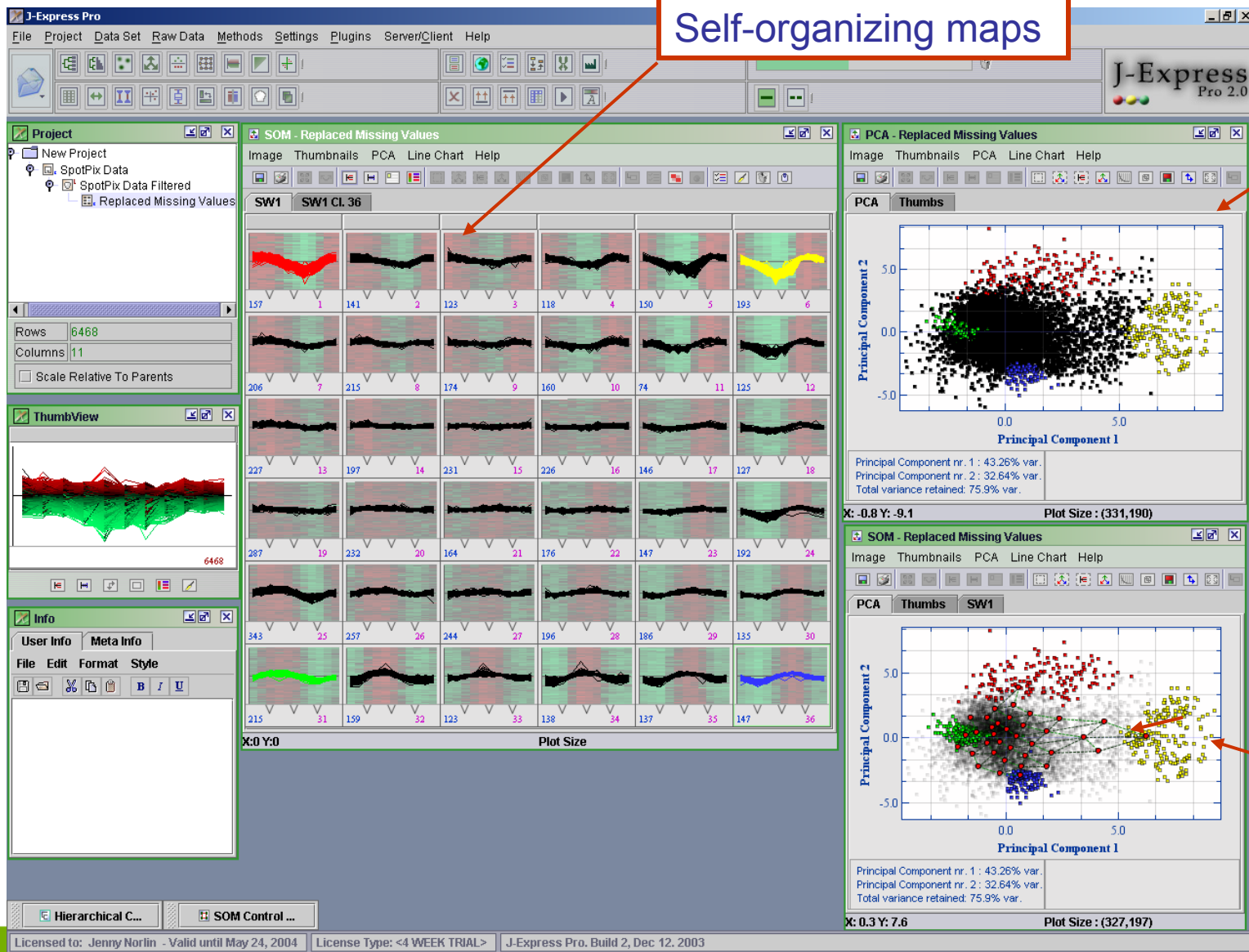
Plot Size: (561,483)

Principal Component nr. 1 : 1
Principal Component nr. 2 : 9
Total variance retained: 26.9%

X: 0.3 Y: 2.6

X: 6.2 Y: 9.7

The different clustering/projection methods often produce similar results:



Self-organizing maps

Principal component analysis

microarray.no

Self-organizing maps

AND
Principal component analysis

Questions?

microarray.no