Statistics for Bioinformatics

Marseille, Nov 29, 2012

Analysis of microarray data

Web site for this course:
http://www.bigre.ulb.ac.be/courses/statistics_bioinformatics/

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

Introduction

- Regulation of gene expression
  - See Jean Imbert presentation, yesterday
- Microarray technology
  - See Cathy Nguyen presentation next week
  - I will give a very short summary
- Statistical analysis of microarrays
  - With microarrays, relatively complex methods of multivariate analysis became necessary for biologists.
  - Not part of classical introductory courses of statistics.
- Content of this course
  - Microarrays: study cases
  - Selecting differentially expressed genes
  - Measuring differences between gene
  - Unsupervised classification: hierarchical clustering, k-means clustering
  - Supervised classification: discriminant analysis, cross-validation, variable selection

Study cases

DNA chip technology

Cell culture,
tissue, ...

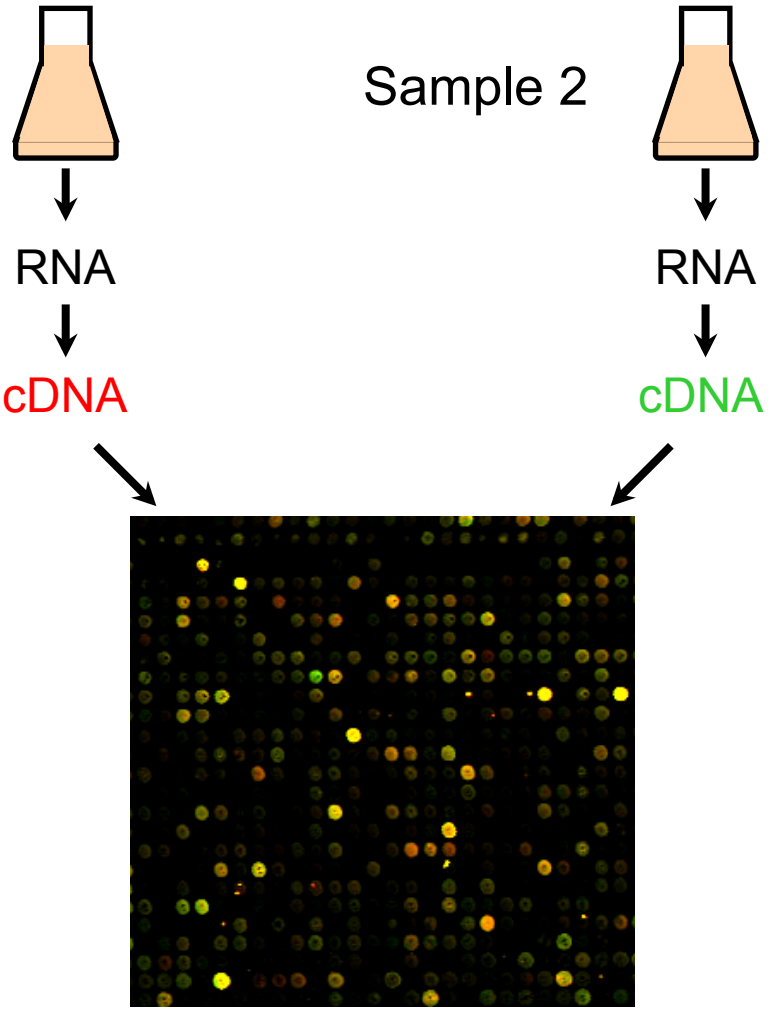
RNA extraction

Synthesis of
fluorescent cDNA

Brightness ↔ Quantity

Color ↔ Specificity

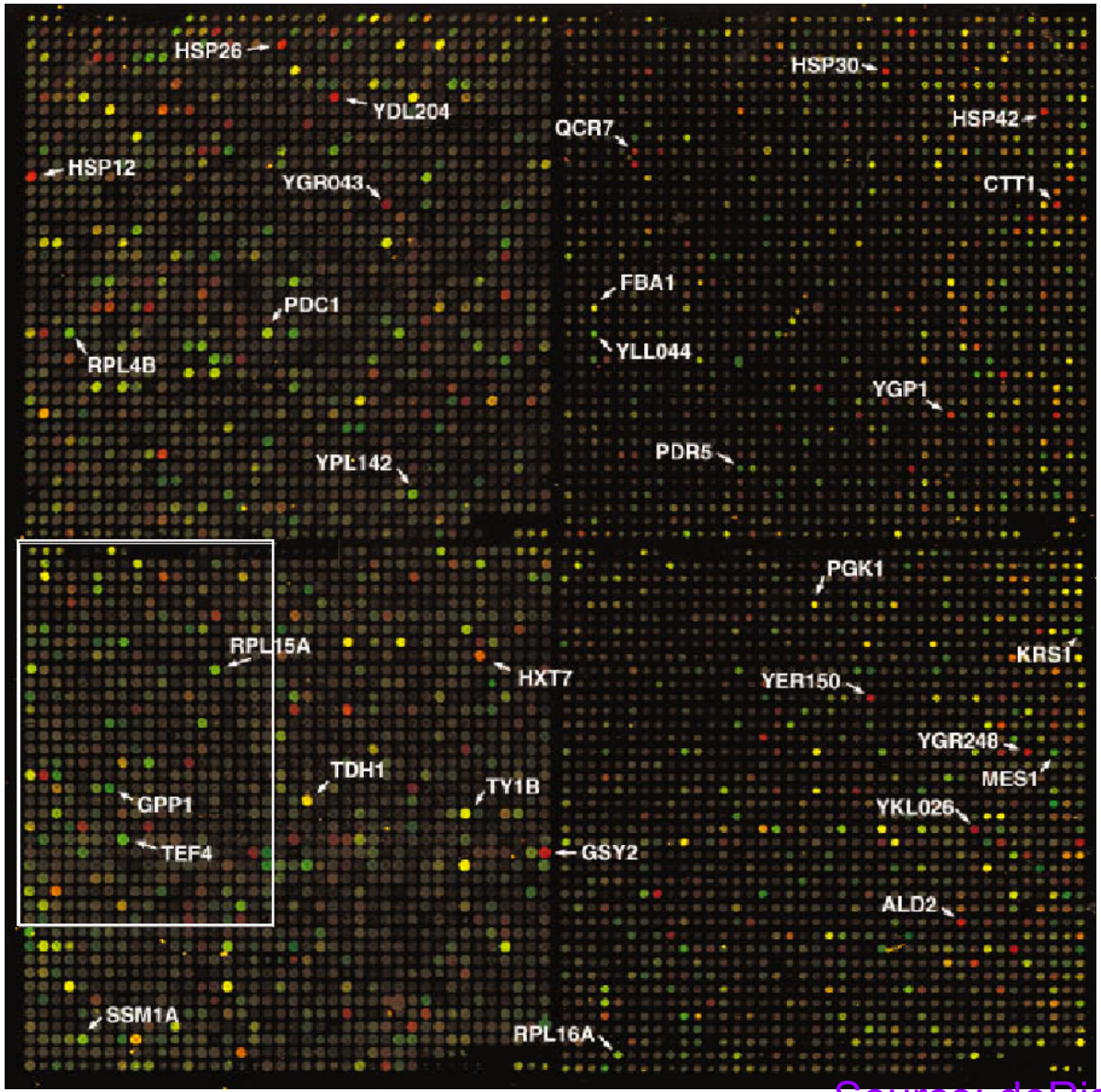
yellowish	not specific
reddish	sample 1 - specific
greenish	sample 2 - specific



DNA chip

Source: deRisi et al., Science 1997

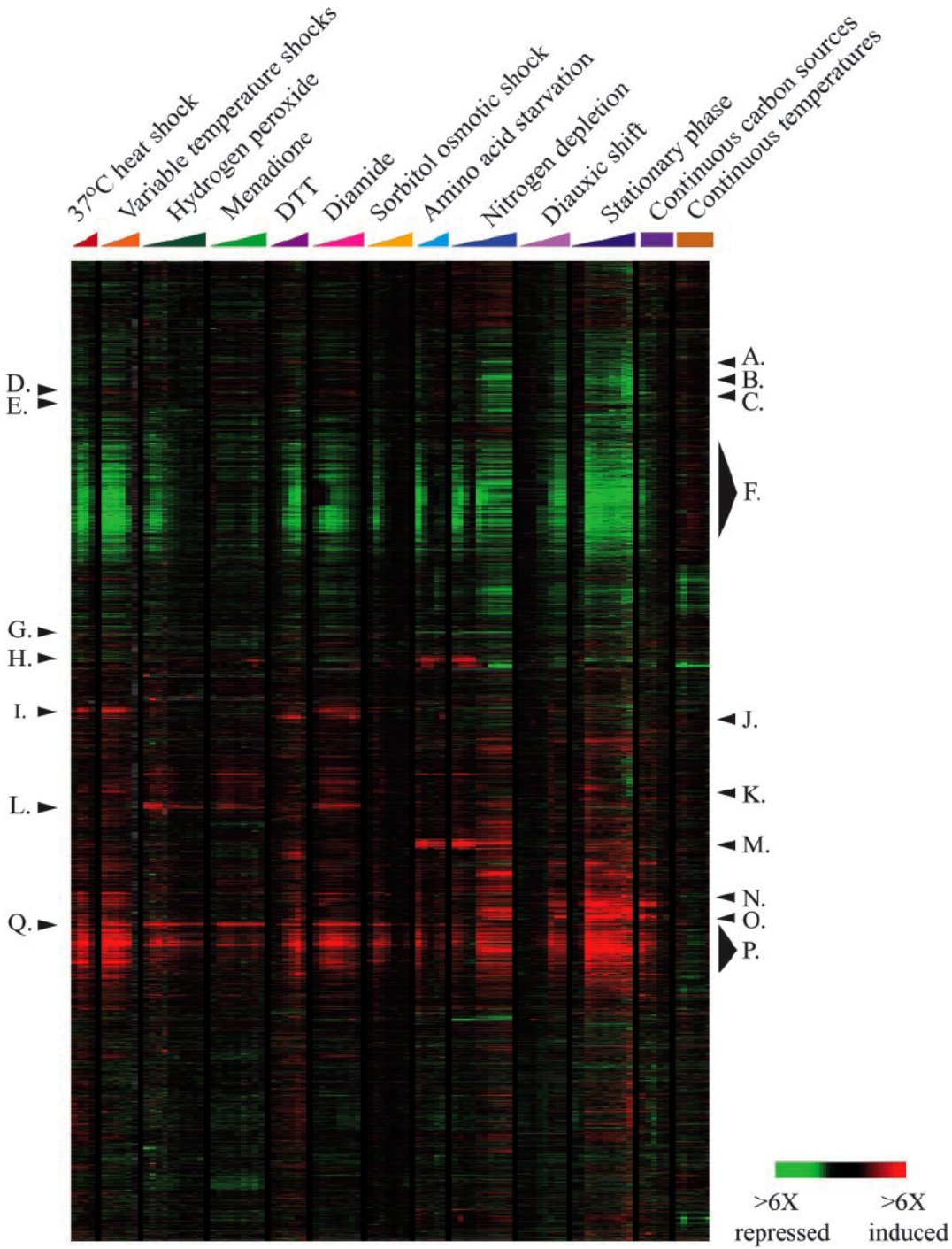
Complete microarray



Source: deRisi et al., Science 1997

Stress response in yeast

- First study case: Gasch et al. (2000
- Rranscriptional response of yeast genome to various stress conditions (heat shock, osmotic shock, …)
  - Drugs
  - Alternative carbon sources
  - …
- Two-channels microarrays.
- Full-length cDNA.



- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eis M. B., Storz, G., Botstein, D. & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241-57.

Den Boer et al., 2009

- Oligonucleotide microarrays (Affymetrix arrays hgu133a)
- Training set: COALL cohort
  - 190 cases of paediatric Acute Lymphoblastic Leukemia
  - Used to define a “transcriptomic signature” for the different subtypes of ALL, using an elaborate procedure relying on an inner and an outer loop of cross-validation.
- Testing set: DCOG cases (107 cases)
  - Used to test the predictive power of the transcriptomic signature.

hyperdiploid	44
pre-B ALL	44
TEL-AML1	43
T-ALL	36
E2A-rearranged (EP)	8
BCR-ABL	4
E2A-rearranged (E-sub)	4
MLL	4
BCR-ABL + hyperdiploidy	1
E2A-rearranged (E)	1
TEL-AML1 + hyperdiploidy	1

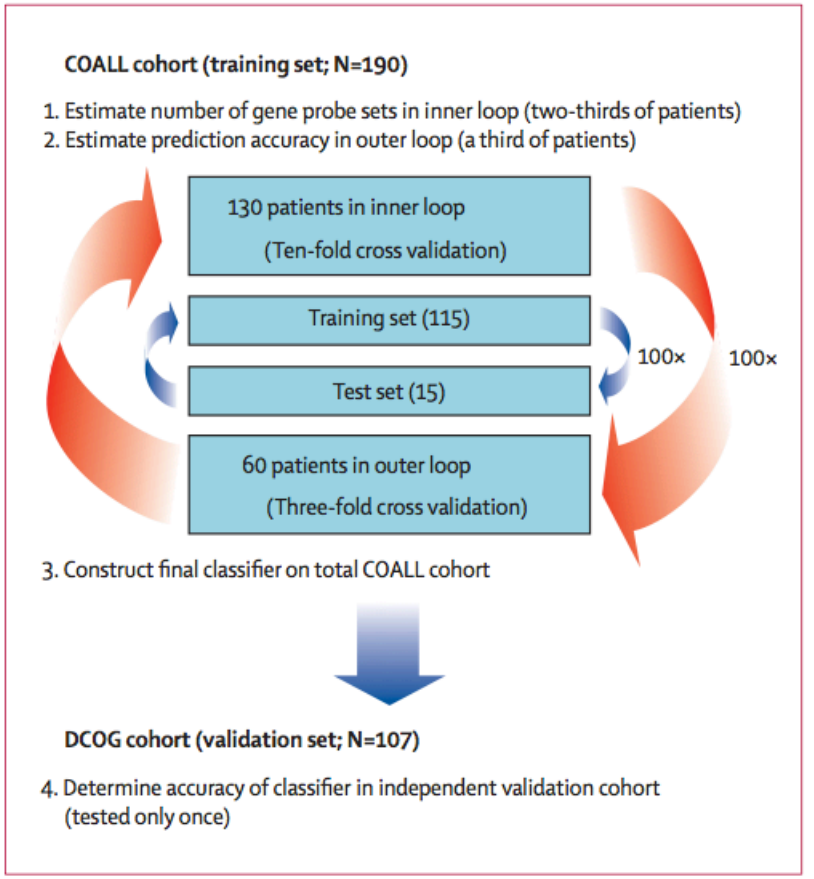


Figure 1: **Identification of a gene-expression signature enabling classification of paediatric ALL**

- Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

The transcriptomic signature

- The training procedure selects 100 gens whose combined expression levels can be used to assign samples to cancer subtypes.
- The heatmaps show that the selected genes are differentially expressed
  - between subtypes of the training set (left);
  - between subtypes of the testing set (right).

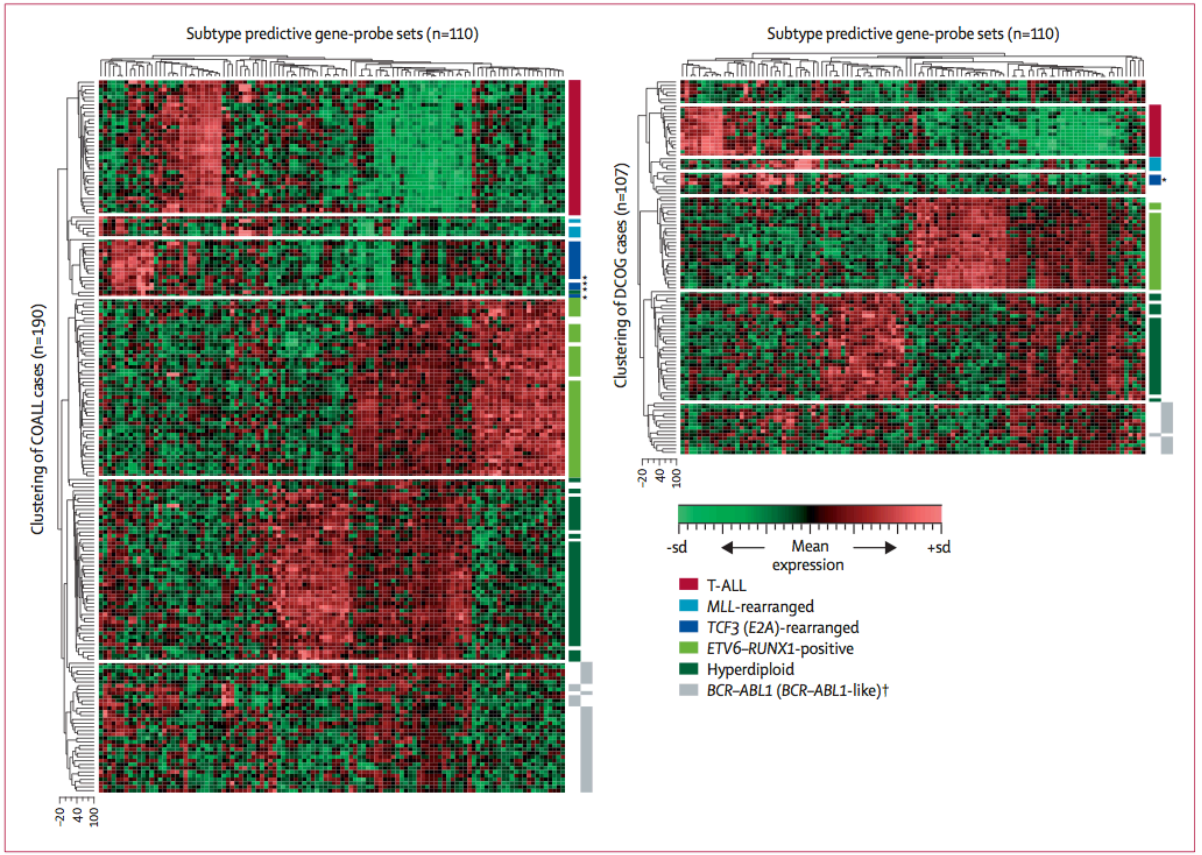


Figure 2: **Clustering of ALL subtypes by gene-expression profiles**
Hierarchical clustering of patients from the COALL (left) and DCOG (right) studies with 110 gene-probe sets selected to classify paediatric ALL. Heat map shows which gene-probe sets are overexpressed (in red) and which gene probe sets are underexpressed (in green) relative to mean expression of all gene-probe sets (see scale bar). *Patients with *E2A*-rearranged subclone (15–26% positive cells). †Right column of grey bar denotes *BCR–ABL1*-like cases.

- Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Accuracy of the classifier

- The signature has an excellent diagnostic value: for the well-represented cancer types, the sensitivity and specificity are >90%.
- **Note:** accuracy is misleading some subtypes have 98% accuracy with 0% sensitivity.

hyperdiploid	44
pre-B ALL	44
TEL-AML1	43
T-ALL	36
E2A-rearranged (EP)	8
BCR-ABL	4
E2A-rearranged (E-sub)	4
MLL	4
BCR-ABL + hyperdiploidy	1
E2A-rearranged (E)	1
TEL-AML1 + hyperdiploidy	1

$$Sn = TP / (TP + FN)$$

$$Sp = TN / (TN + FP)$$

$$PPV = TP / (VP + FP)$$

	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Accuracy (%)
T-lineage ALL	100 (100–100)	100 (100–100)	100 (100–100)	100 (100–100)	100 (100–100)
ETV6-RUNX1-positive	100 (100–100)	97·8 (95·7–97·8)	93·3 (87·5–93·3)	100 (100–100)	98·3 (96·7–98·3)
Hyperdiploid	100 (92·9–100)	97·8 (95·7–97·8)	92·6 (86·7–93·3)	100 (97·8–100)	96·7 (95·0–98·3)
E2A-rearranged	100 (75·0–100)	100 (98·2–100)	100 (80·0–100)	100 (98·2–100)	98·3 (98·3–100)
BCR–ABL1-positive	0 (0–0)	100 (100–100)	0 (0–0)	98·3 (98·3–98·3)	98·3 (98·3–98·3)
MLL-rearranged	0 (0–0)	100 (100–100)	0 (0–0)	98·3 (98·3–98·3)	98·3 (98·3–98·3)
Overall values	93·5 (93·5–95·7)	78·6 (78·6–85·7)	93·6 (93·2–95·6)	80·0 (76·4–84·6)	90·0 (88·3–91·7)

Data from the COALL study. Data are median (25th–75th percentile). Accuracy is for 100 iterations that include 130 cases to build the classifier and 60 other patients to determine the diagnostic test values in each interation (three-fold cross validation). Overall values based on the classification of all cases, including the B-other group.

Table 1: **Diagnostic test values for the classification of acute lymphoblastic leukaemia by three-fold cross-validation approach**

	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
T-lineage ALL	15/15 (100%)	92/92 (100%)	15/15 (100%)	92/92 (100%)	107/107 (100%)
ETV6–RUNX1-positive	24/24 (100%)	81/83 (97·6%)	24/26 (92·3%)	81/81 (100%)	105/107 (98·1%)
Hyperdiploid	28/28 (100%)	74/79 (93·7%)	28/33 (84·8%)	74/74 (100%)	102/107 (95·3%)
E2A-rearranged	2/2 (100%)	104/105 (99·0%)	2/3 (66·7%)	104/104 (100%)	106/107 (99·1%)
BCR–ABL1-positive	0/1 (0%)	106/106 (100%)	0/0	106/107 (99·1%)	106/107 (99·1%)
MLL-rearranged	0/4 (0%)	103/103 (100%)	0/0	103/107 (96·3%)	103/107 (96·3%)
Overall values	69/74 (93·2%)	25/33 (75·8%)	69/77 (89·6%)	25/30 (83·3%)	94/107 (87·9%)

Data are number of predicted cases/total per subtype (%). DCOG cohort (107 patients) used to validate independently the predictive value of classification by gene expression signature (tested only once). Overall values based on the classification of all cases, including the B-other group. The specificity, positive predictive value, and accuracy are 100% for *E2A*-rearranged cases if the B-other case with an *E2A*-rearranged subclone (21% positive cells) is included as true positive case (webappendix).

Table 2: **Diagnostic test values for independent validation group**

- Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Goal of the course

Goal of this course

- During this course, we will introduce the statistical methods used to analyse microarray profiles.
- The analysis will be mainly based on the dataset from Den Boer.
- However, since our goal is to discuss fundamental concepts, we will not strictly follow their complex procedure.
- We will rather explain the different questions that can be made, and the classical methods to address them.

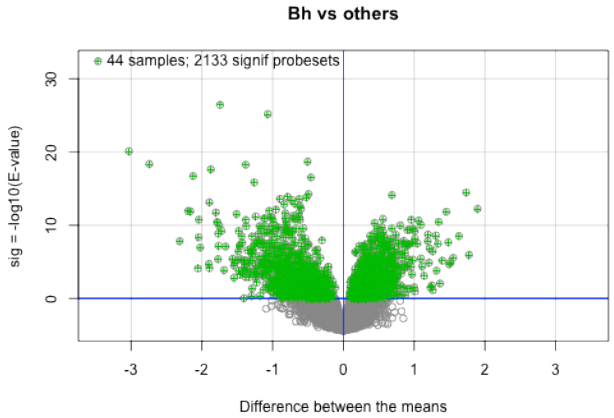
Selecting differentially expressed genes

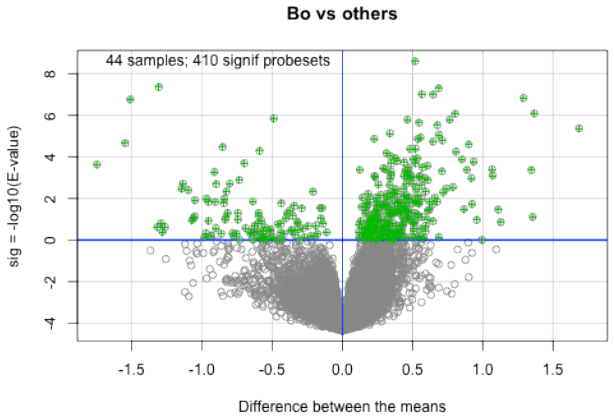
Principle of differential analysis

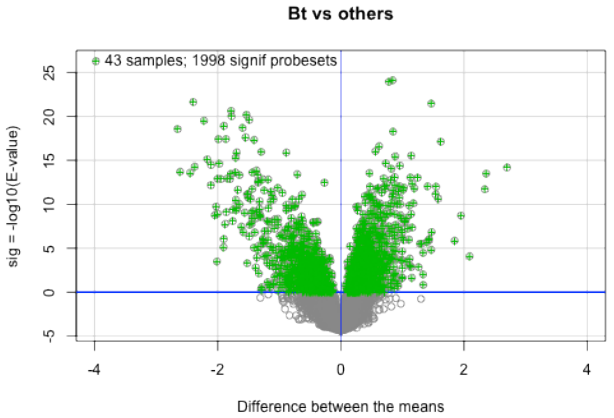
- **Two-groups differential analysis**
  - Tests: T-test de Student, Welch, Wilcoxon.
  - Principle: define a group of interest ("goi", for example hyperdiploidy), and compare it to all other cancer subtypes.
  - For each gene *I,* test the null hypothesis of mean equality
    - H_0: $m_{i,goi} = m_{i,others}$
    - H_A: $m_{i,goi} <> m_{i,others}$
  - A priori, we expect that differential expression will cause a difference between group variances -> we apply Welch rather than Student test.
- **Multi-groups differential analysis**
  - Test: ANOVA
  - Test the hypothesis of mean equality between all groups.
  - For each gene, analyze the variance and compare the inter-group variance with the intra-group (residual) variance.
- **Multiple testing corrections**
  - The data set from Den Boer (2009) contains 22,283 probes. We are thus challenging 22,283 times the risk of false positive (considering a gene as significant whereas it is "truly null").
  - Different methods have been proposed to control the number of false positives:
    - Bonferoni correction : decrease the significance threshold to alpha / N
    - E-value: compute the expected number of false positives: e-value = p-value * N
    - FWER: compute P(FP >= 1)
    - q-value: estimate the false discovery rate (proportion of FP among the genes declared significant).

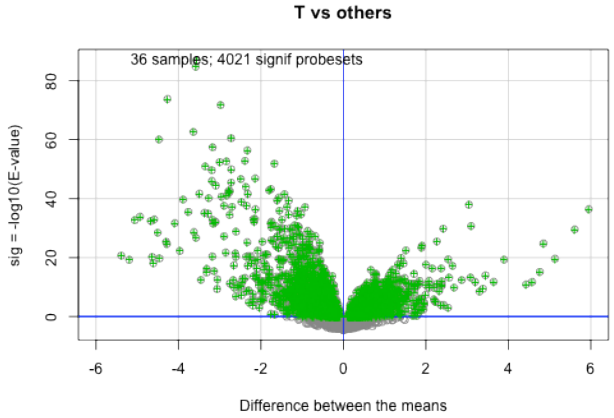
hyperdiploid	44
pre-B ALL	44
TEL-AML1	43
T-ALL	36
E2A-rearranged (EP)	8
BCR-ABL	4
E2A-rearranged (E-sub)	4
MLL	4
BCR-ABL + hyperdiploidy	1
E2A-rearranged (E)	1
TEL-AML1 + hyperdiploidy	1

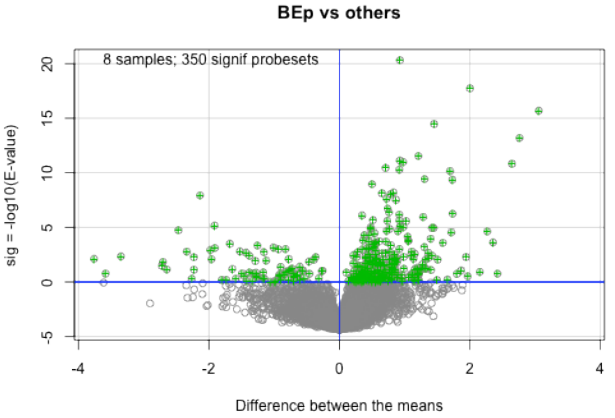
Welch test results for two-groups differential analysis

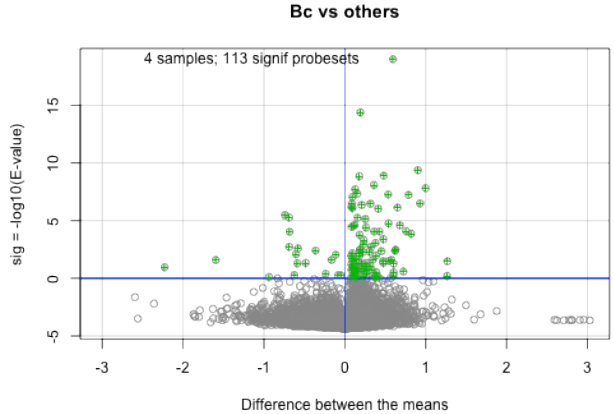


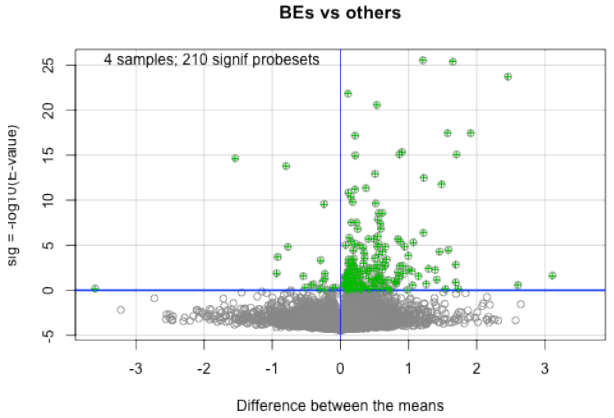


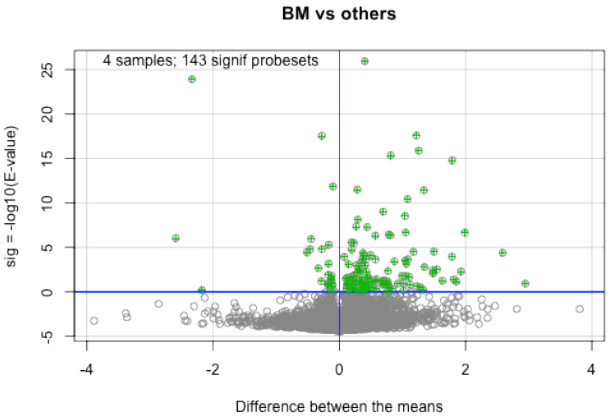








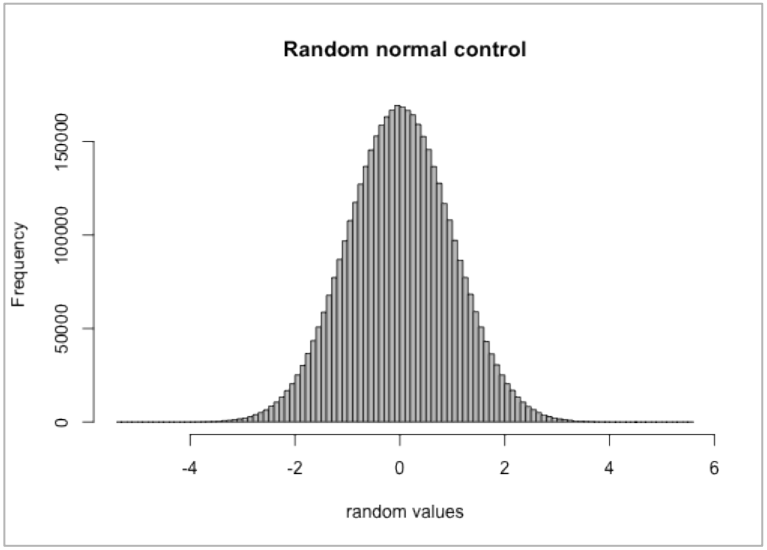


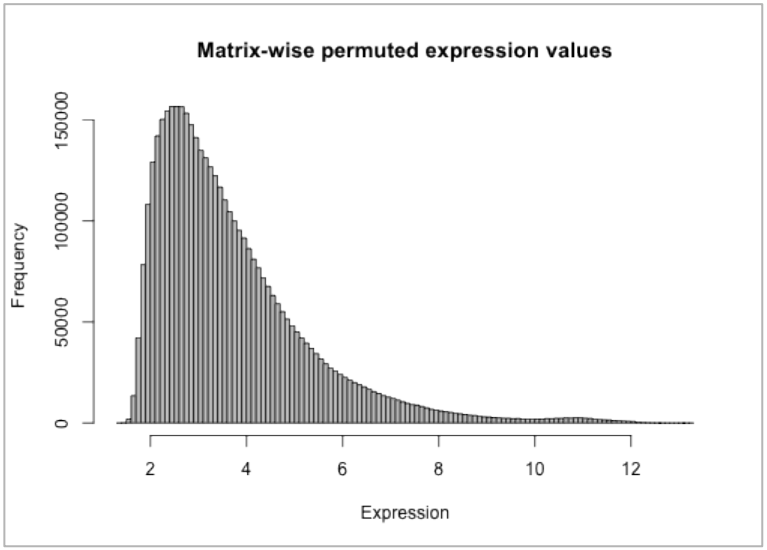


Bh	hyperdiploid	44
Bo	pre-B ALL	44
Bt	TEL-AML1	43
T	T-ALL	36
BEp	E2A-rearranged (EP)	8
Bc	BCR-ABL	4
BEs	E2A-rearranged (E-sub)	4
BM	MLL	4
Bch	BCR-ABL + hyperdiploidy	1
BE	E2A-rearranged (E)	1
Bth	TEL-AML1 + hyperdiploidy	1

Negative controls

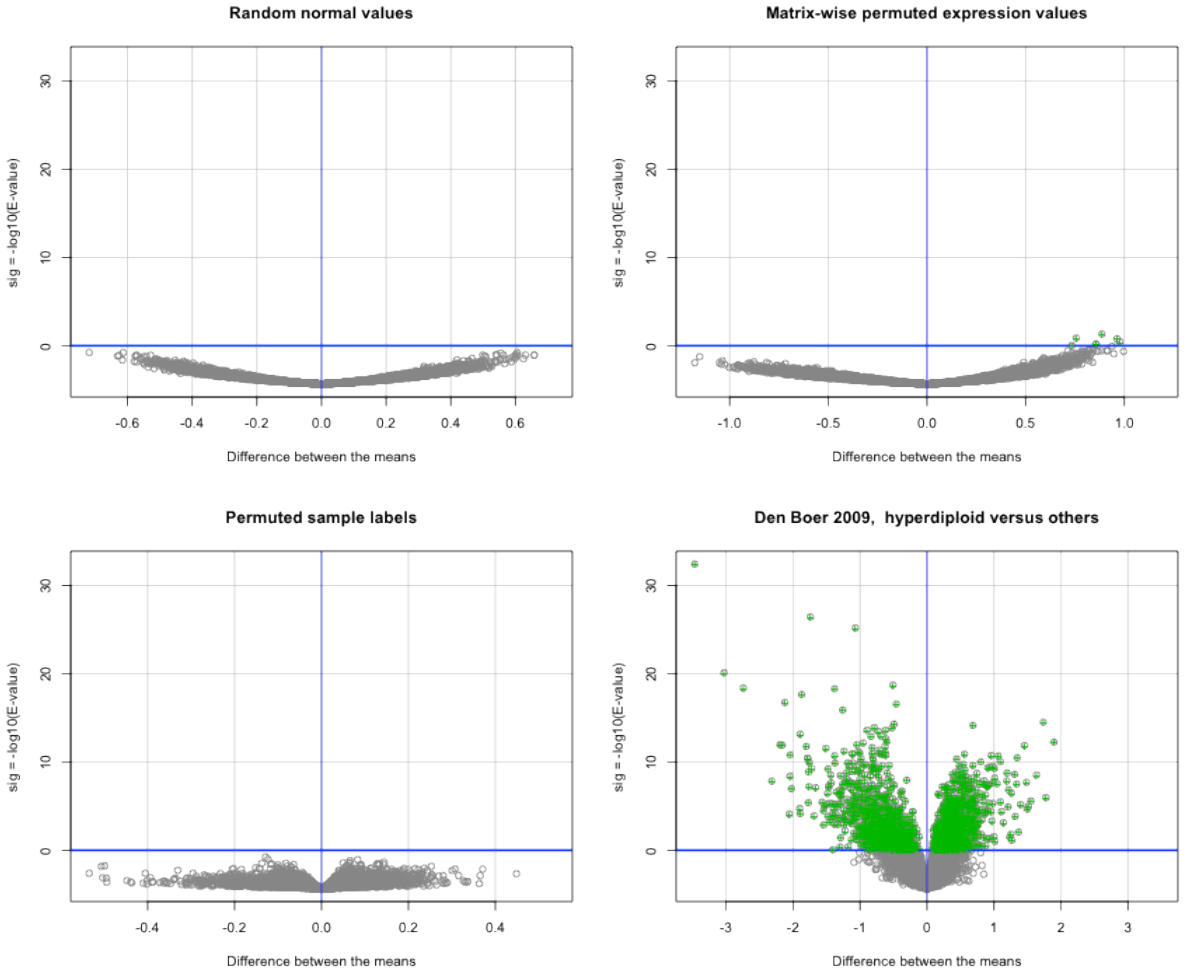
- It is always useful to check empirically the significance of a selection procedure.
- For this, we can build negative controls, i.e. datasets where no difference is expected between groups.
- 3 negative controls
  - **Random normal values**. We build a fake expression matrix by generating random numbers following a normal distribution. This perfectly fits the working hypotheses underlying statistical tests (Student, ANOVA, …) but is not a very realistic image of the biological data.
  - **Matrix-wise random permutation of expression values**. The distribution of values corresponds to the typical Affymetrix expression sets: left-skewed distribution.
  - **Permutation of sample labels**. We maintain the structure of the original expression matrix, but the sample labels are re-assigned at random. In principle, the labels are balanced between all the cancer subtypes, and there should be no significant difference between the randomized groups.





Distribution of P-values from Welch test

- Data set: Den Boer et al. (2009).
- Welch test: hyperdiploid versus other types of Acute Lymphoblastic Leukemia.
- 3 negative controls
  - Random normal values.
    - All significances are negative.
  - Matrix-wise random permutation of expression values.
    - 7 probesets are slightly significant.
  - Permutation of sample labels, analysis of the original expression matrix.
    - All significances are negative.
- Original expression matrix.
  - 2133 probesets are declared significant (differentially expressed) with E-value <= 1.



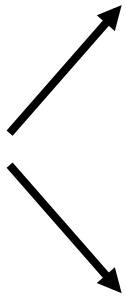
- Data source: Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Measuring differences between expression profiles

From profiles to (dis)similarity matrix

Expression profiles
(gene/sample)

	Sample 1	Sample2	...	Sample j	...	Sample p
Gene 1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
Gene 2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
...	...	...	...	...	...	...
Gene i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{xp}
...	...	...	...	...	...	...
Gene n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}



(Dis)similarity matrix
(gene/gene)

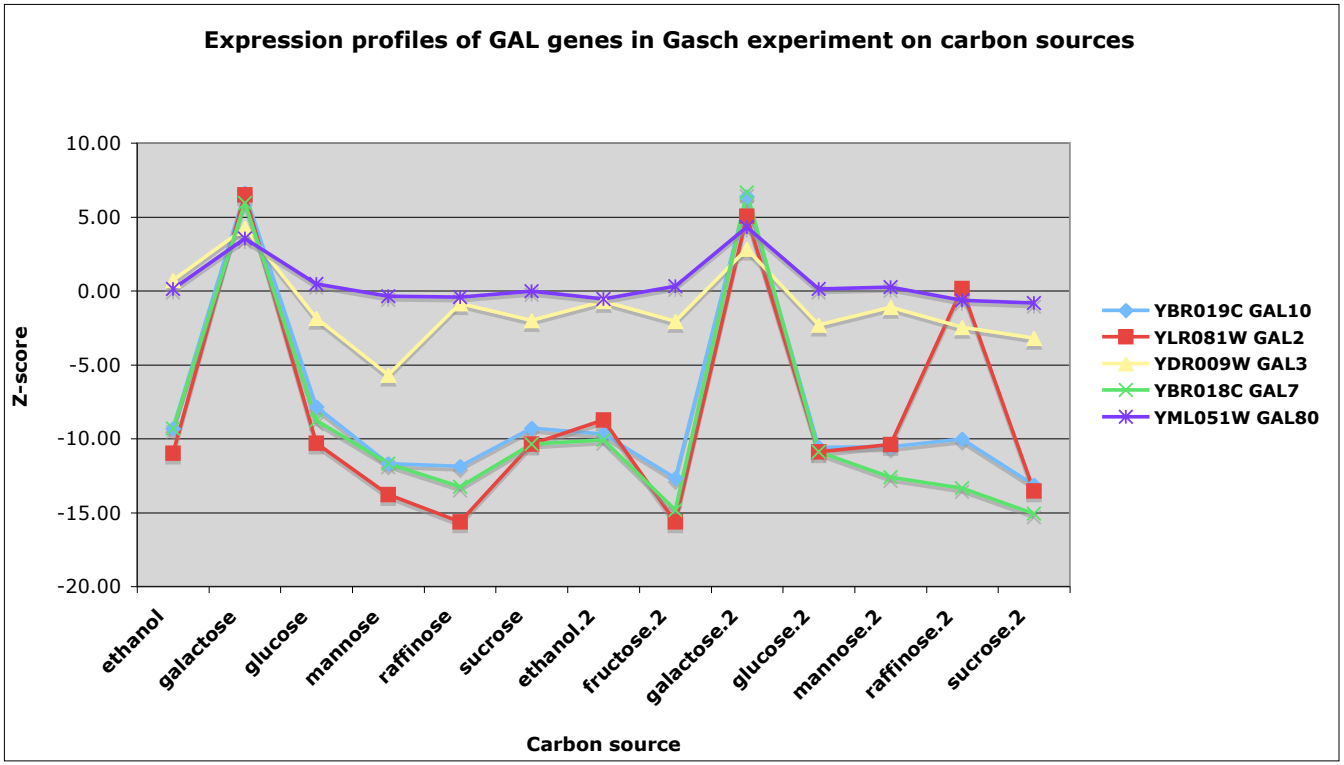
	Gene 1	Gene 2	...	Gene j	...	Gene n
Gene 1	d_{11}	d_{12}	...	d_{1j}	...	d_{1n}
Gene 2	d_{21}	d_{22}	...	d_{2j}	...	d_{2n}
...	...	...	...	...	...	...
Gene i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{dn}
...	...	...	...	...	...	...
Gene n	d_{n1}	d_{n2}	...	d_{nj}	...	d_{nn}

(Dis)similarity matrix
(sample/sample)

	Sample 1	Sample 2	...	Sample j	...	Sample n
Sample 1	d_{11}	d_{12}	...	d_{1j}	...	d_{1p}
Sample 2	d_{21}	d_{22}	...	d_{2j}	...	d_{2p}
...	...	...	...	...	...	...
Sample i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{dp}
...	...	...	...	...	...	...
Sample n	p_{n1}	d_{p2}	...	d_{pj}	...	d_{pp}

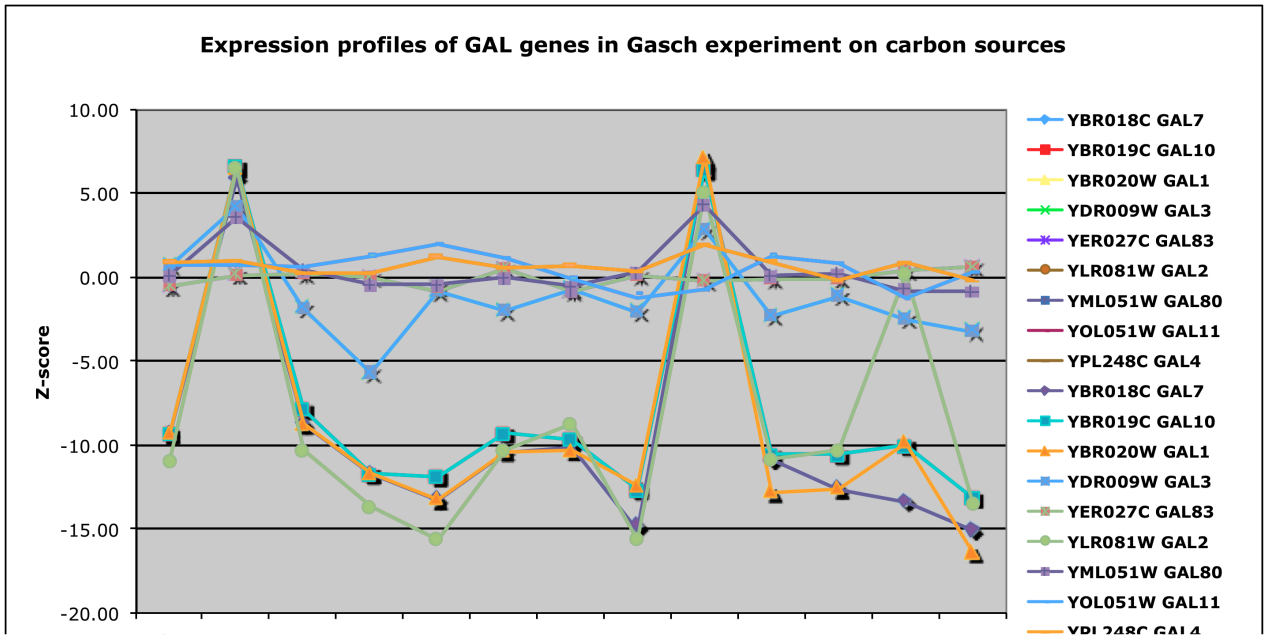
Example: response of GAL genes to carbon sources

- Some GAL genes show obvious similarities in their expression profiles
- For some other ones, it is less obvious.
- How can we quantify this ?



Example: response of GAL genes to carbon sources

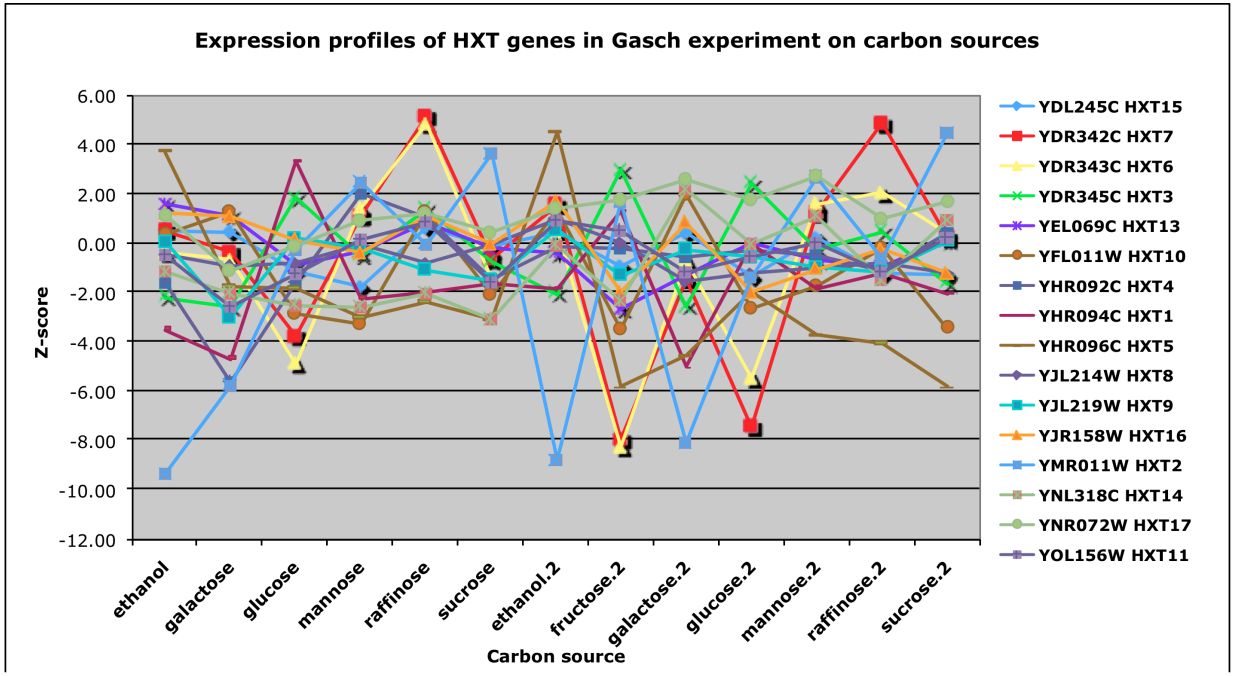
Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2
YBR018C	GAL7	-9.33	5.94	-8.76	-11.71	-13.25	-10.37	-10.08	-14.84	6.63	-10.89	-12.61	-13.37	-15.08
YBR019C	GAL10	-9.33	6.62	-7.86	-11.71	-11.88	-9.30	-9.69	-12.66	6.37	-10.58	-10.57	-10.03	-13.15
YBR020W	GAL1	-9.33	6.50	-8.76	-11.71	-13.25	-10.37	-10.32	-12.44	7.17	-12.82	-12.61	-9.89	-16.40
YDR009W	GAL3	0.70	4.25	-1.85	-5.69	-0.85	-2.01	-0.69	-2.07	2.83	-2.29	-1.10	-2.46	-3.20
YER027C	GAL83	-0.57	0.16	0.22	0.00	-0.59	0.52	-0.84	0.11	-0.22	-0.10	-0.06	0.40	0.64
YLR081W	GAL2	-10.99	6.483	-10.31	-13.78	-15.62	-10.37	-8.75	-15.64	5.042	-10.89	-10.4	0.161	-13.54
YML051W	GAL80	0.117	3.544	0.45	-0.374	-0.426	-0.018	-0.562	0.303	4.312	0.116	0.245	-0.644	-0.82
YOL051W	GAL11	0.699	0.685	0.589	1.287	1.988	1.127	-0.112	-1.019	-0.774	1.253	0.818	-1.228	0.265
YPL248C	GAL4	0.899	0.987	0.217	0.228	1.231	0.573	0.712	0.33	1.946	0.944	-0.061	0.846	-0.159
YBR018C	GAL7	-9.326	5.94	-8.761	-11.71	-13.25	-10.37	-10.08	-14.84	6.634	-10.89	-12.61	-13.37	-15.08
YBR019C	GAL10	-9.326	6.624	-7.861	-11.71	-11.88	-9.295	-9.686	-12.66	6.369	-10.58	-10.57	-10.03	-13.15
YBR020W	GAL1	-9.326	6.503	-8.761	-11.71	-13.25	-10.37	-10.32	-12.44	7.165	-12.82	-12.61	-9.886	-16.4
YDR009W	GAL3	0.699	4.248	-1.845	-5.686	-0.852	-2.014	-0.693	-2.065	2.831	-2.293	-1.104	-2.456	-3.201
YER027C	GAL83	-0.566	0.161	0.217	0	-0.592	0.517	-0.843	0.11	-0.221	-0.096	-0.061	0.403	0.635
YLR081W	GAL2	-10.99	6.483	-10.31	-13.78	-15.62	-10.37	-8.75	-15.64	5.042	-10.89	-10.4	0.161	-13.54
YML051W	GAL80	0.117	3.544	0.45	-0.374	-0.426	-0.018	-0.562	0.303	4.312	0.116	0.245	-0.644	-0.82
YOL051W	GAL11	0.699	0.685	0.589	1.287	1.988	1.127	-0.112	-1.019	-0.774	1.253	0.818	-1.228	0.265
YPL248C	GAL4	0.899	0.987	0.217	0.228	1.231	0.573	0.712	0.33	1.946	0.944	-0.061	0.846	-0.159



- Some GAL genes show obvious similarities in their expression profiles
- For some other ones, it is less obvious.
- How can we quantify this ?

Example: response of HXT genes to carbon sources

Gene ID	Gene Name	ethanol	galactose	glucose	mannose	raffinose	sucrose	ethanol.2	fructose.2	galactose.2	glucose.2	mannose.2	raffinose.2	sucrose.2
YDL245C	HXT15	0.53	0.48	-1.12	-1.74	0.66	-0.13	0.43	-0.91	0.44	-1.35	0.25	-1.25	-1.27
YDR342C	HXT7	0.55	-0.32	-3.74	1.12	5.18	-0.61	1.59	-7.93	-1.42	-7.42	1.25	4.87	0.24
YDR343C	HXT6	-0.35	-0.64	-4.85	1.43	4.88	-1.46	0.54	-8.23	-0.69	-5.45	1.64	2.05	0.40
YDR345C	HXT3	-2.13	-2.56	1.86	-0.48	1.42	-0.72	-1.99	3.00	-2.48	2.45	-0.18	0.50	-1.59
YEL069C	HXT13	1.62	1.15	-0.92	-0.35	0.85	-0.22	-0.38	-2.64	-1.28	NA	-0.65	-0.95	0.13
YFL011W	HXT10	0.383	1.309	-2.791	-3.238	1.302	-2.07	1.368	-3.469	1.99	-2.64	-1.676	-0.201	-3.386
YHR092C	HXT4	-1.599	-5.617	-1.566	2.034	1.041	-1.478	-0.094	-0.193	-0.531	-0.251	-0.491	-1.107	0.45
YHR094C	HXT1	-3.464	-4.651	3.349	-2.2	-1.941	-1.589	-1.817	1.321	-4.976	NA	-1.778	-1.208	-2.037
YHR096C	HXT5	3.731	-1.711	-1.799	-2.968	-2.343	-3.068	4.534	-5.781	-4.556	-1.946	-3.72	-4.027	-5.766
YJL214W	HXT8	-0.5	-0.946	-0.853	-0.062	-0.852	-0.074	0.993	0	-1.526	-1.253	-1.022	-0.846	-1.27
YJL219W	HXT9	0.067	-2.96	0.202	-0.208	-1.065	-1.497	0.581	-1.294	-0.221	-0.54	-0.981	-1.128	0.106
YJR158W	HXT16	1.249	1.087	0.14	-0.374	1.136	0	1.742	-1.955	0.862	-1.946	-1.022	-0.242	-1.217
YMR011W	HXT2	-9.326	-5.738	-0.248	2.511	0	3.64	-8.75	1.679	-8.05	-1.349	2.698	-0.604	4.497
YNL318C	HXT14	-1.166	-2.013	-2.496	-2.594	-2.012	-3.068	-0.056	-2.34	2.211	-0.019	1.124	-1.49	0.926
YNR072W	HXT17	1.132	-1.107	-0.109	0.913	1.183	0.425	1.424	1.789	2.632	1.773	2.739	1.027	1.693
YOL156W	HXT11	-0.466	-2.557	-1.209	0.145	0.876	-1.571	0.937	0.523	-1.128	-0.54	0.02	-1.128	0.238



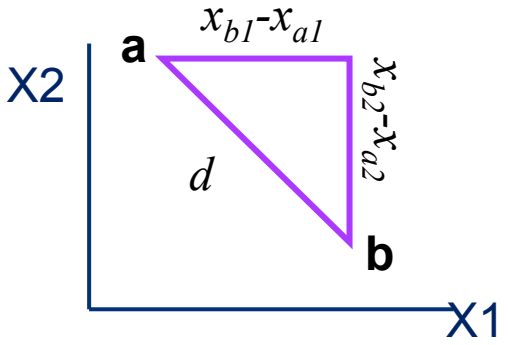
- The HXT genes are annotated as « hexose transporters », « glucose transporters », « putative hexose transporters », ...
- Some of them are strongly activated or repressed by particular carbon sources.
- Their profiles differ from each other, suggesting substrate specificity.
- Can we learn something about the function of these genes by analyzing their expression profiles?

Euclidian distance

- You are probably familiar with the computation of Euclidian distance in a 2-dimensional space.
- The concept naturally extends to spaces with higher dimension (p-dimensional space).
- Two typical applications
  - The distance between genes is calculated in the space of samples:
    objects = genes (or probes), variables = samples (or chips)
  - The distance between samples is calculated in the space of genes:
    objects = samples (or chips), variables = genes (or probes)
- Notations
  - X_{aj} , X_{bj} values taken by the j^{th} variable for the objects *a* and *b*, respectively.
  - p number of dimensions.

Euclidian distance in a 2D space

$$D_{ab} = \sqrt{\left(x_{a1} - x_{b1}\right)^2 + \left(x_{a2} - x_{b2}\right)^2}$$



Euclidian distance
in a p-dimensional space

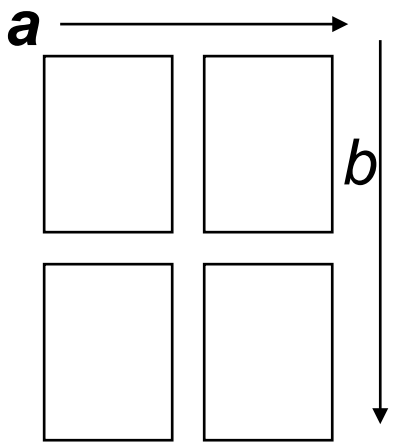
$$D_{ab} = \sqrt{\sum_{j=1}^{p}\left(x_{aj} - x_{bj}\right)^2}$$

Mean Euclidian distance

$$D_{ab} = \frac{1}{p}\sqrt{\sum_{i=1}^{p}\left(x_{ai} - x_{bi}\right)^2}$$

Manhattan distance

- The Manhattan distance between two points a and b is the weighted sum of the absolute differences in each dimension.
  - a,b two points in the multi-variate space
  - p number of dimensions
  - w_i weight if the i^{th} dimension



$$D_{ab} = \sum_{j=1}^{p} w_j \left| x_{aj} - x_{bj} \right|$$

Correlation-related metrics

- A detailed description of those metrics has been given in the chapter "**Correlation analysis**".
- The coefficient of correlation and several related metrics can be converted to dissimilarity metrics.
  - *mdp* mean dot product
  - *cor* correlation
  - *Ucor* uncentered correlation

$$mdp_{ab} = \frac{1}{p}\mathbf{x_a} \cdot \mathbf{x_b} = \frac{1}{p}\sum_{i=1}^{p}\left(x_{ai} \cdot x_{bi}\right)$$

$$mdpd_{ab} = k - dmp_{ab}$$

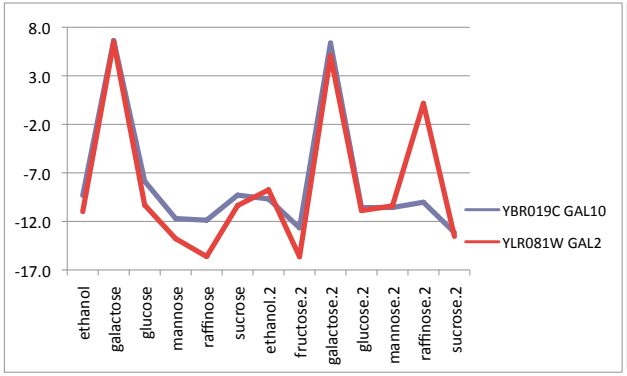
$$cor_{ab} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{x_{ai} - \hat{m}_a}{\hat{\sigma}_a}\right)\left(\frac{x_{bi} - \hat{m}_b}{\hat{\sigma}_b}\right)$$

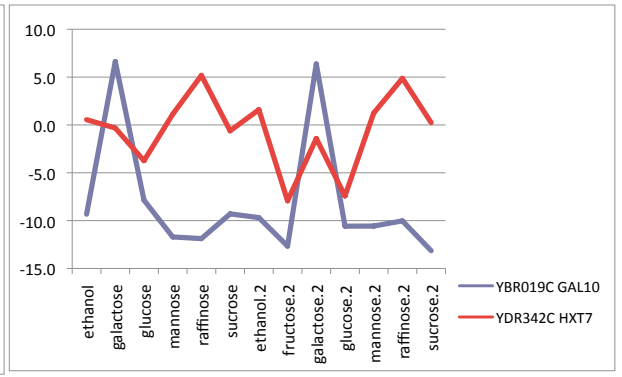
$$Dcor_{ab} = 1 - cor_{ab}$$

$$Ucor_{ab} = \frac{\sum_{i=1}^{p}\left(x_{ai}x_{bi}\right)}{\sqrt{\sum_{j=1}^{p}x_{aj}^{2}}\sqrt{\sum_{j=1}^{p}x_{bj}^{2}}}$$

Examples of comparisons between expression profiles

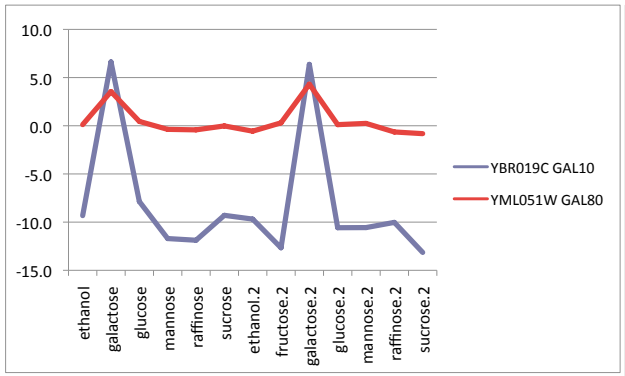
GAL10 - GAL2
Eucl dist 12.0
Dot product 1386.2
Cor 0.8

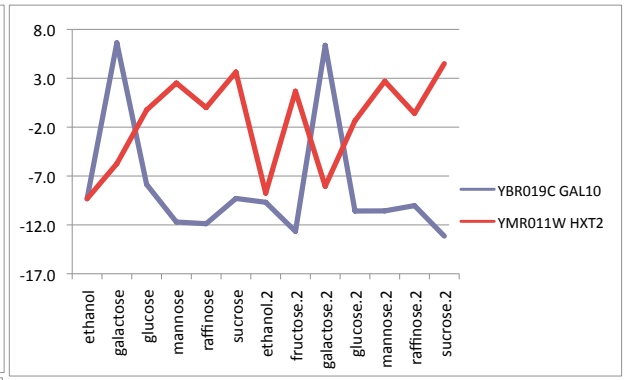




GAL10 - HXT7
Eucl dist 38.1
Dot product 42.4
Cor 0.0

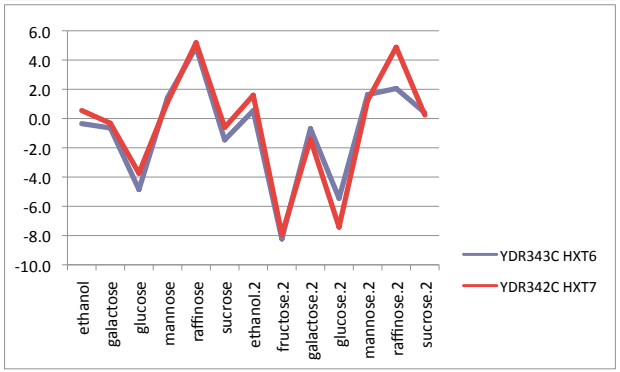
GAL10 - GAL80
Eucl dist 35.2
Dot product 70.9
Cor 0.9





GAL10 - HXT2
Eucl dist 42.4
Dot product -67.4
Cor -0.5

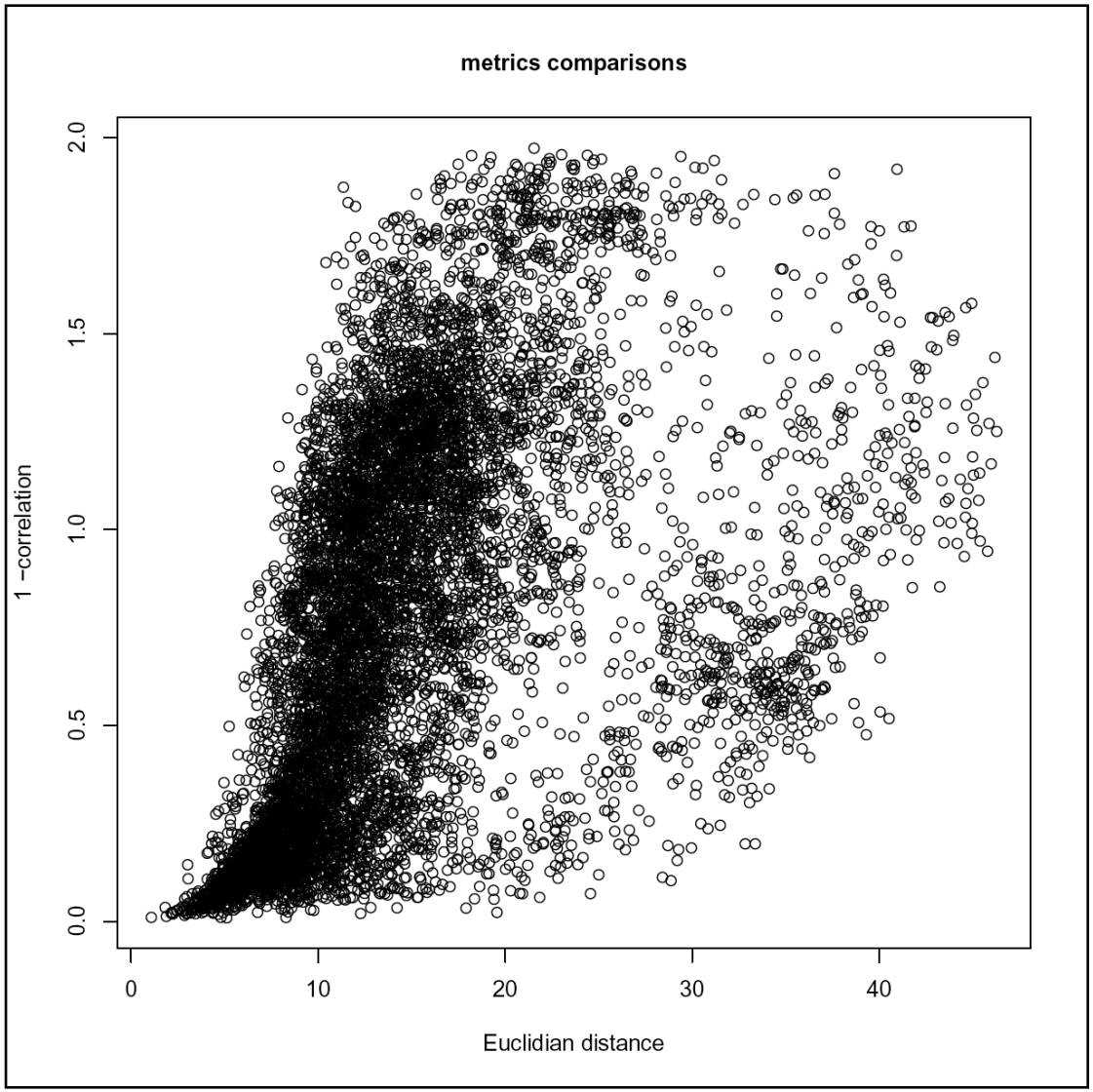
HXT7 - HXT6
Eucl dist 4.1
Dot product 165.6
Cor 0.9



Choice of a metric for the clustering of gene expression data

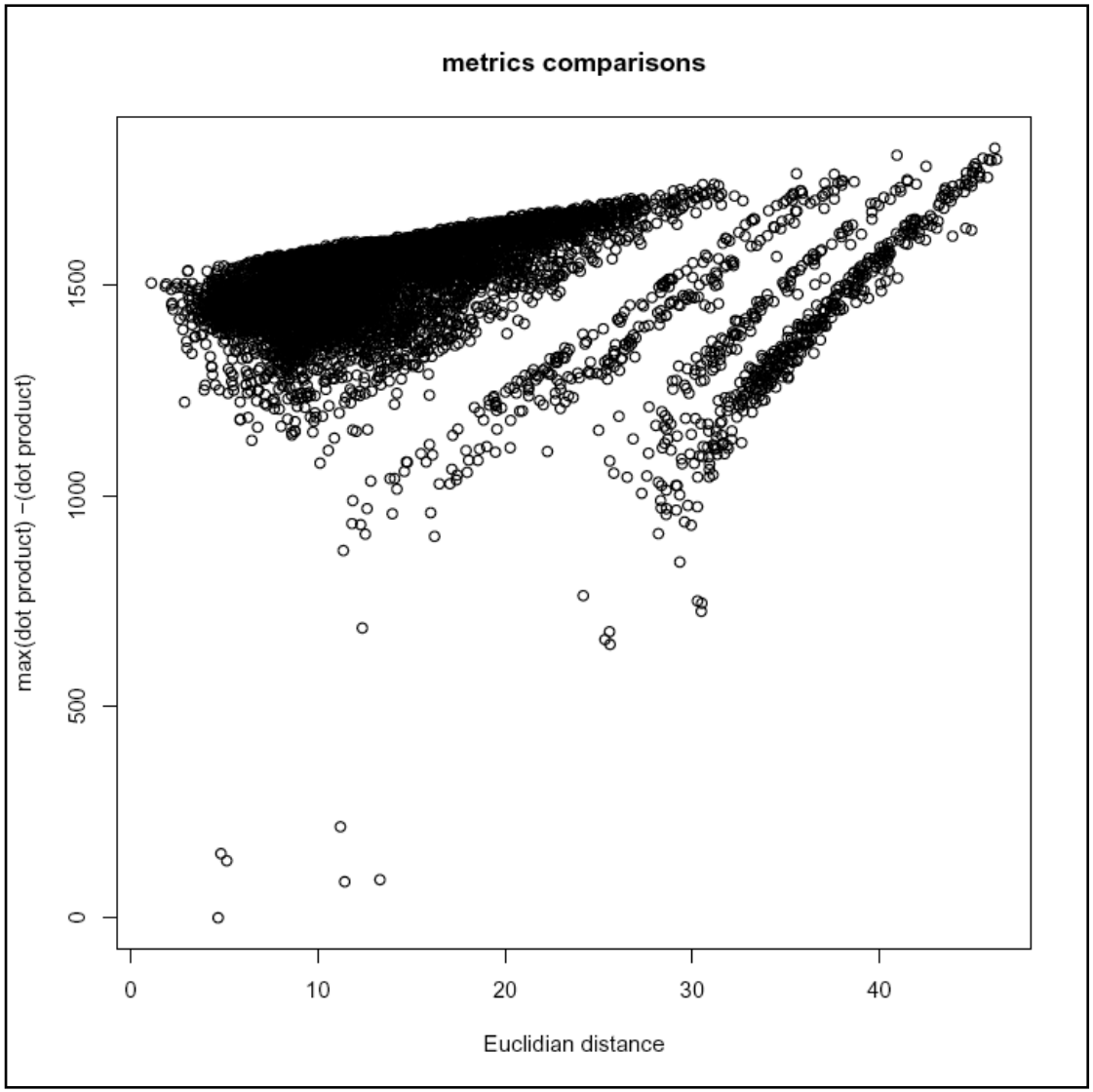
- Euclidian distance
  - takes into account the absolute level of regulation (provided the genes have not been standardized)
  - Does not distinguish anti-correlation from absence of correlation
- Pearson's correlation
  - Indicates anti-correlation as well as correlation.
  - Does not indicate the absolute level of regulation.
  - Problem : assumes that the reference for each gene is the mean of its profile -> implicitly, one consider that each gene is on the average not regulated in the data set.
- Uncentered correlation
  - Indicates anti-correlation as well as correlation.
  - Does not indicate the absolute level of regulation.
  - Assumes that the reference level is 0, i.e. the level of the control experiment (the contribution of the green measurement to the log ratio)
- Dot product
  - In principle, the dot product combines advantages of the Euclidian distance and of the coefficient of correlation
  - It takes positive values to represent co-regulation, and negative values to represent anti-regulation (as the coefficient of correlation)
  - It reflects the strength of the regulation of both genes, since it uses the real values (as the Euclidian distance) rather than the standardized ones (as the coefficient of correlation).

Carbon sources – Euclidian versus correlation



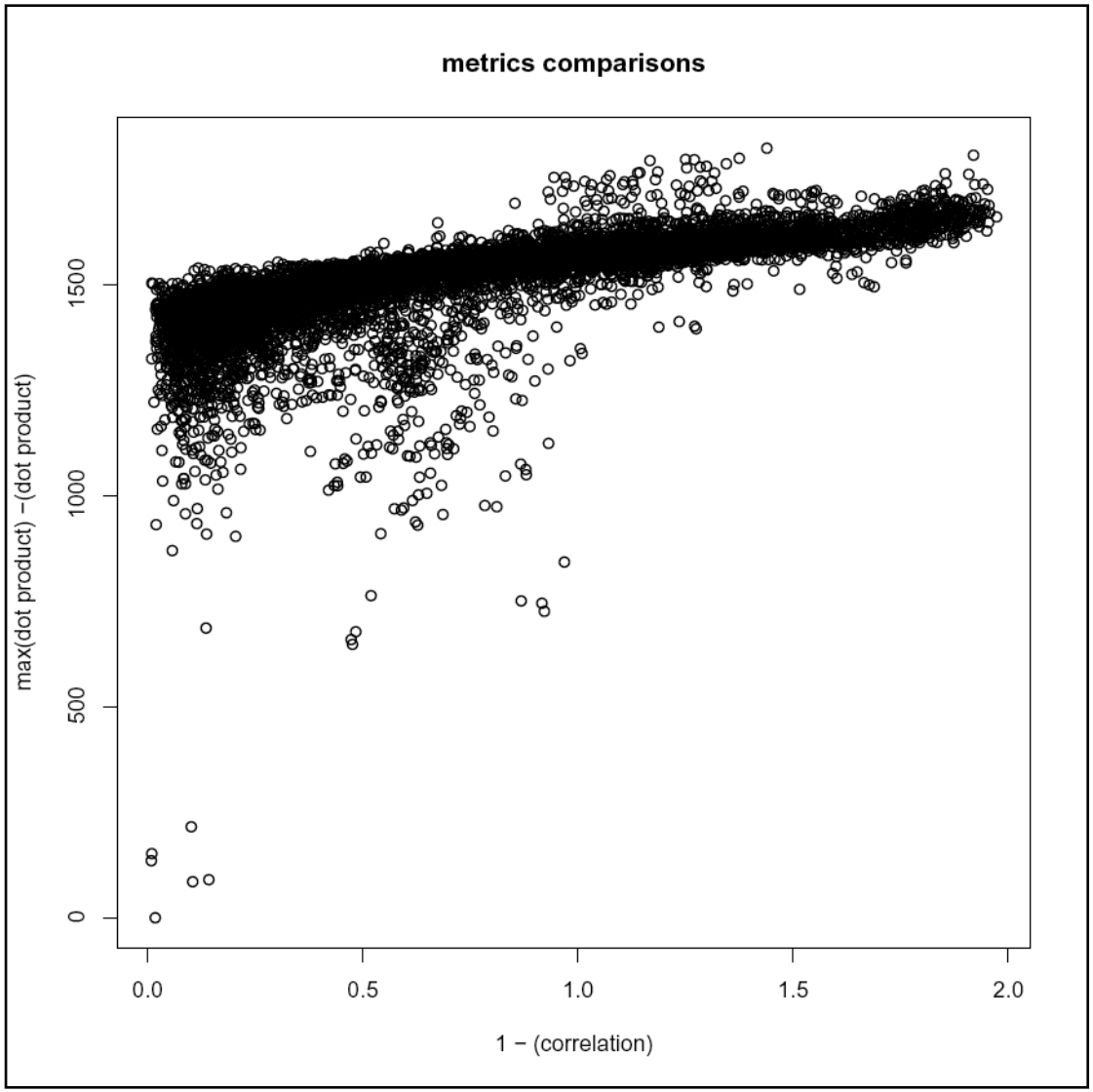
- On this figure, each dot represents a pair of genes from the carbon source experiment.
- We selected the 133 genes showing a significant response in at least one of the 13 chips.
- For each pair of genes, we calculated the **Euclidian distance** (X axis) and Pearson's **centred coefficient of correlation** (Y axis).
- The plot shows that the two metrics are related but distinct.
- The cloud of points seems to be inhomogeneous: there are at least two separate trends.

Carbon sources – Euclidiant versys dot product



- **Euclidian distance** and **dot product** are clearly distinct.
- Some gene pairs will be considered as very similar or very dissimilar depending on the chosen metrics.
- There are however several apparent groups of gene pairs showing linear relationships between Euclidiant distance and dot product.
- Why ? (not obvious for me, I would need to dig out)

Carbon sources – correlation versus dot product



- **Coefficient of correlation (centred)** and **dot product.**
- Correlation and dot product show similar trends (not surprizing, the formulae are related).
- The rankings will nevertheless be different between these two metrics.

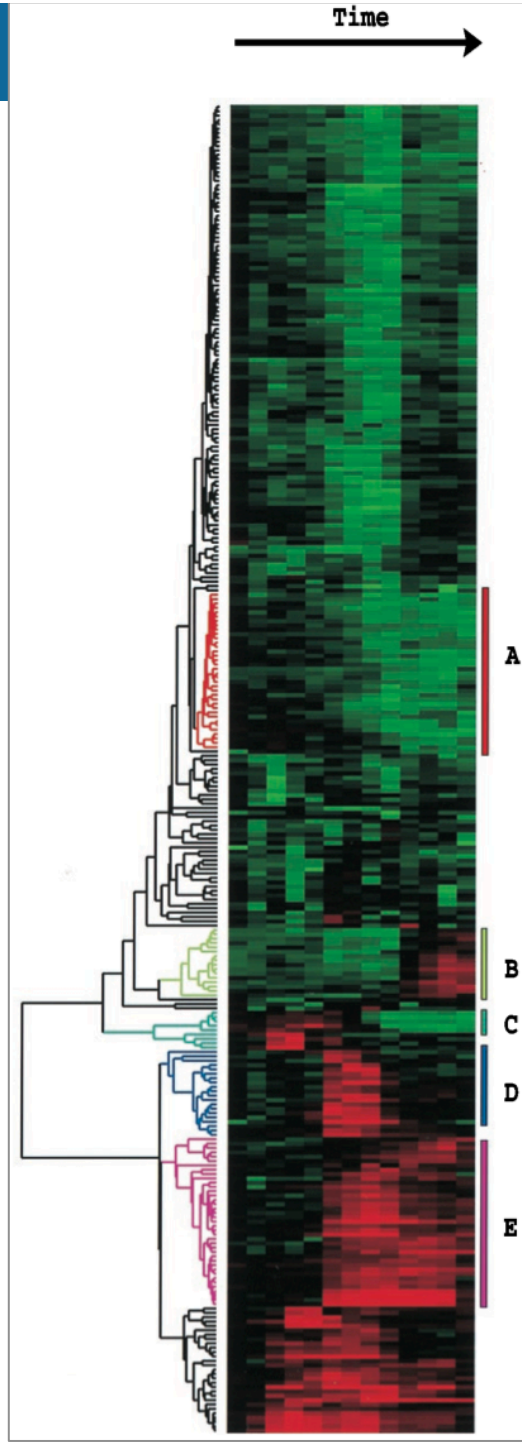
Clustering

Introduction to clustering

- Clustering is an *unsupervised* approach
  - Class discovery: starting from a set of objects, group them into classes, without any prior knowledge of these classes.
- There are many clustering methods
  - hierarchical
  - k-means
  - self-organizing maps (SOM)
  - knn
  - ...
- The results vary drastically depending on
  - clustering method
  - similarity or dissimilarity metric
  - additional parameters specific to each clustering method (e.g. number of centres for the k-mean, agglomeration rule for hierarchical clustering, ...)

Hierarchical clustering of expression profiles

- In 1998, Eisen et al.
  - Implemented a software tool called *Cluster*, which combine hierarchical clustering and heatmap visualization.
  - Applied it to extract clusters of co-expressed genes from various types of expression profiles.



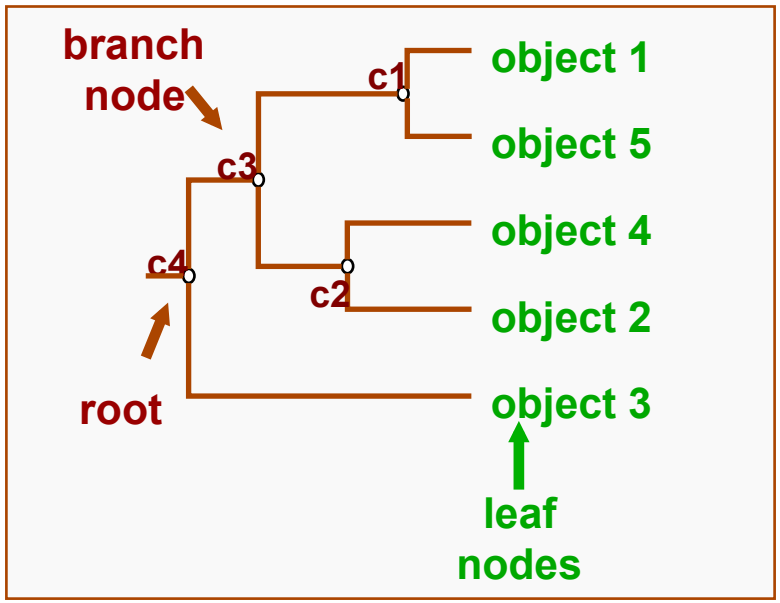
- Eisen et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A (1998) vol. 95 (25) pp. 14863-8

Principle of tree building

Distance matrix

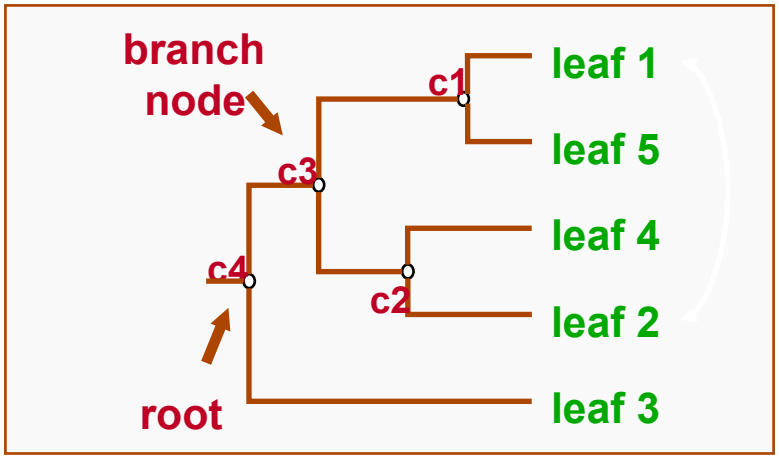
	object 1	object 2	object 3	object 4	object 5
object 1	0.00	4.00	6.00	3.50	1.00
object 2	4.00	0.00	6.00	2.00	4.50
object 3	6.00	6.00	0.00	5.50	6.50
object 4	3.50	2.00	5.50	0.00	4.00
object 5	1.00	4.50	6.50	4.00	0.00

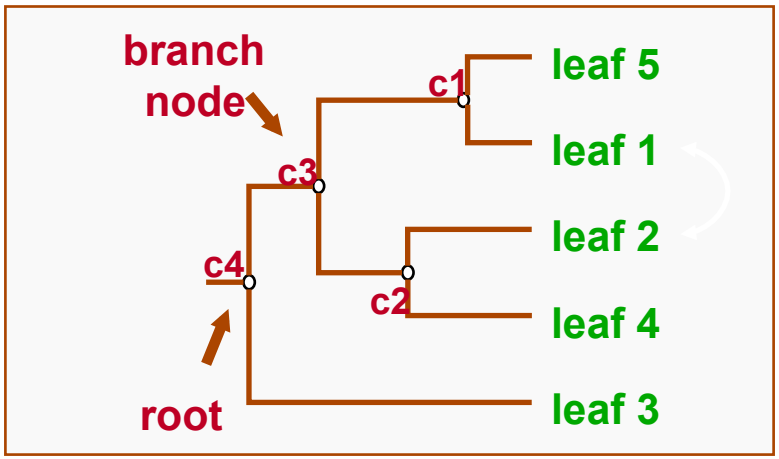
Tree representation



- Hierarchical clustering is an aggregative clustering method
  - takes as input a distance matrix
  - progressively regroups the closest objects/ groups
- One needs to define a (dis)similarity metric between two groups. There are several possibilities
  - **Average linkage**: the average distance between objects from groups A and B
  - **Single linkage**: the distance between the closest objects from groups A and B
  - **Complete linkage**: the distance between the most distant objects from groups A and B
- Algorithm
  - (1) Assign each object to a separate cluster.
  - (2) Find the pair of clusters with the shortest distance, and regroup them in a single cluster
  - (3) Repeat (2) until there is a single cluster
- The result is a tree, whose intermediate nodes represent clusters
  - N objects → N-1 intermediate nodes
- Branch lengths represent distances between clusters

Isomorphism on a tree

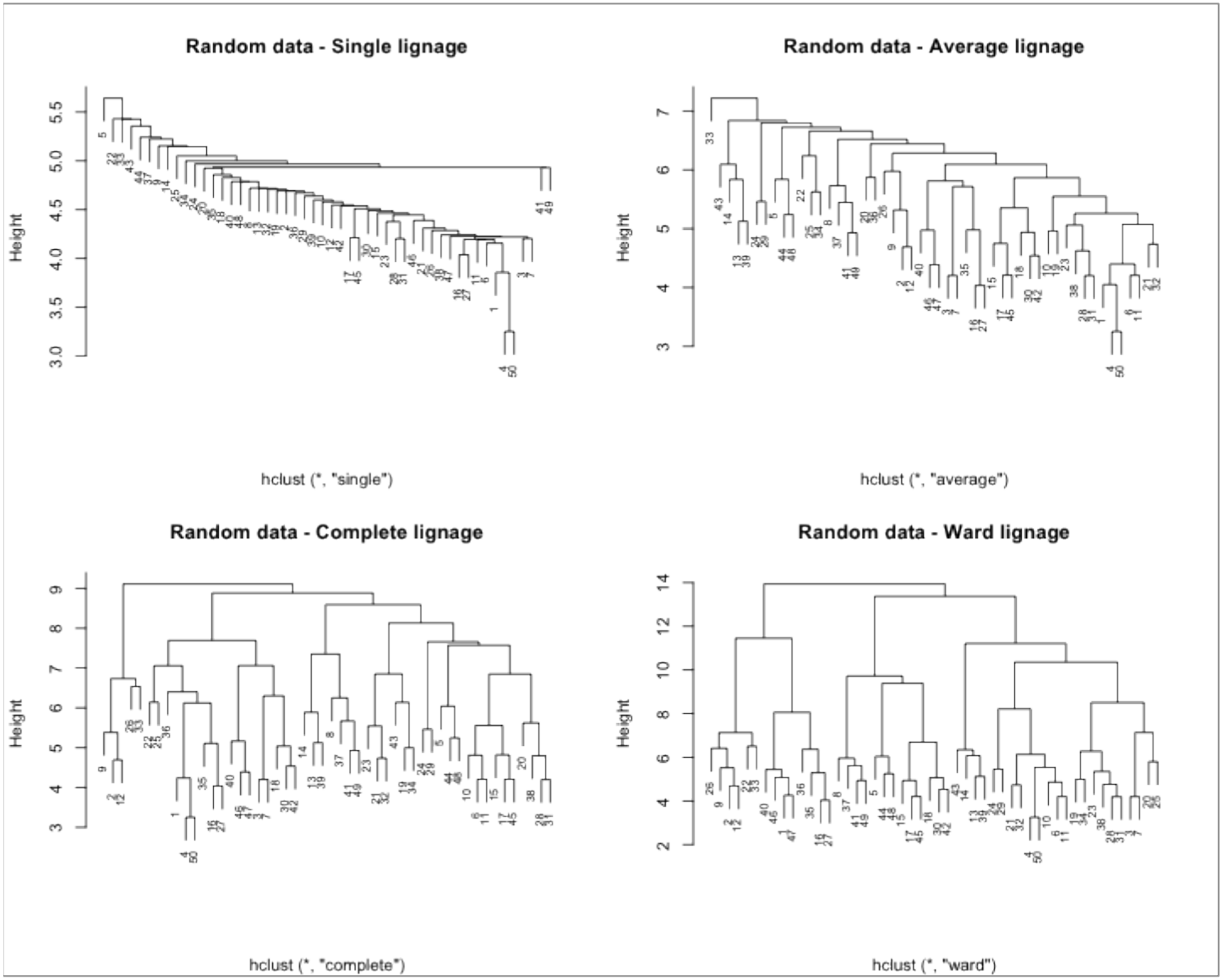




- In a tree, the two children of any branch node can be swapped. The result is an ***isomorphic tree***, considered as equivalent to the intial one.
- The two trees shown here are equivalent, however
  - Top tree: leaf 1 is far away from leaf 2
  - Bottom tree: leaf 1 is neighbour from leaf 2
- The vertical distance between two nodes does NOT reflect their actual distance !
- The distance between two nodes is the ***sum of branch lengths***.

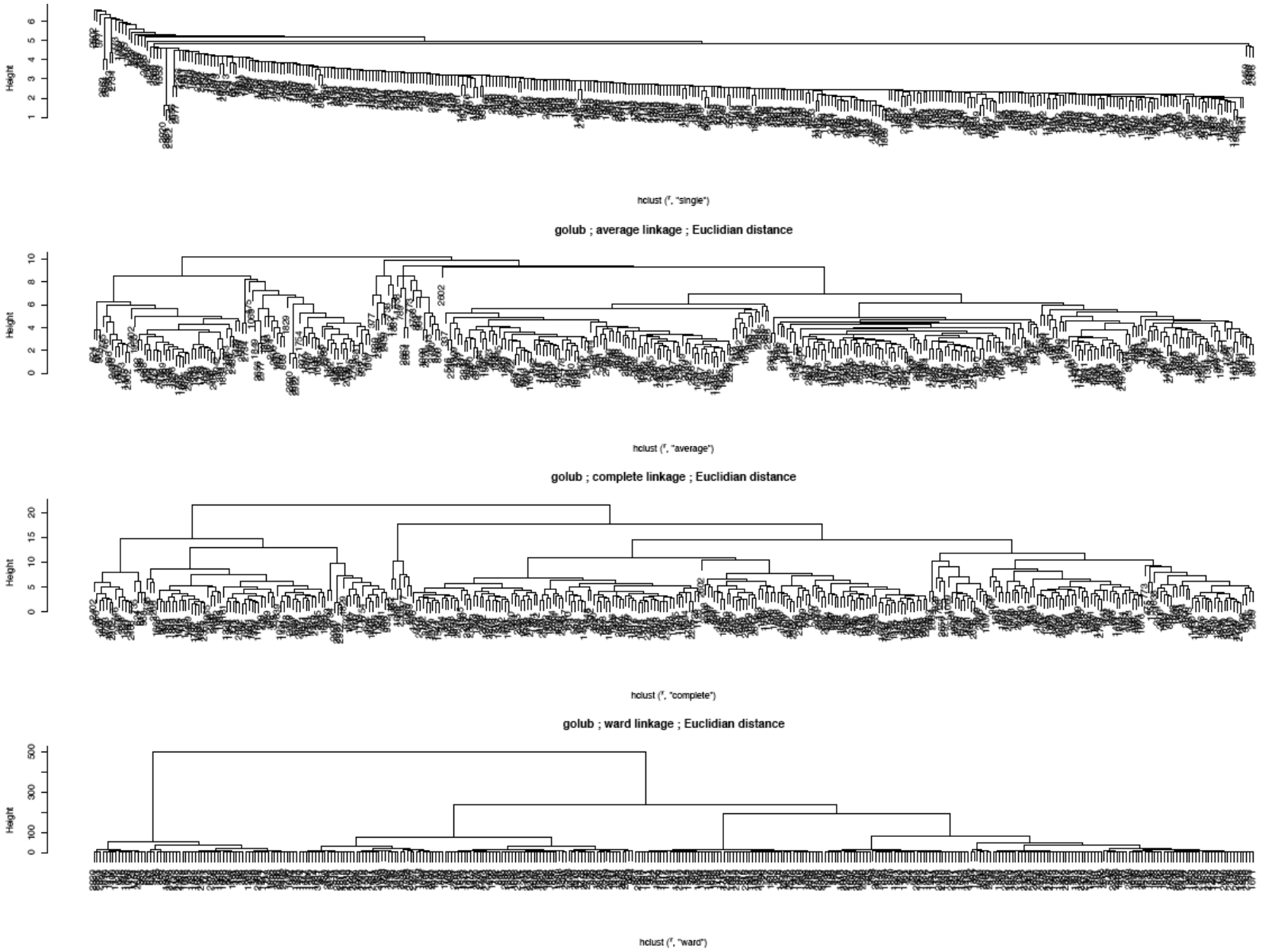
Impact of the agglomeration rule

- The choice of the agglomeration rule has a strong impact on the structure of a tree resulting from hierarchical clustering.

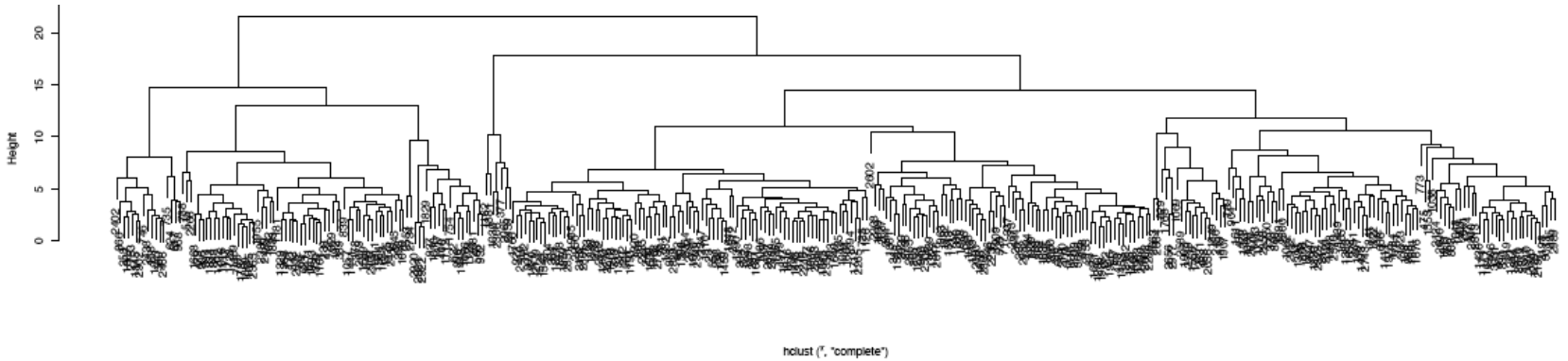


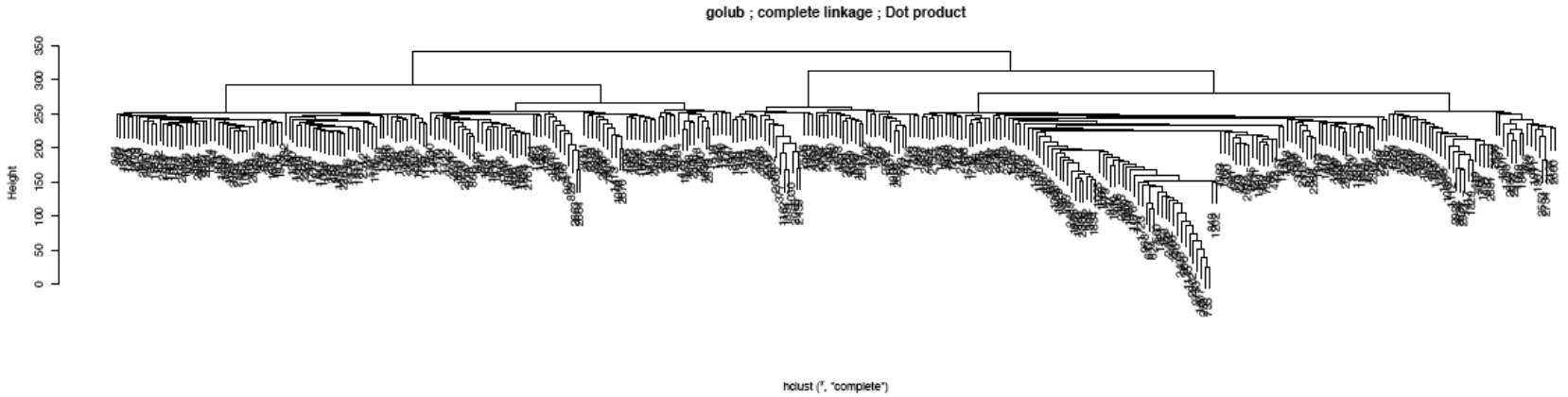
- Those four trees were built from the same distance matrix, using 4 different agglomeration rules.
- The clustering order is completely different.
- Single-linkage typically creates nesting clusters (“Matryoshka dolls”).
- Complete and Ward linkage create more balanced trees.
- Note: the matrix was computed from a matrix of random numbers. The subjective impression of structure are thus complete artifacts.

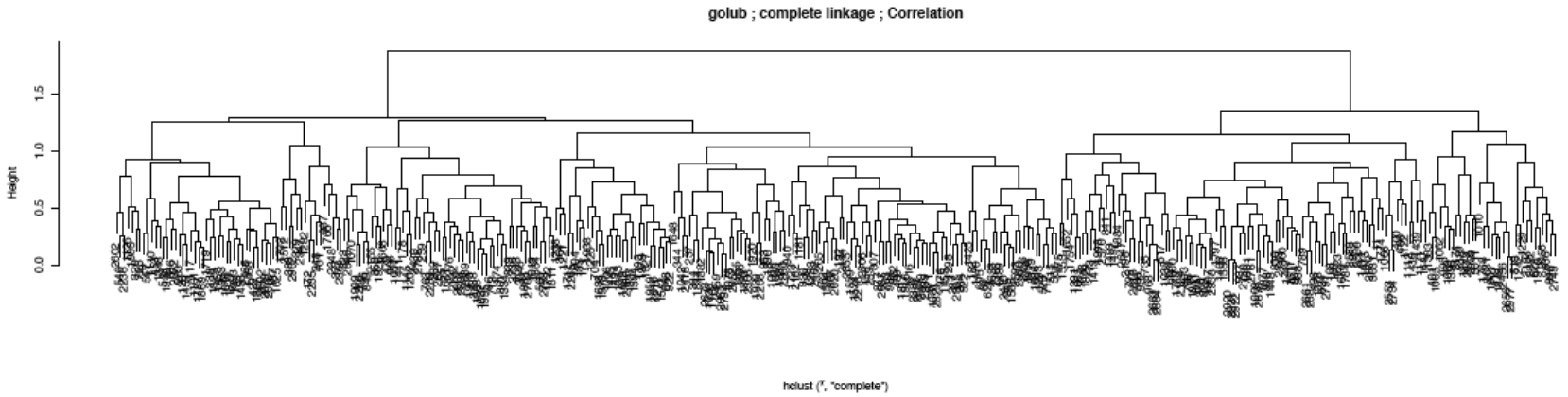
Golub 1999 - Impact of the linkage method (Euclidian distance for all the trees)



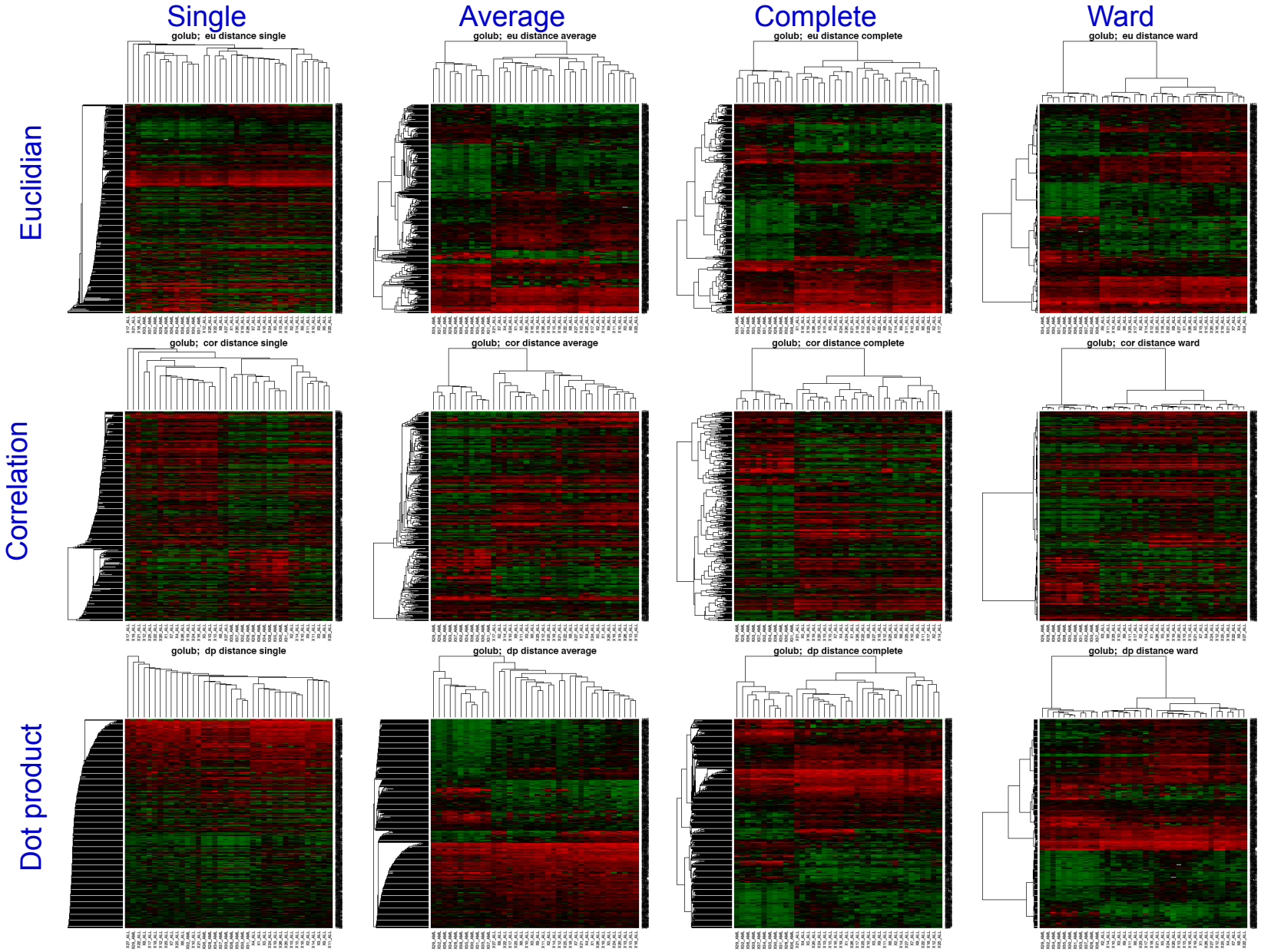
Golub 1999 - Effect of the distance metrics (complete linkage for all the trees)



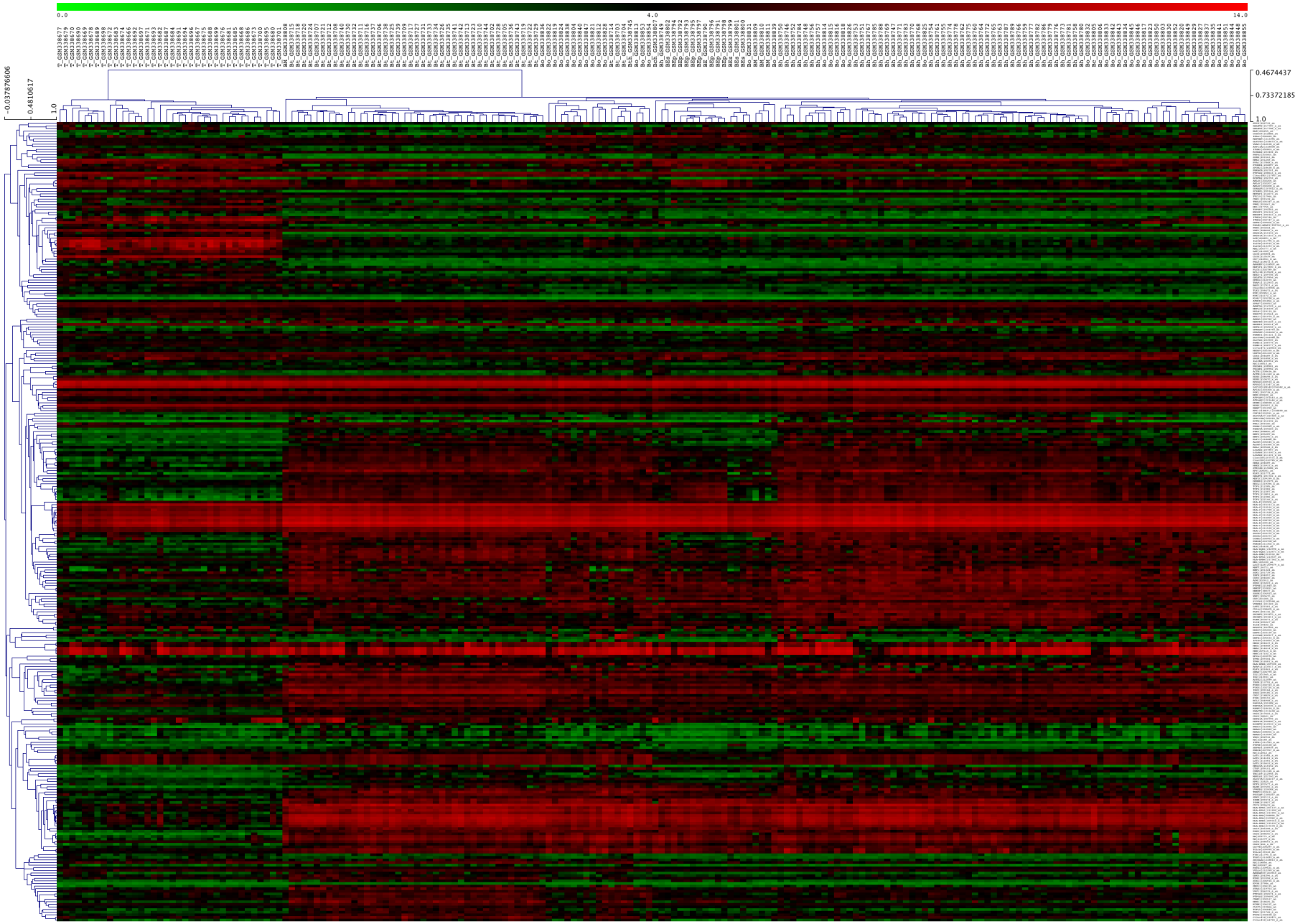




Impact of distance metrics and agglomeration rules



Den Boer 2009 – Hierarchical clustering



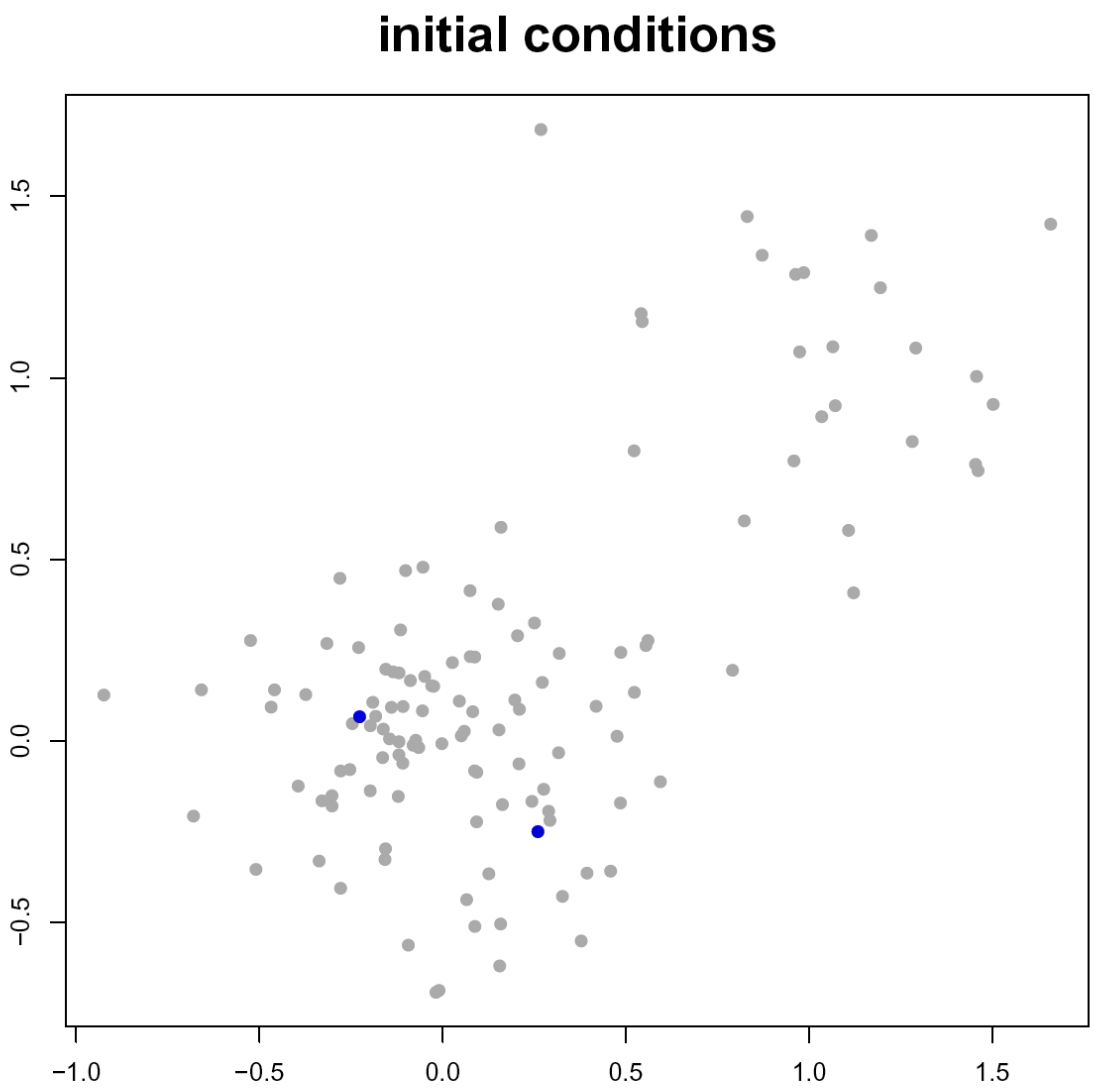
K-means clustering

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Clustering around mobile centres

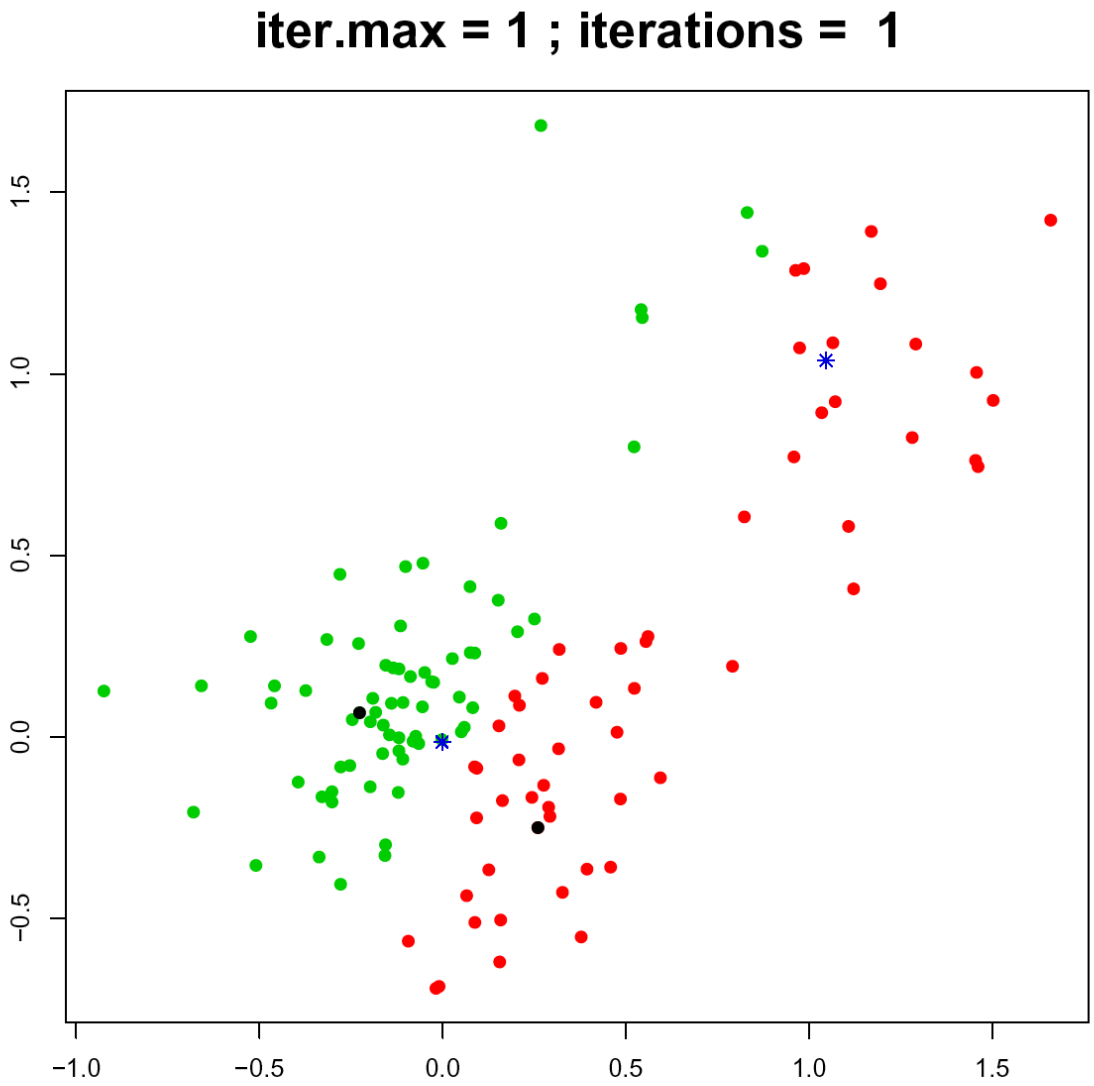
- The number of centres (k) has to be specified a priori
- Algorithm
  - (1) Arbitrarily select k initial centres
  - (2) Assign each element to the closest centre
  - (3) Re-calculate centres (mean position of the assigned elements)
  - (4) Repeat (2) and (3) until one of the stopping conditions is reached
    - the clusters are the same as in the previous iteration
    - the difference between two iterations is smaller than a specified threshold
    - the max number of iterations has been reached

Mobile centres example - initial conditions



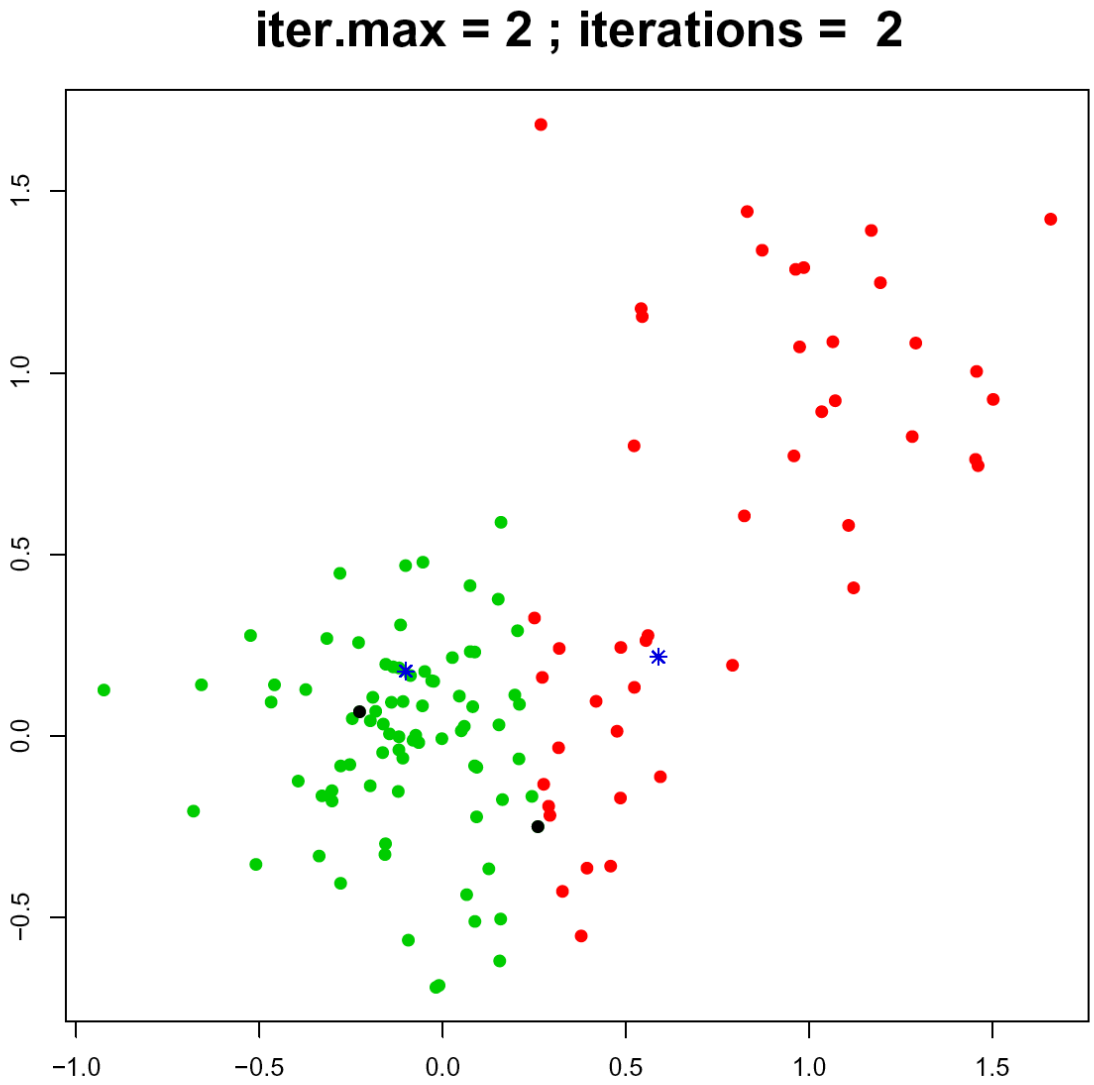
- Two sets of random points are randomly generated
  - 200 points centred on (0,0)
  - 50 points centred on (1,1)
- Two points are randomly chosen as seeds (blue dots)

Mobile centres example - first iteration



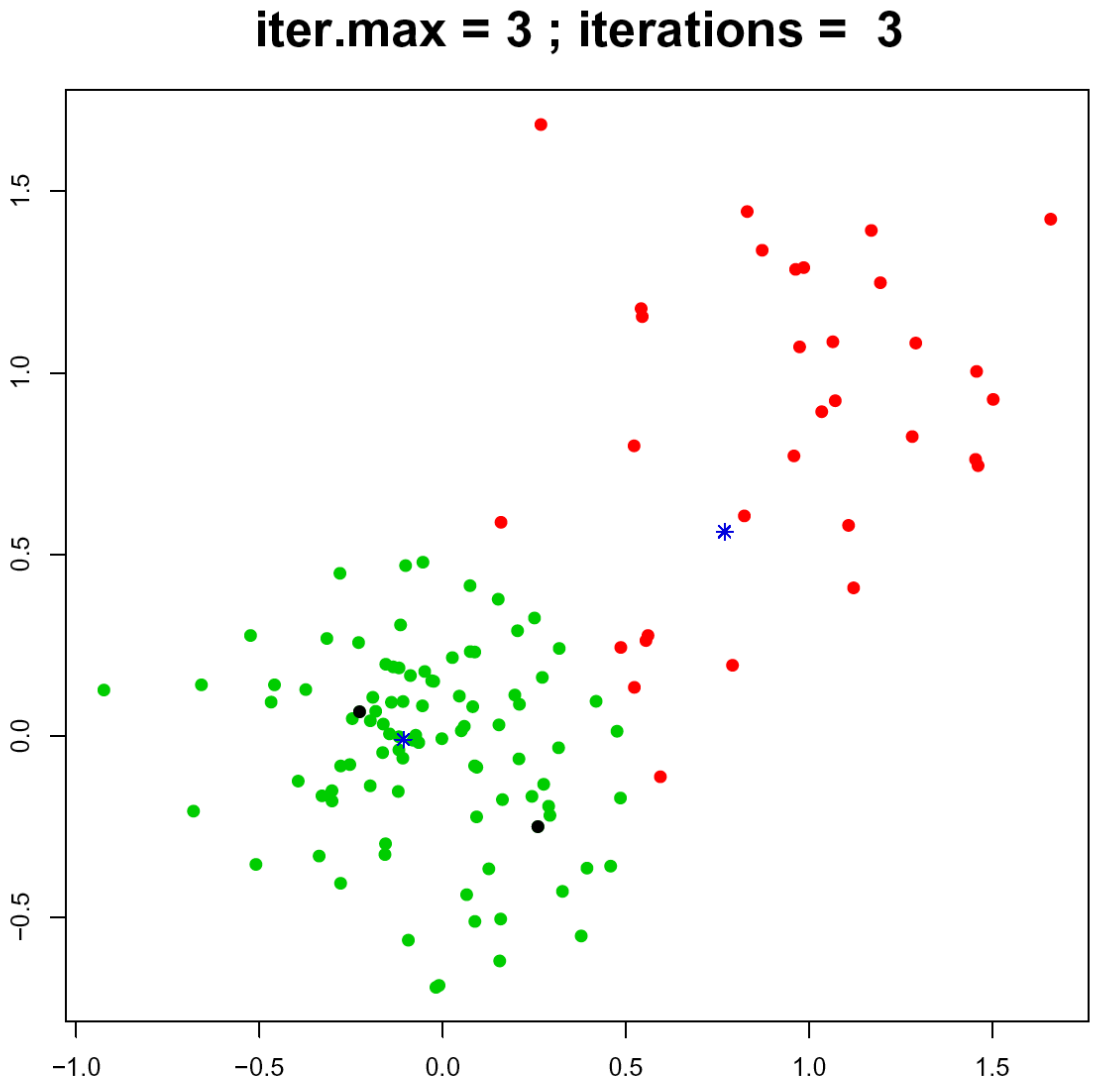
- Step 1
  - Each dot is assigned to the cluster with the closest centre
  - Centres are re-calculated (blue star) on the basis of the new clusters

Mobile centres example - second iteration



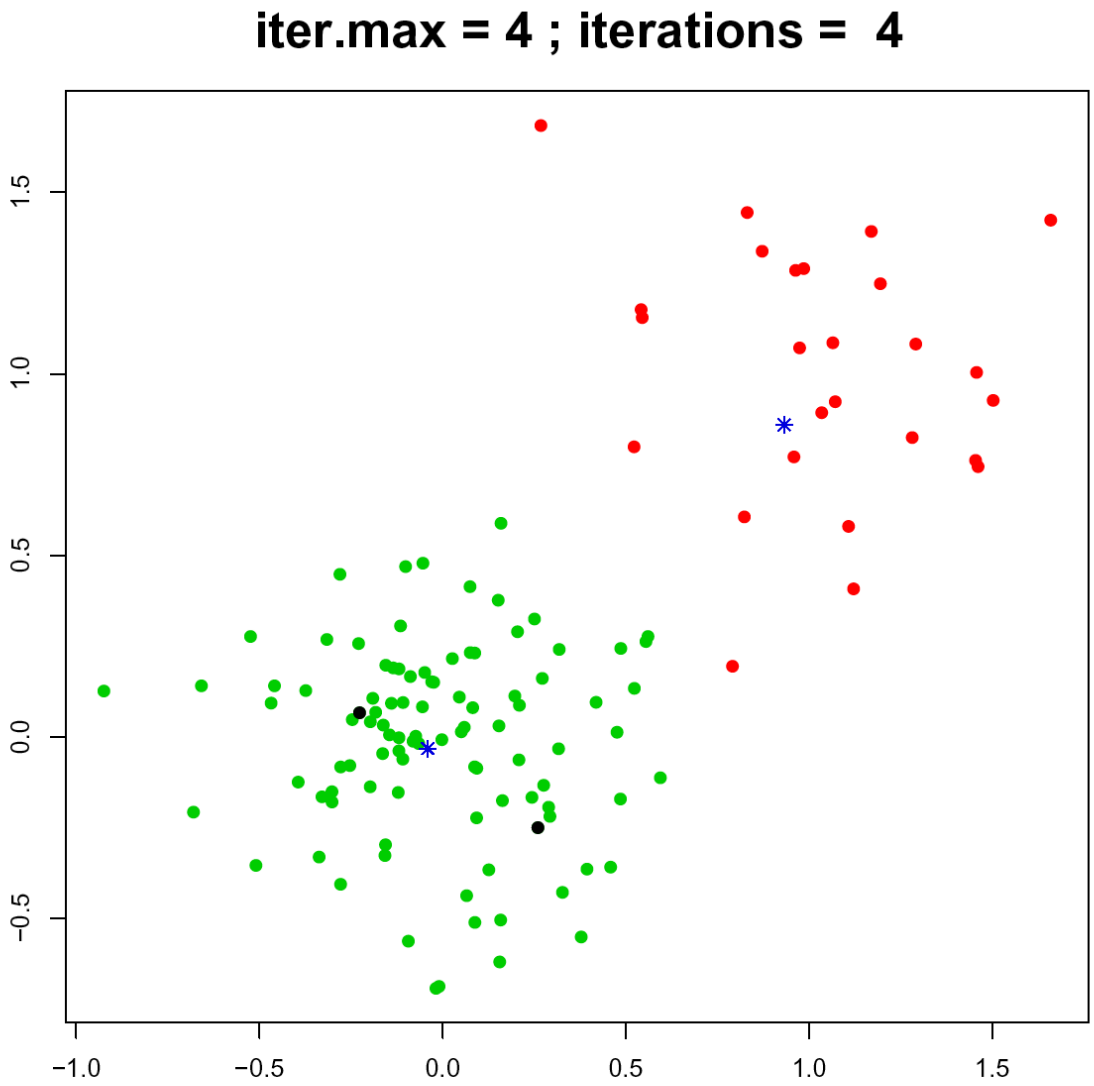
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 3 iterations



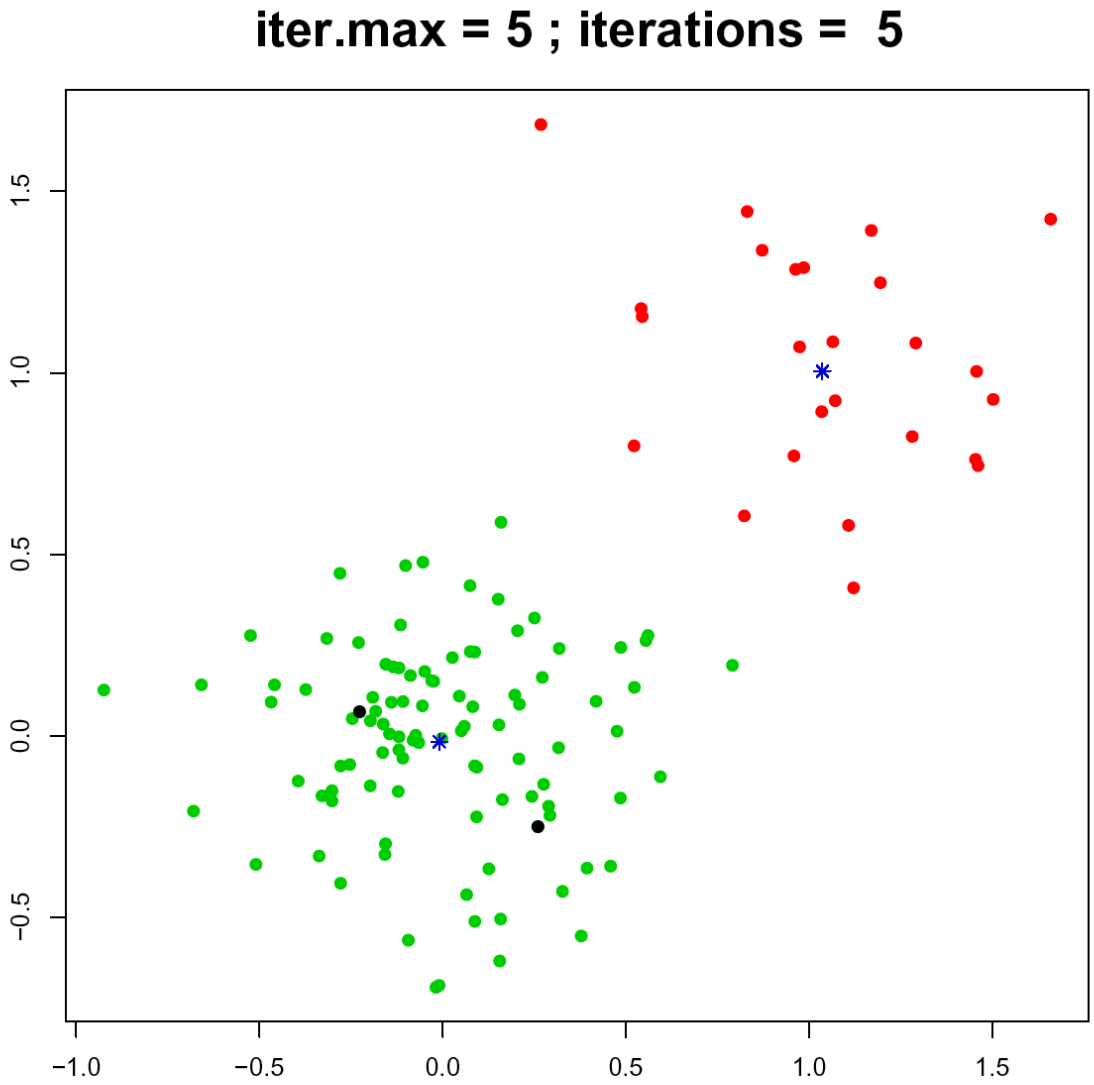
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 4 iterations



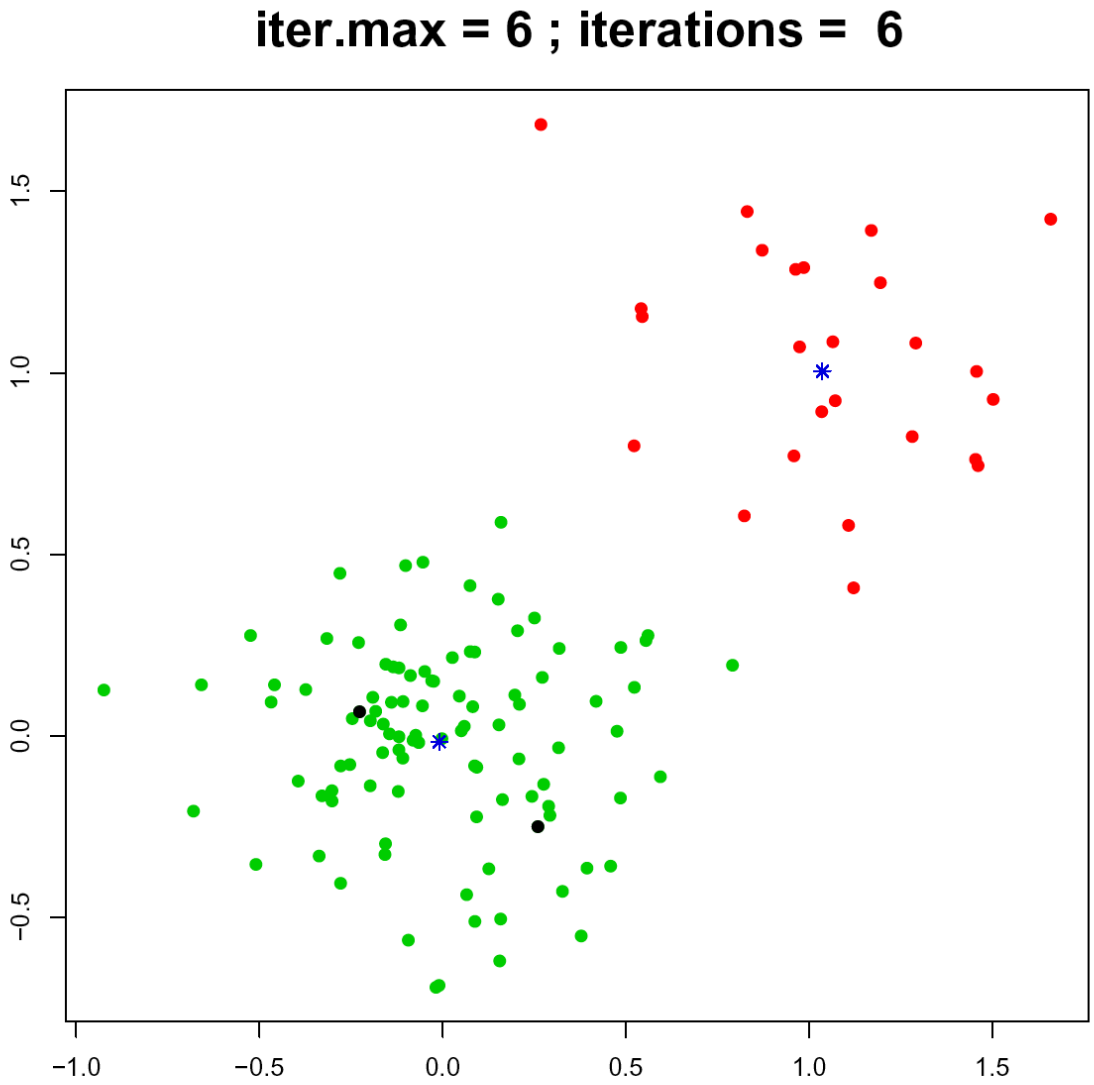
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 5 iterations



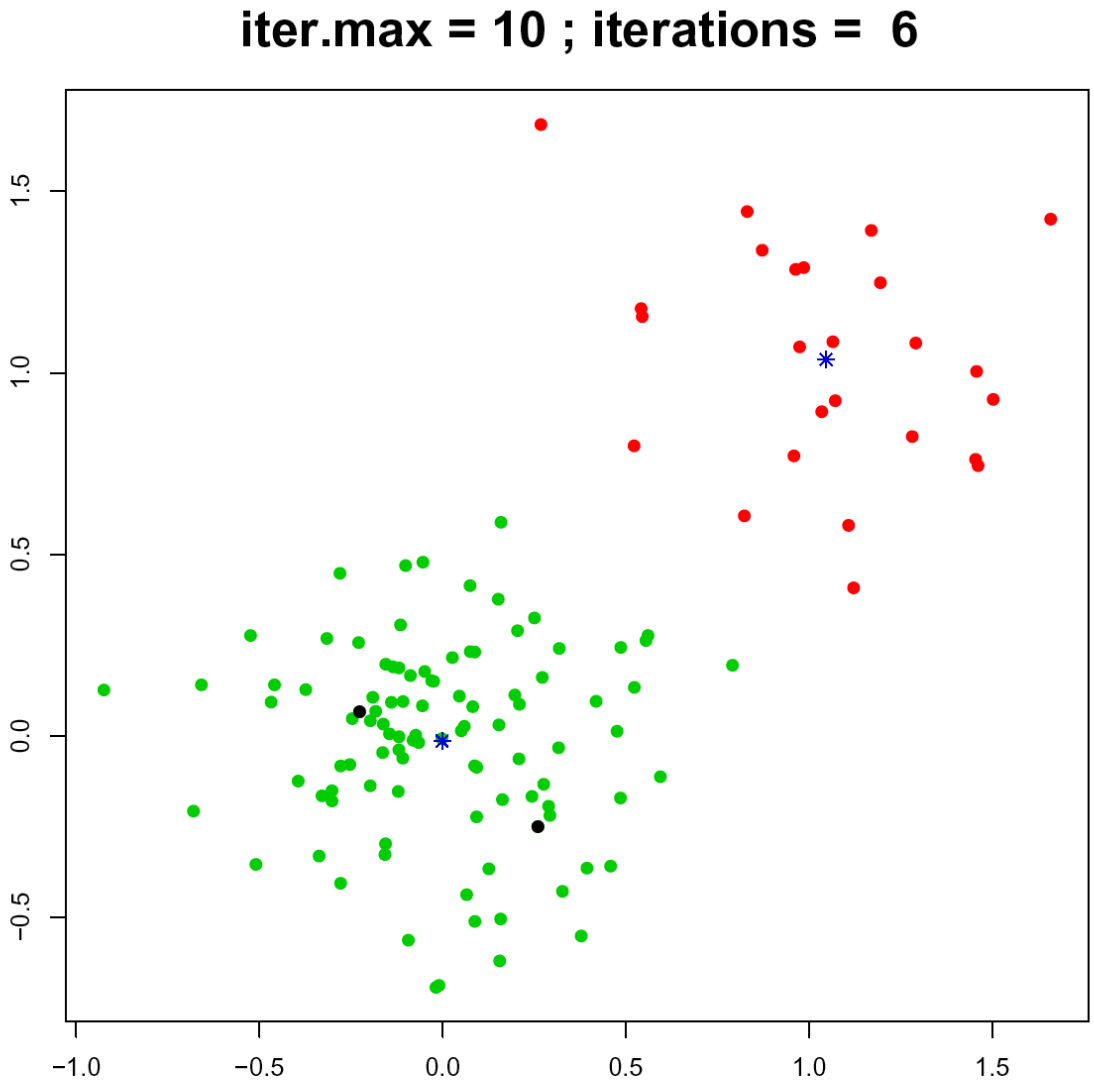
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 6 iterations



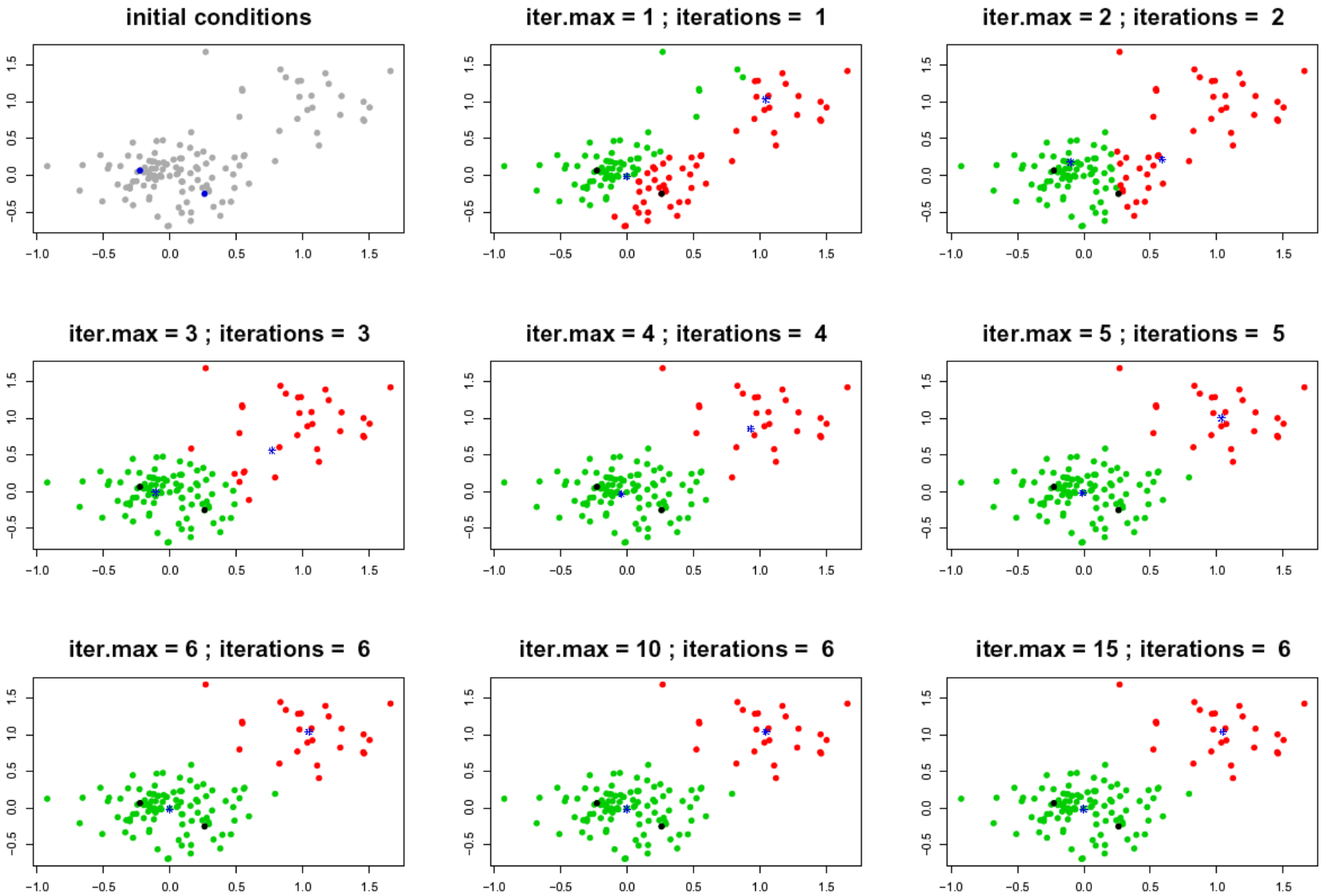
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 10 iterations

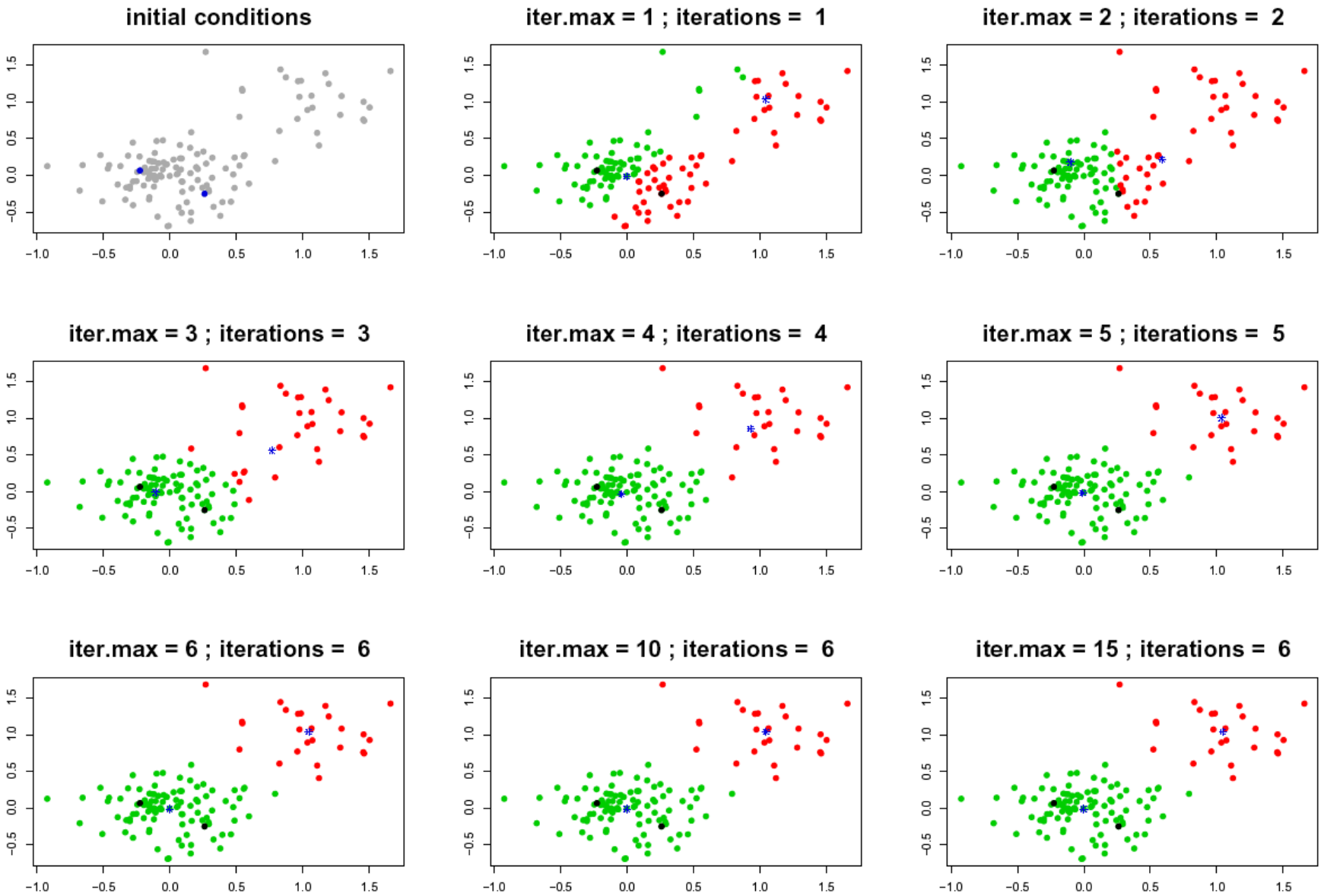


- After some iterations (6 in this case), the clusters and centres do not change anymore

Mobile centres example - random data



Mobile centres example - random data



K-means clustering

- K-means clustering is a variant of clustering around mobile centres
- After each assignation of an element to a centre, the position of this centre is re-calculated
- The convergence is much faster than with the basic mobile centre algorithm
  - after 1 iteration, the result might already be stable
- K-means is time- and memory-efficient for very large data sets (e.g. thousands of objects)

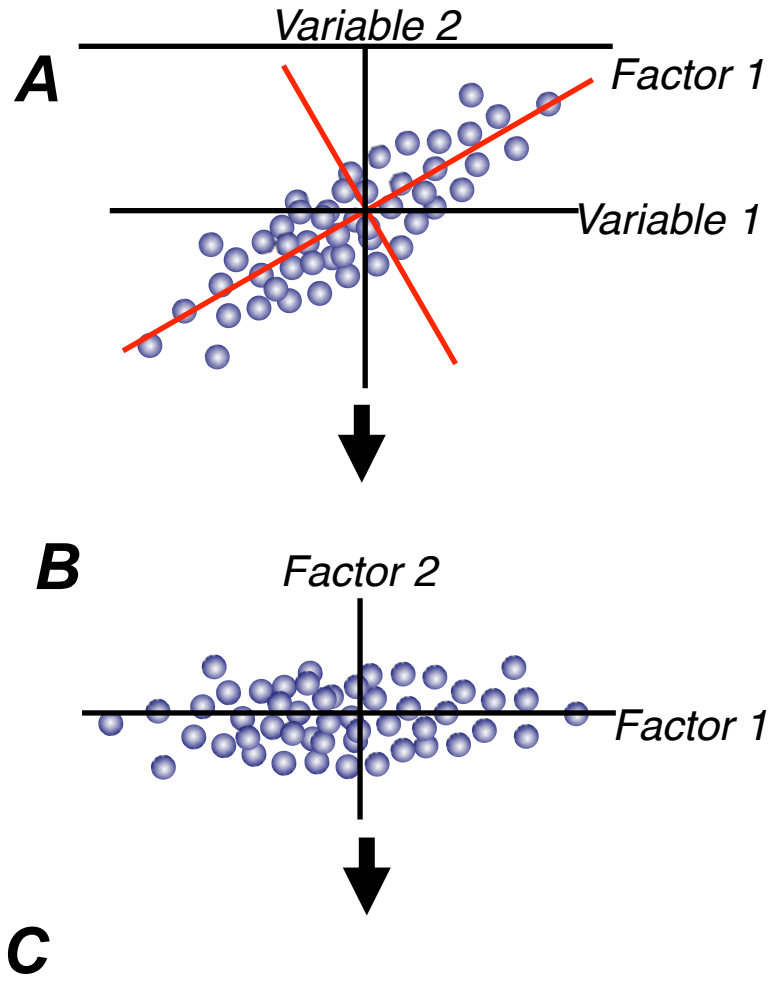
Exploring the data with multidimensional scaling

Principal component analysis

A. Multidimensional data
   - n objects, p variables (in this case p=2)

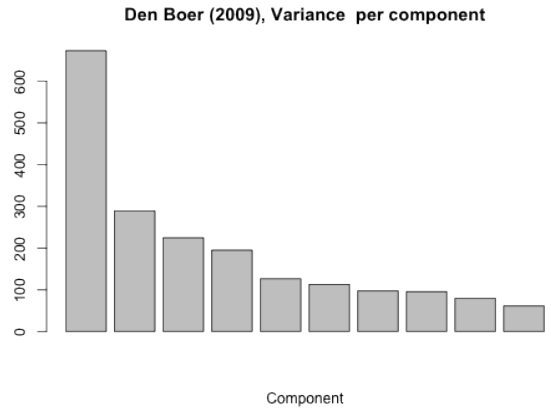
B. Principal components
   - n objects, p factors
   - Each factor is a linear combination of variables

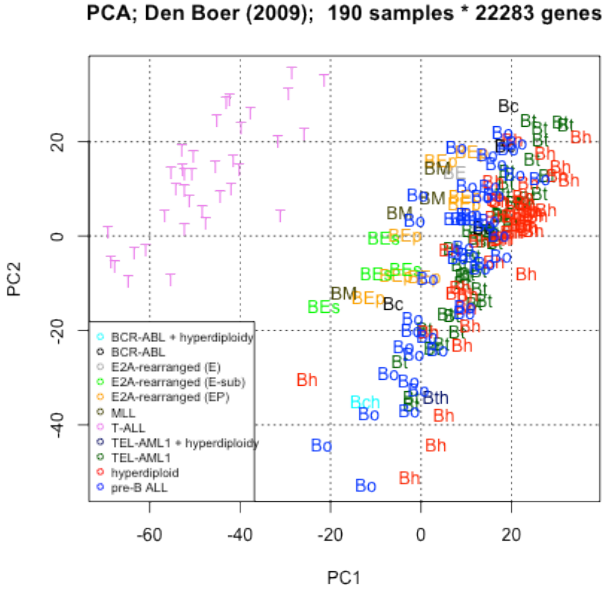
C. Reduction in dimensions
   - Selection of a subset of principal components
   - q factors, with q < p (in this case, q=1)

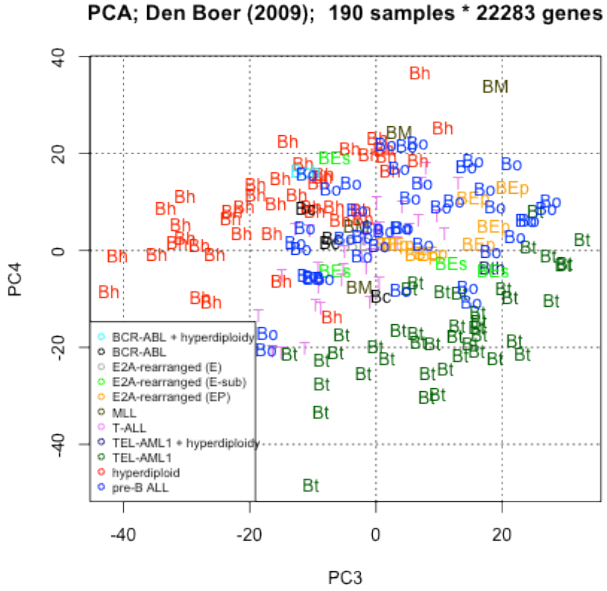


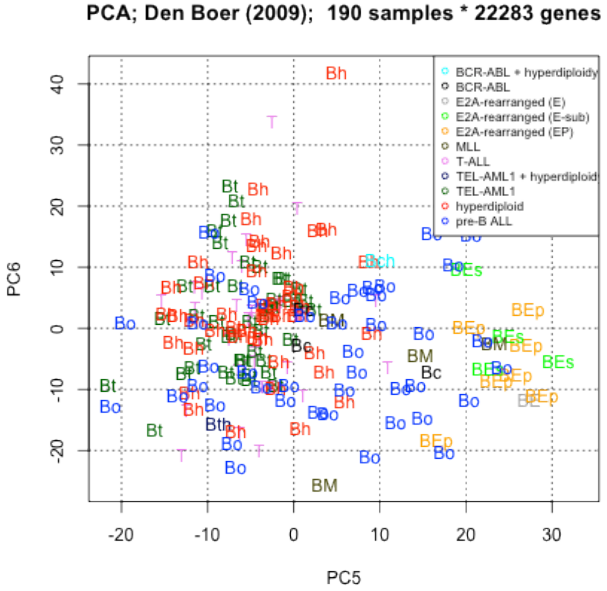
- Gilbert, D., Schroeder, M. & van Helden, J. (2000). Trends in Biotechnology 18) 487-495.

PCA – Den Boer (2009)

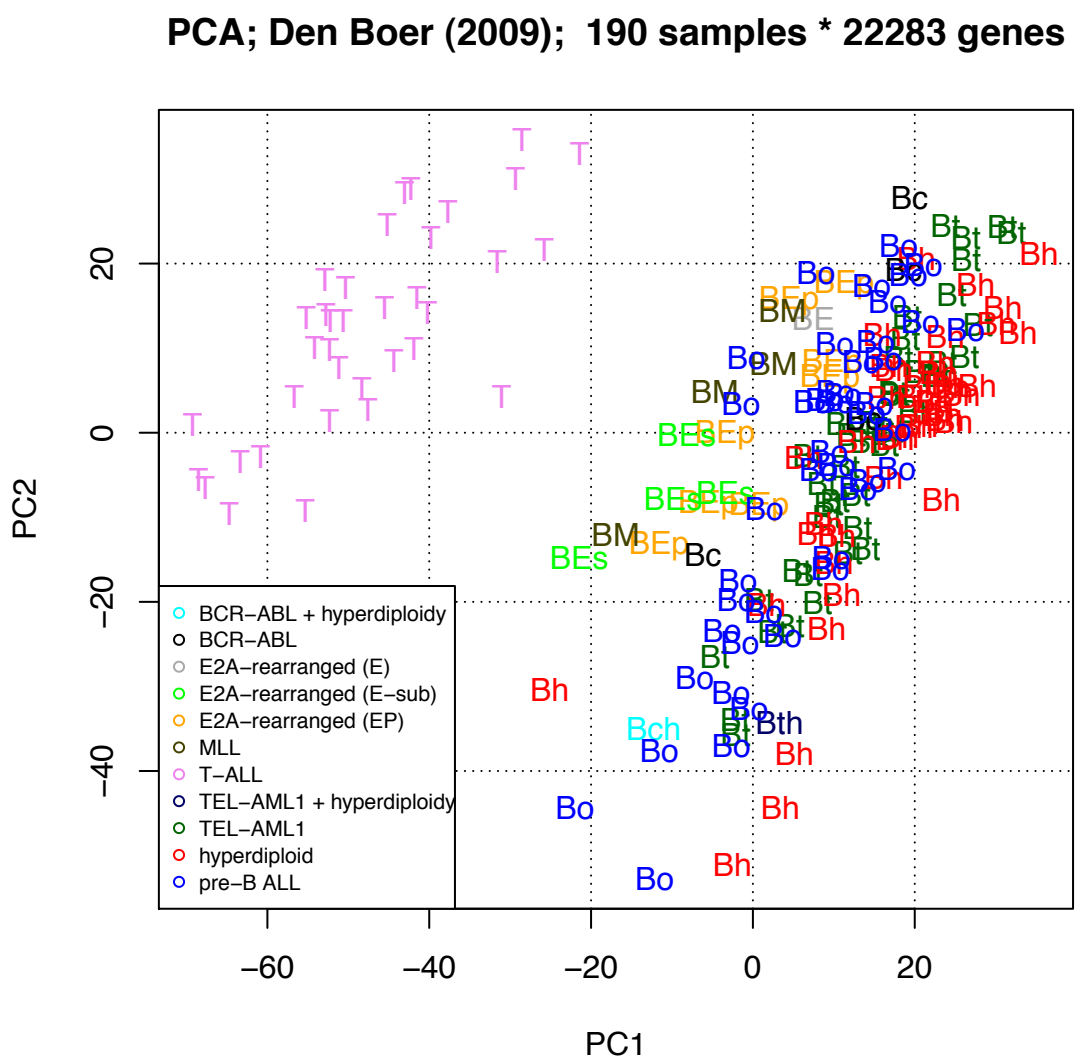




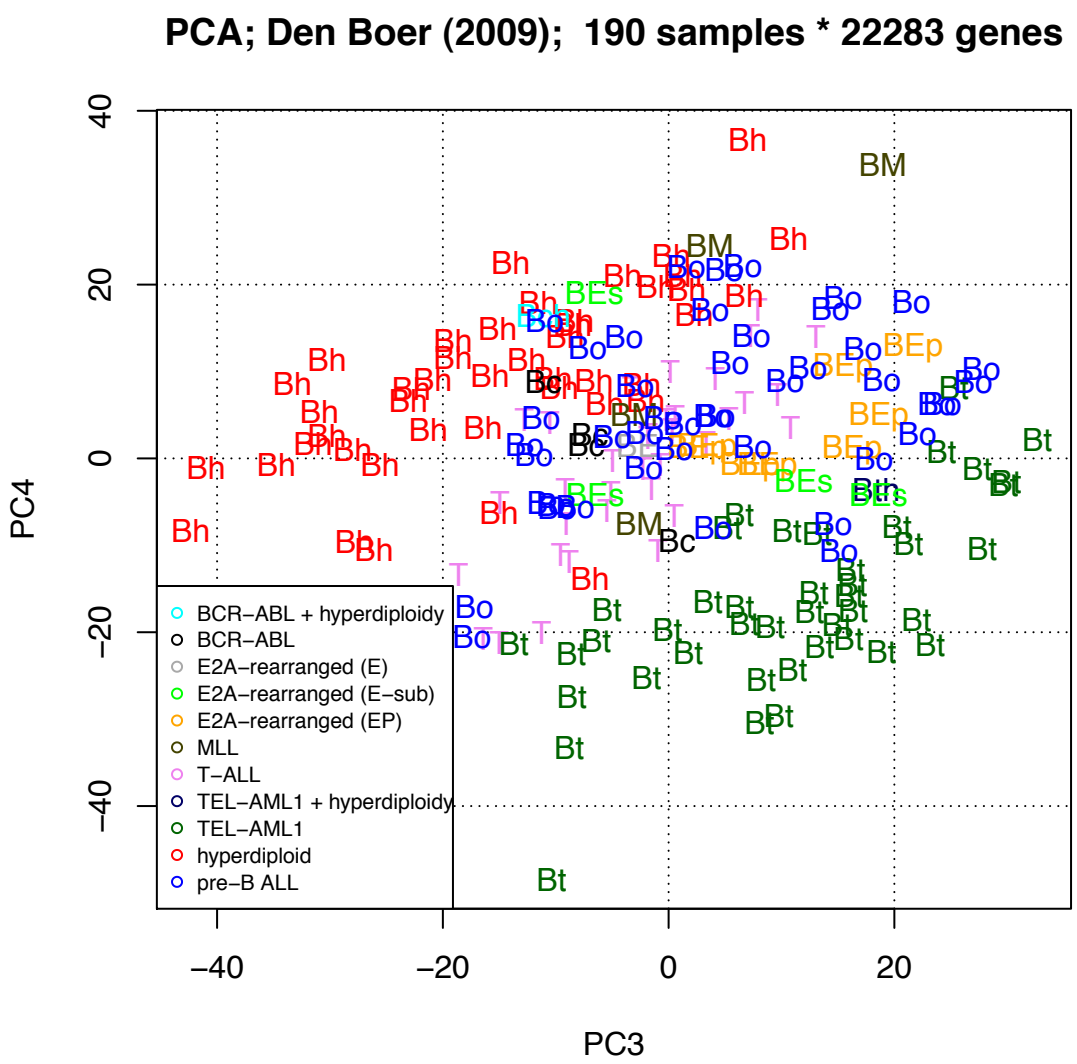




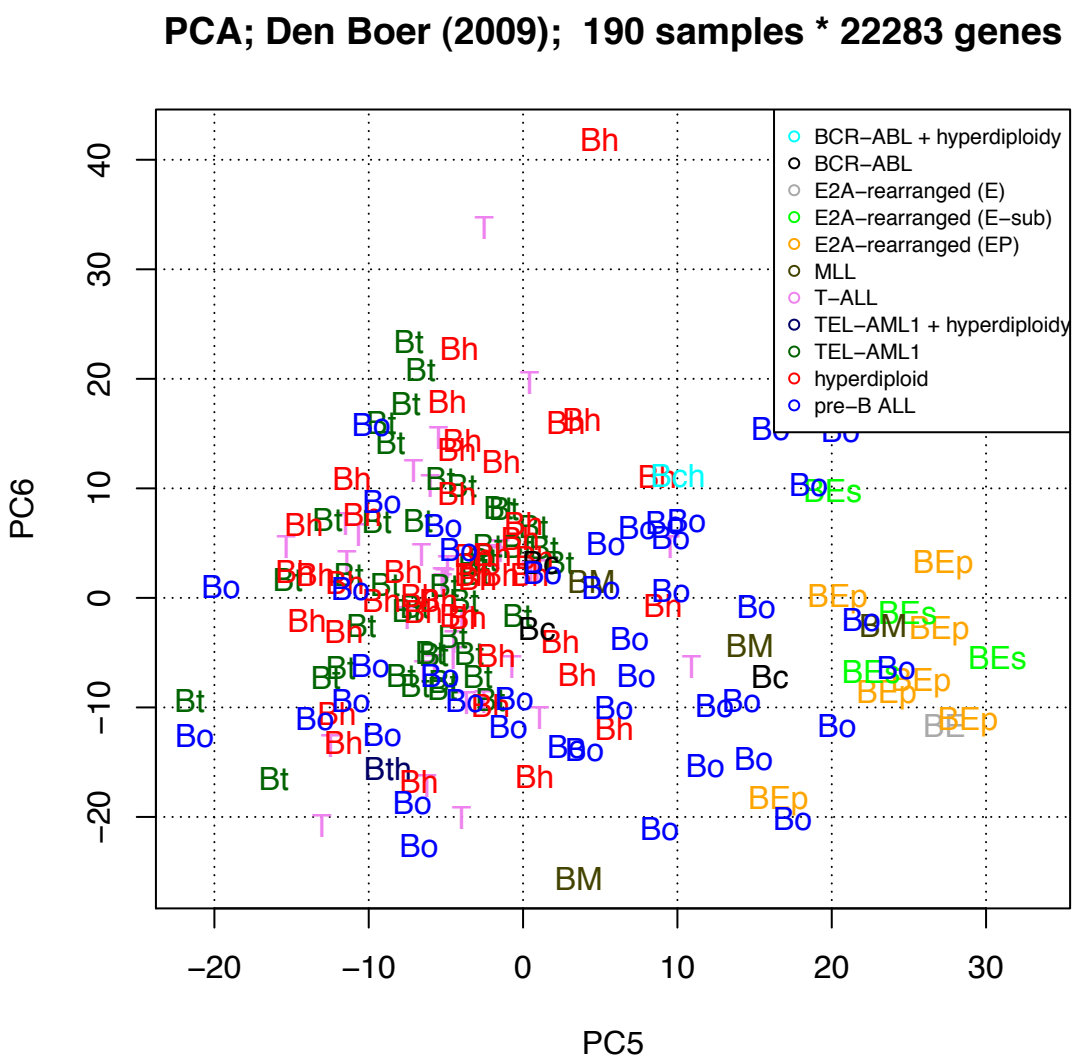
PCA – Den Boer (2009) – PC1 versus PC2



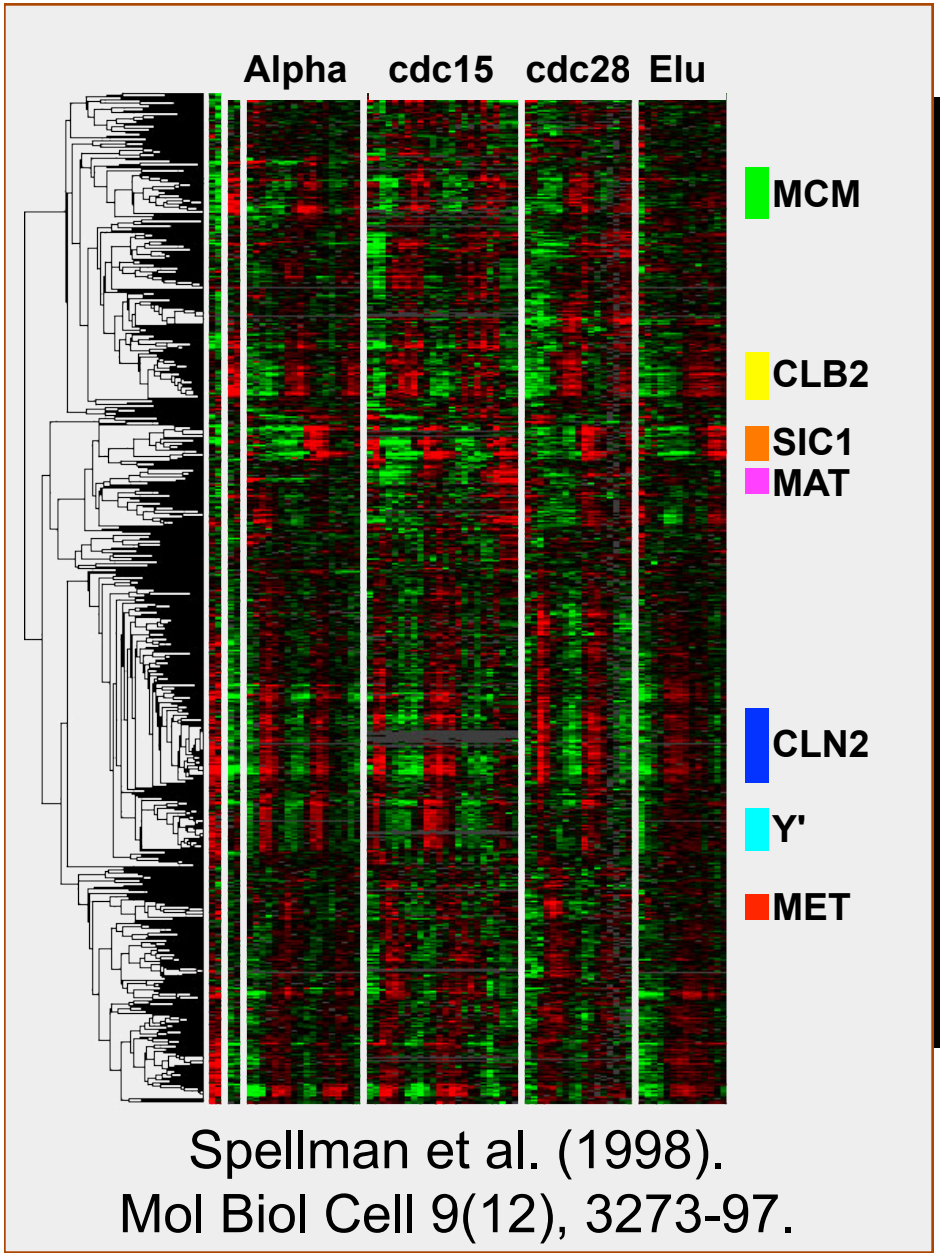
PCA – Den Boer (2009) – PC3 versus PC4



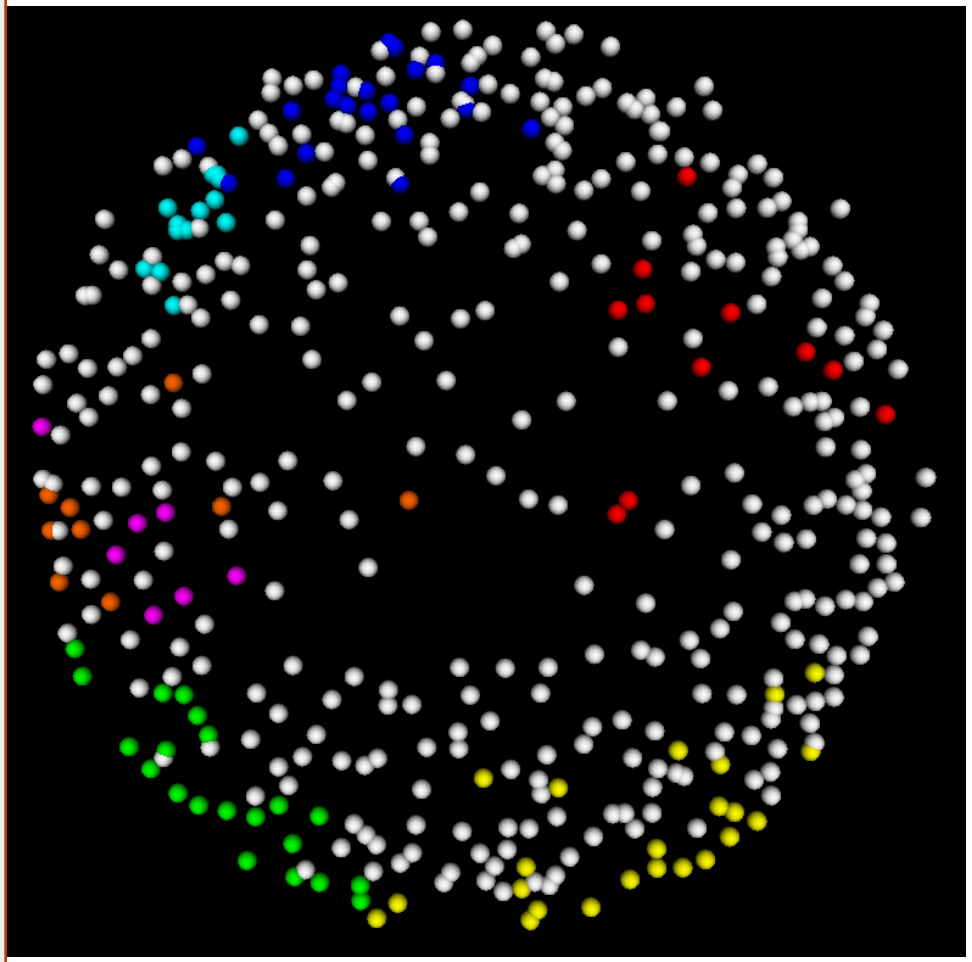
PCA – Den Boer (2009) – PC5 versus PC6



Singular value decomposition



Spellman et al. (1998).
Mol Biol Cell 9(12), 3273-97.

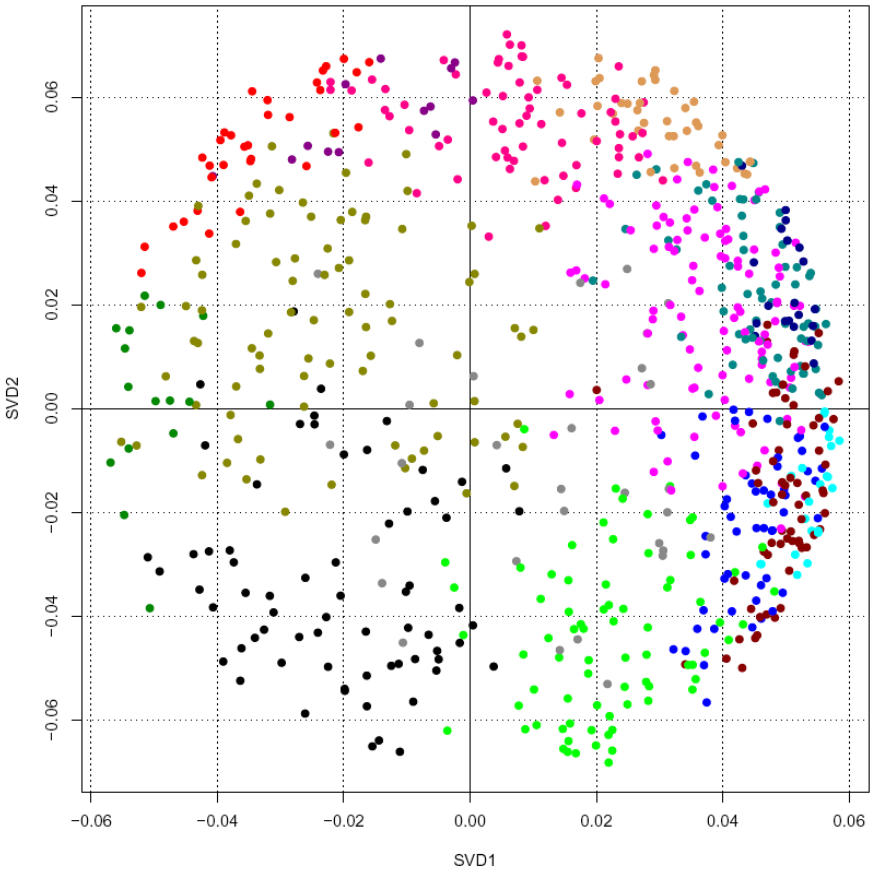


Gilbert et al. (2000).
Trends Biotech. 18(Dec), 487-495.

Singular value decomposition - Cell cycle

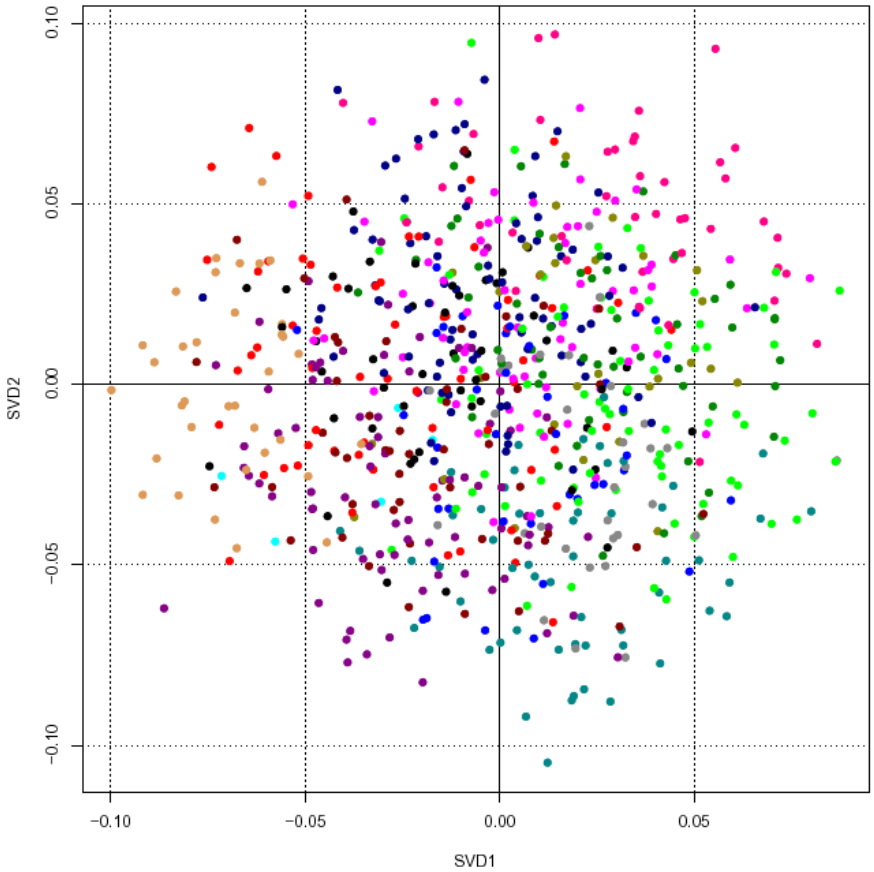
Cell cycle data

Spellman 98 elu



Randomized data

Spellman 98 elu permuted



- Calculate a distance matrix between objects
  - in this case Pearson's coefficient of correlation
- Assign 2D-coordinates which reflect at best the distances

Supervised classification

Supervised classification - Introduction

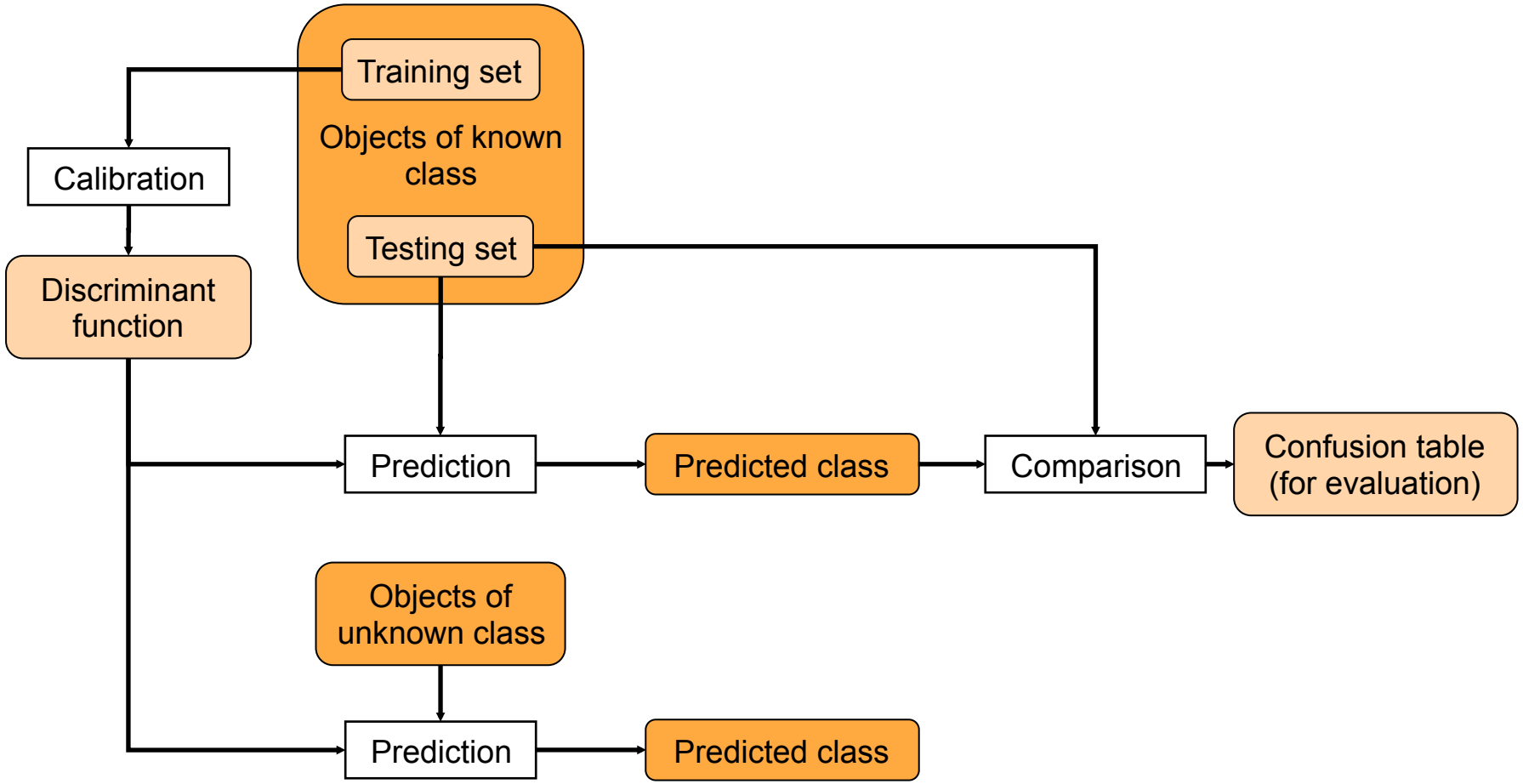
- In the previous chapter, we presented the problem of ***clustering***, which consists in grouping objects *without any a priori definition of the groups*. The group definition emerge from the clustering itself (class discovery). Clustering is thus *unsupervised*.

- In some cases, one would like to focus on some **pre-defined classes** :
  - classifying tissues as cancer or non-cancer
  - classifying tissues between different cancer types
  - classifying genes according to pre-defined functional classes (e.g. metabolic pathway, different phases of the cell cycle, ...)
- The classifier can be built with a **training set**, and used later for classifying new objects. This is called **supervised classification**.

Supervised classification methods

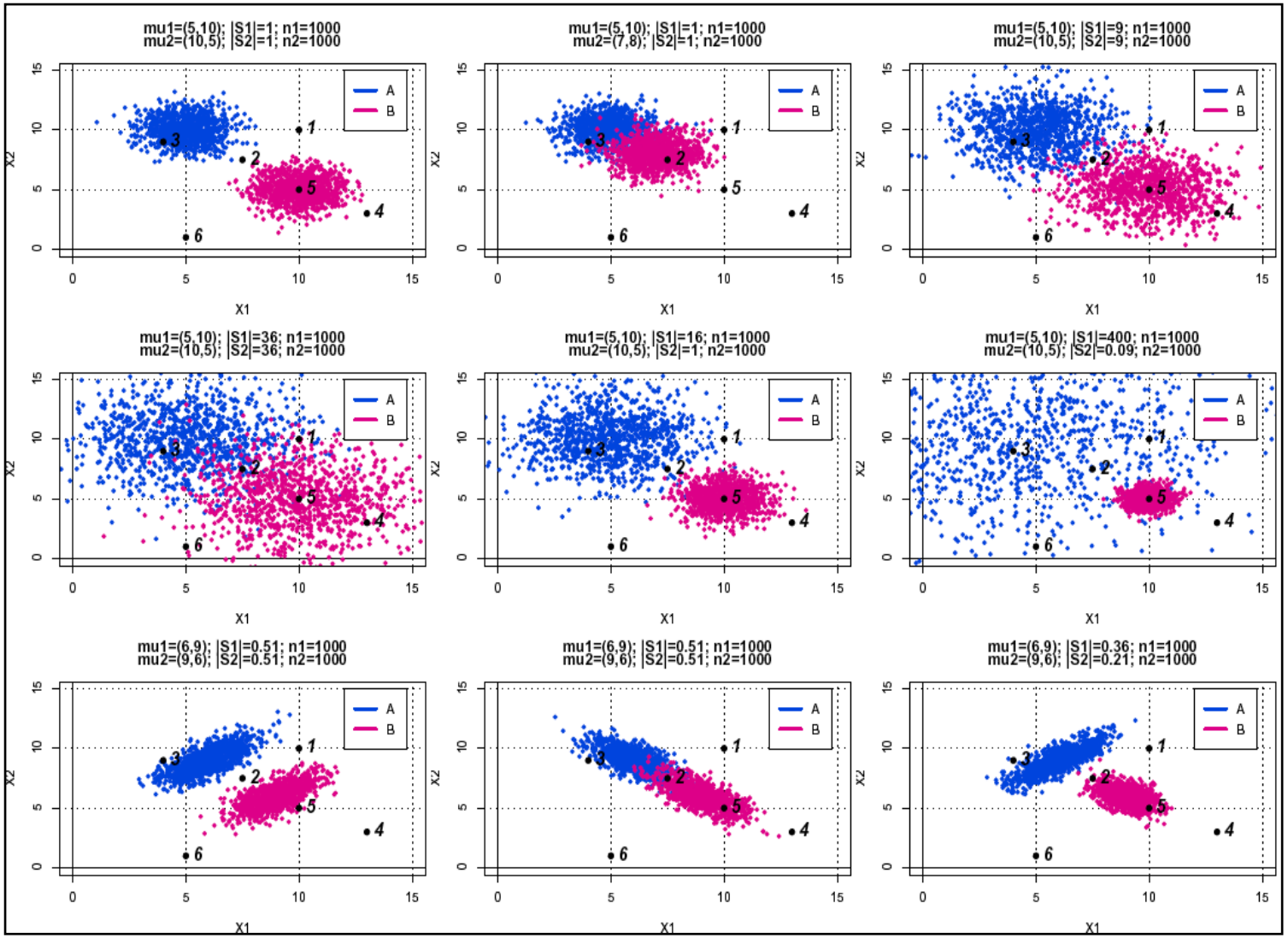
- There are many alternative methods for supervised classification
  - Discriminant analysis (linear or quadratic)
  - Bayesian classifiers
  - K-nearest neighbours (**KNN**)
  - Support Vector Machines (**SVM**)
  - Neural networks
  - ...
- Some methods rely on strong assumptions.
  - Discriminant analysis is based on an assumption of multivariate normality.
  - In addition, linear discriminant analysis assumes that all the classes have the same variance.
- Some methods require a large training set, to avoid over-fitting.
- The choice of the method thus depends on the structure and on the size of the data sets.

Discriminant analysis



Conceptual illustration with two predictor variables

- Given two predefined classes (A and B), try intuitively to assign a class to each new object (black dots).
- How confident do you feel for each of your predictions ?

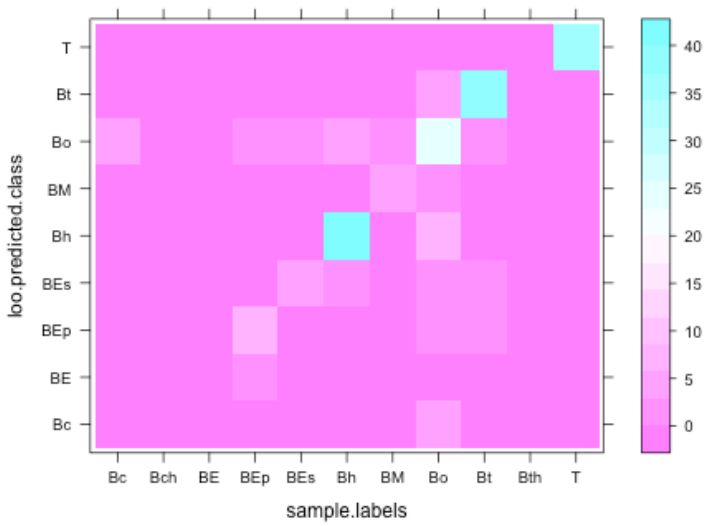


- What is the effect of the respective means ?
- What is the effect of the respective standard deviations ?
- What is the effect of the correlations between the two variables ?
- Note that the two population can have distinct correlations (orientations of the clouds)

Leave-one-out cross-validation – 20 genes , top-ranking by variance

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: **20** top-ranking probesets **sorted by decreasing variances**.
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 152
  - Total: 187
  - Hit rate: 81.3%
  - Error rate: 18.7%

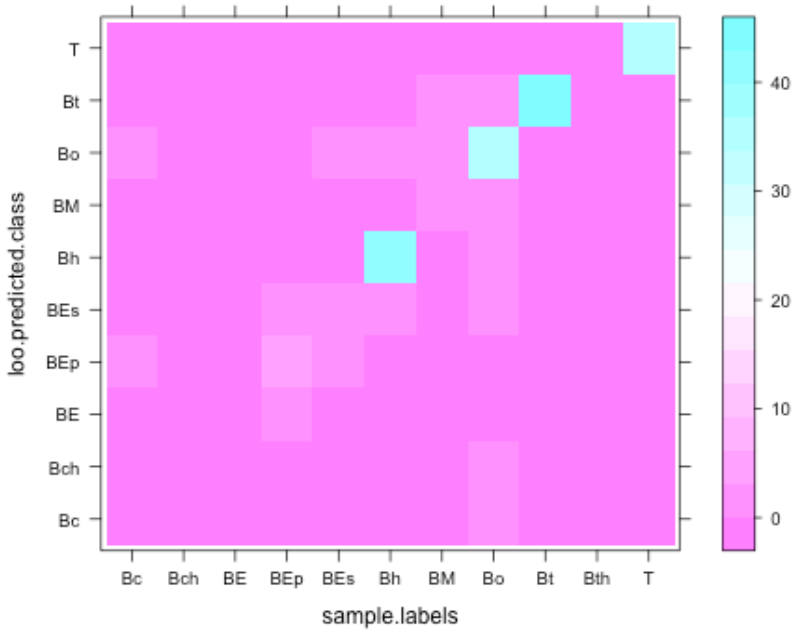
	loo.predicted.class								
sample.labels	Bc	BE	BEp	BEs	Bh	BM	Bo	Bt	T
Bc	0	0	0	0	0	0	4	0	0
Bch	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0
BEp	0	1	6	0	0	0	1	0	0
BEs	0	0	0	3	0	0	1	0	0
Bh	0	0	0	1	40	0	3	0	0
BM	0	0	0	0	0	3	1	0	0
Bo	3	0	2	2	7	1	25	4	0
Bt	0	0	2	1	0	0	1	39	0
Bth	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	36



Leave-one-out cross-validation – 20 genes top-ranked by various criteria

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: 20 top-ranking probesets **sorted by multi-criterion rank** (variance + two-groups Welch tests).
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 164
  - Total: 187
  - Hit rate: **87.7**%
  - Error rate: 12.3%

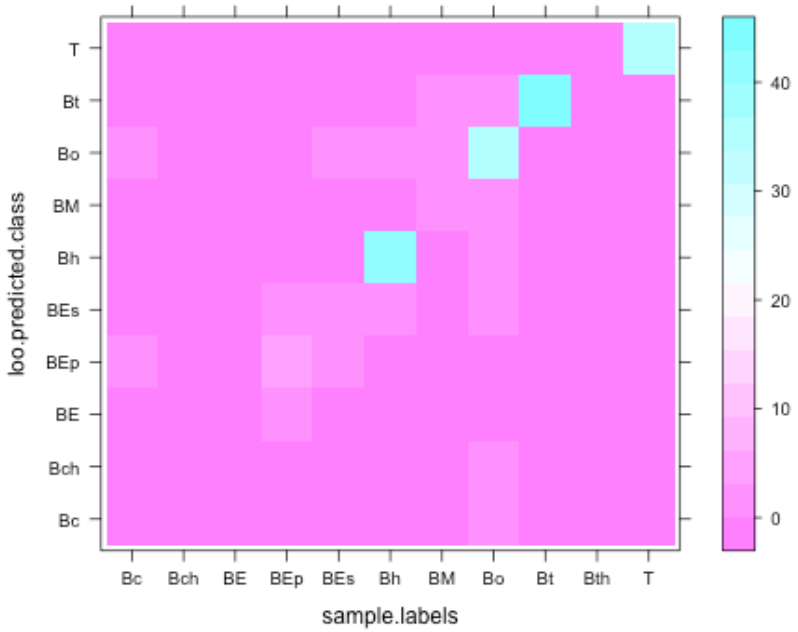
loo.predicted.class										
sample.labels	Bc	Bch	BE	BEp	BEs	Bh	BM	Bo	Bt	T
Bc	0	0	0	1	0	0	0	3	0	0
Bch	0	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0	0
BEp	0	0	1	6	1	0	0	0	0	0
BEs	0	0	0	1	2	0	0	1	0	0
Bh	0	0	0	0	1	41	0	2	0	0
BM	0	0	0	0	0	0	2	1	1	0
Bo	2	1	0	0	1	3	1	34	2	0
Bt	0	0	0	0	0	0	0	0	43	0
Bth	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	36



Leave-one-out cross-validation – 100 genes top-ranked by various criteria

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: **100** top-ranking probesets sorted by multi-criterion rank (variance + two-groups Welch tests).
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 168
  - Total: 187
  - Hit rate: **89.9**%
  - Error rate: 10.2%

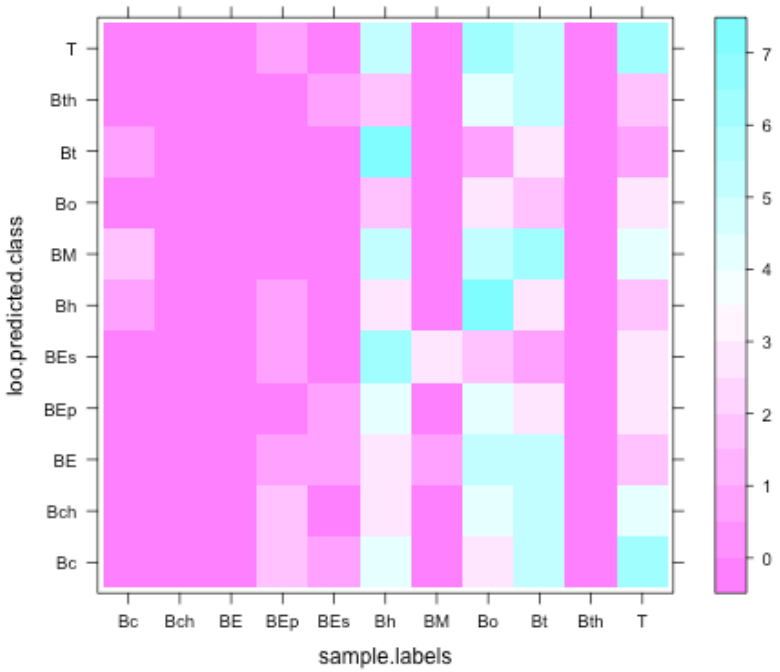
loo.predicted.class								
sample.labels	Bc	BEp	BEs	Bh	BM	Bo	Bt	T
Bc	0	0	0	2	0	2	0	0
Bch	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0
BEp	0	8	0	0	0	0	0	0
BEs	0	0	**3**	0	0	1	0	0
Bh	0	0	1	**41**	0	2	0	0
BM	0	0	1	0	**2**	1	0	0
Bo	2	1	0	5	0	**35**	1	0
Bt	0	0	0	0	0	0	**43**	0
Bth	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	**36**



Leave-one-out cross-validation – 200 genes top-ranked by various criteria

- Cross-validation of Linear Discriminant Analysis with Den Boer (2009).
- Variables: **200** top-ranking probesets sorted by multi-criterion rank (variance + two-groups Welch tests).
- Hit rate: proportion of correct predictions
  - Correct (diagonal): 15
  - Total: 187
  - Hit rate: 8%
  - Error rate: **92**%

loo.predicted.class											
sample.labels	Bc	Bch	BE	BEp	BEs	Bh	BM	Bo	Bt	Bth	T
Bc	0	0	0	0	0	1	2	0	1	0	0
Bch	0	0	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0	0	0
BEp	2	2	1	0	1	1	0	0	0	0	1
BEs	1	0	1	1	0	0	0	0	0	1	0
Bh	4	3	3	4	6	3	5	2	7	2	5
BM	0	0	1	0	3	0	0	0	0	0	0
Bo	3	4	5	4	2	7	5	3	1	4	6
Bt	5	5	5	3	1	3	6	2	3	5	5
Bth	0	0	0	0	0	0	0	0	0	0	0
T	6	4	2	3	3	2	4	3	1	2	6



Technical note: approach followed by DenBoer (differs from here)

- Multi-groups discrimination with 6 subtypes only (T-ALL, ETV6–RUNX1-positive, hyperdiploid, E2A- rearranged, BCR–ABL1-positive and MLL-rearranged)
- Training: 190 cases (COALL)
- Inner loop
  - Three-fold cross-validation: 2/3 cases for training, 1/3 for evaluation.
  - 100 iterations
- Variable filtering:
  - for each subtype, selection of the 50 lowest p-values with Wilcoxon's test.
  - For BCR-ABL1 and MLL, used 40 probesets from another source.
- Learning algorithm: radial-kernal support vector machine.
- Selection of the least number of probes by backward selection.

Bh	hyperdiploid	44
Bo	pre-B ALL	44
Bt	TEL-AML1	43
T	T-ALL	36
BEp	E2A-rearranged (EP)	8
Bc	BCR-ABL	4
BEs	E2A-rearranged (E-sub)	4
BM	MLL	4
Bch	BCR-ABL + hyperdiploidy	1
BE	E2A-rearranged (E)	1
Bth	TEL-AML1 + hyperdiploidy	1