JGB71E - Bioinformatique appliquée

Analyse statistique du transcriptome

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.univ-amu.fr/

Reminder: sampling

Estimating a parameter of the population from a sample

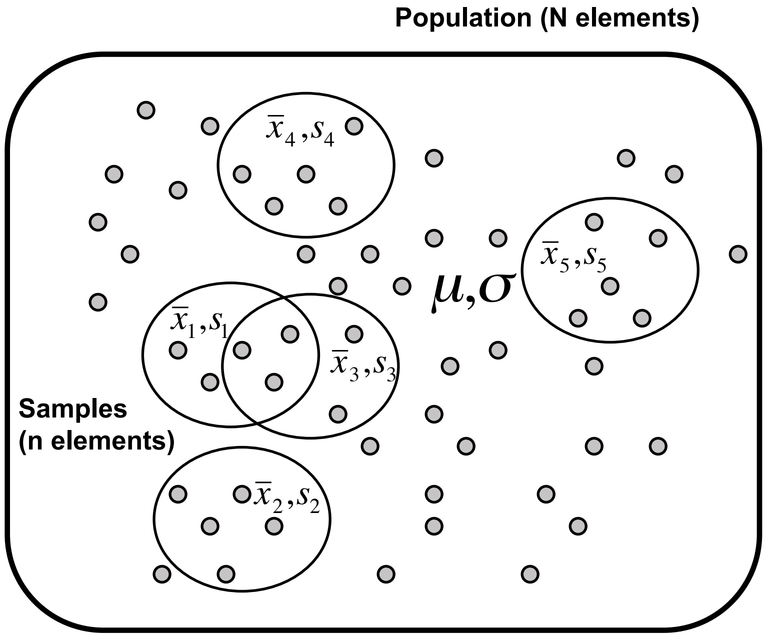
- Situation
  - We dispose of a sample drawn randomly from a population.
  - We can easily compute the ***sample parameters*** such as mean, standard deviation.

$\bar{x}, s$

  - From these, we would like to ***estimate*** the corresponding ***population parameters***.

μ, σ

- Problem
  - The sample mean and standard deviation will vary depending on the particular sample.
  - The population mean and standard deviation are however constant.
- Question
  - To which extent can we rely on the mean and standard deviation of the sample to estimate the mean and standard deviation of the population ?



Population mean (unknown) μ

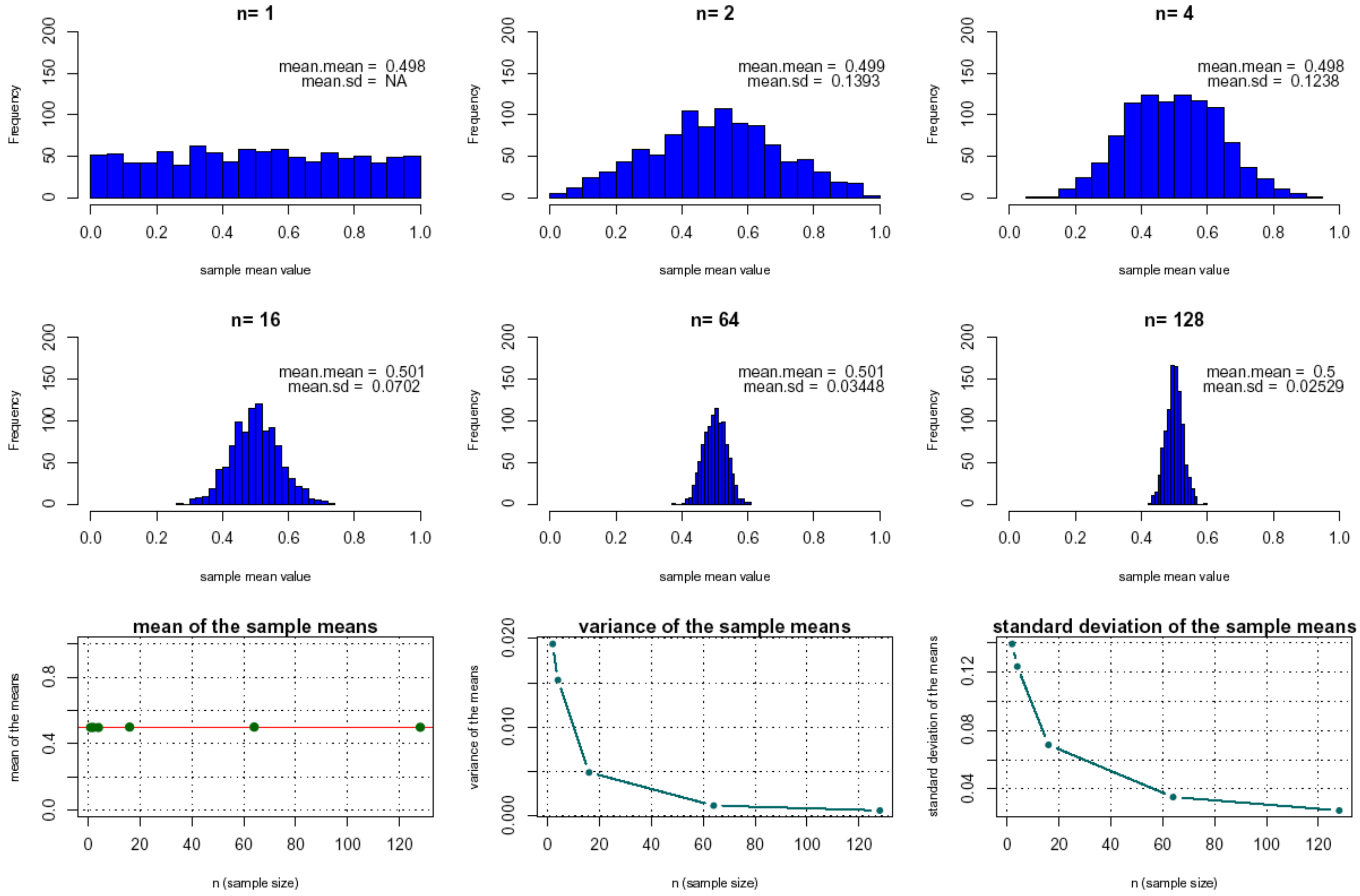
Population sd (unknown) σ

Sample $\{x_1, x_2, ..., x_n\}$

Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample sd $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

Sampling distribution of the mean



Sampling distribution of the mean

- In this simulation, the population is drawn randomly from a uniform distribution.
- When the sample size (n) increases, the sample mean tends towards a normal distribution. This is an application of the ***central limit theorem***.
- On the histograms of the previous slide, the distribution of the sample means is always centred around 0.5, irrespective of the sample size. The mean of the sample is an ***unbiased estimate*** of the population mean: its expected value equals the mean of the population.
- Note: the variance and standard deviation of the sample mean decrease as the sample size (n) increase.

- The expectation for the sample mean is the population mean. The sample mean is thus an **unbiased** estimator of the population mean.

$$\boxed{E\left(\overline{X}\right) = m}$$

$$\boxed{\hat{m} = \overline{X}}$$

(the hat means "estimate")

Sampling distribution - Sample variance

- The sample variance is a **biased** estimator of the population variance.

$$\boxed{E\left(S^2\right) = \frac{(n-1)}{n}\sigma^2}$$

$$\boxed{E\left(S\right) = \sqrt{\frac{(n-1)}{n}}\sigma}$$

- For this reason, one has to introduce a **corrective factor n/(n-1)** when one tries to estimate the population variance from the sample variance.

$$\boxed{\hat{\sigma}^2 = \frac{n}{n-1}s^2}$$

$$\boxed{\hat{\sigma} = \sqrt{\frac{n}{n-1}}s}$$

- Remarks
  - This correction only matters for small samples.
    For large samples, n/(n-1) ~ 1.
  - This correction is already included in some packages (e.g. R): when you compute the variance of a vector, the function var() returns the estimate for population variance rather than the actual variance of the input numbers (the sample).

Sampling distribution - The standard error

- The expectation for the sample mean is the population mean. The sample mean is thus an **unbiased** estimator of the population mean.

$$\boxed{E\left(\overline{X}\right) = m}$$

$$\boxed{\hat{m} = \overline{X}}$$

(the hat means "estimate")

- The variance of the sample mean distribution **differs** from the population variance.

for a finite population

$$\boxed{\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$$

for an infinite population

$$\boxed{\sigma_{\overline{X}}^2 = \sigma^2 / n}$$

- The standard deviation of the sample mean is called **standard error**. The standard error decreases when n increases. The larger is the sample, the more reliable is the estimation of the mean.

For a finite population

$$\boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}}$$

For an infinite population

$$\boxed{\sigma_{\overline{X}} = \sigma / \sqrt{n}}$$

Reminder: mean comparison testing

Introduction

- H_0 null hypothesis
- H_A alternative hypothesis
- AH_0 acceptation of the null hypothesis
- RH_0 rejection of the null hypothesis
- α $P(RH_0|H_0)$ probability to reject the null hypothesis when it is true
- β $P(AH_0|H_A)$ probability to accept the null hypothesis when it is false
- $1\text{-}\beta$ $P(RH_0|H_A)$ Power of the test (also called rejection power)

	H0	HA
AH0	Correct acception	Type II error β risk
RH0	Type I error α risk	Correct rejection

Two-tailed test of mean equality

- Two-tailed test
  - $H_0 : m_1 = m_2$
  - $H_1 : m_1 \neq m_2$
- Principle of the test
  - Estimate the difference between m_1 and m_2
  - Compare this estimation with the theoretical distribution
- Usually, the variance is a priori not know, and has to be estimated
  - Warning: the variance of a difference is the sum of variances
  - The formula for estimating the whether the populations are supposed to have or not similar variances
- The theoretical distribution is thus the *Student* (t)
  - $k = n_1 + n_2 - 2$ degrees of freedom
  - α is shared between the two tails → use the value for $t_{1-\alpha/2}$ in Student's table

$$\boxed{t_{obs} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\hat{\sigma}_{m_1 - m_2}}}$$

Reject H_0 if $t_{obs} \geq t_{1-\alpha/2}$

Estimating the variance of the difference between sample means

- Generally, the variance of the populations (σ1 and σ2) are not known a priori.
  - They have thus to be estimated from the two samples.
  - The variance of the sample is a biased estimation of the variance of the population (see chapter on estimation).
  - Each variance estimate needs thus to be corrected by a factor *n/(n-1)*.
- The estimation of the variance will raise an error, which has to be taken into account for the calculation of significance. This will be done differently depending on two considerations
  - Can we assume that the two populations have the same variance ?
  - Do the two sample have the same size ?
- Note: the estimators of variance have to be corrected for the bias (hence the n_i-*1* numerator in the formula).

$$\boxed{\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\hat{\sigma}^2_{\bar{X}_1} + \hat{\sigma}^2_{\bar{X}_2}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

Populations with the same variance

- When one can assume that the two populations have the same variance (Student test), the variance of the difference is estimated as follows.

$$\widehat{\sigma}^2 = \widehat{\sigma}_1^2 = \widehat{\sigma}_2^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$\widehat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}} = \sqrt{\widehat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

If the two samples have the same size (n_1=n_2=n), this formula can be simplified.

$$\widehat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{n s_1^2 + n s_2^2}{n + n - 2}\left(\frac{1}{n} + \frac{1}{n}\right)} = \sqrt{\frac{s_1^2 + s_2^2}{n - 1}}$$

Population with different variances

- When one cannot assume that the two populations have the same variance (Welch test), the variance of the difference is estimated as follows

$$\boxed{\widehat{\sigma}_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{n_1 s_1^2}{n_1(n_1-1)} + \frac{n_2 s_2^2}{n_2(n_2-1)}} = \sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}}}$$

Unequal variances : the Welch test

- If one cannot assume variance equality, the same statistics (t_{obs}) can be used, but the number of degrees of freedom k is calculated with the formula besides.
  - Note: the formula to compute k in a Welch t-test returns positive Real numbers. The "number" of degrees of freedom does not need to be a Natural number anymore.
- This test is called the ***Welch t-test***.

$$t_{obs} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}}$$

$$k = \frac{\left[\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}\right]^2}{\frac{1}{n_1 - 1}\left[\sqrt{\frac{s_1^2}{n_1 - 1}}\right]^2 + \frac{1}{n_2 - 1}\left[\sqrt{\frac{s_2^2}{n_2 - 1}}\right]^2}$$

Statistics for Bioinformatics

Detecting differentially expressed genes

Jacques van Helden

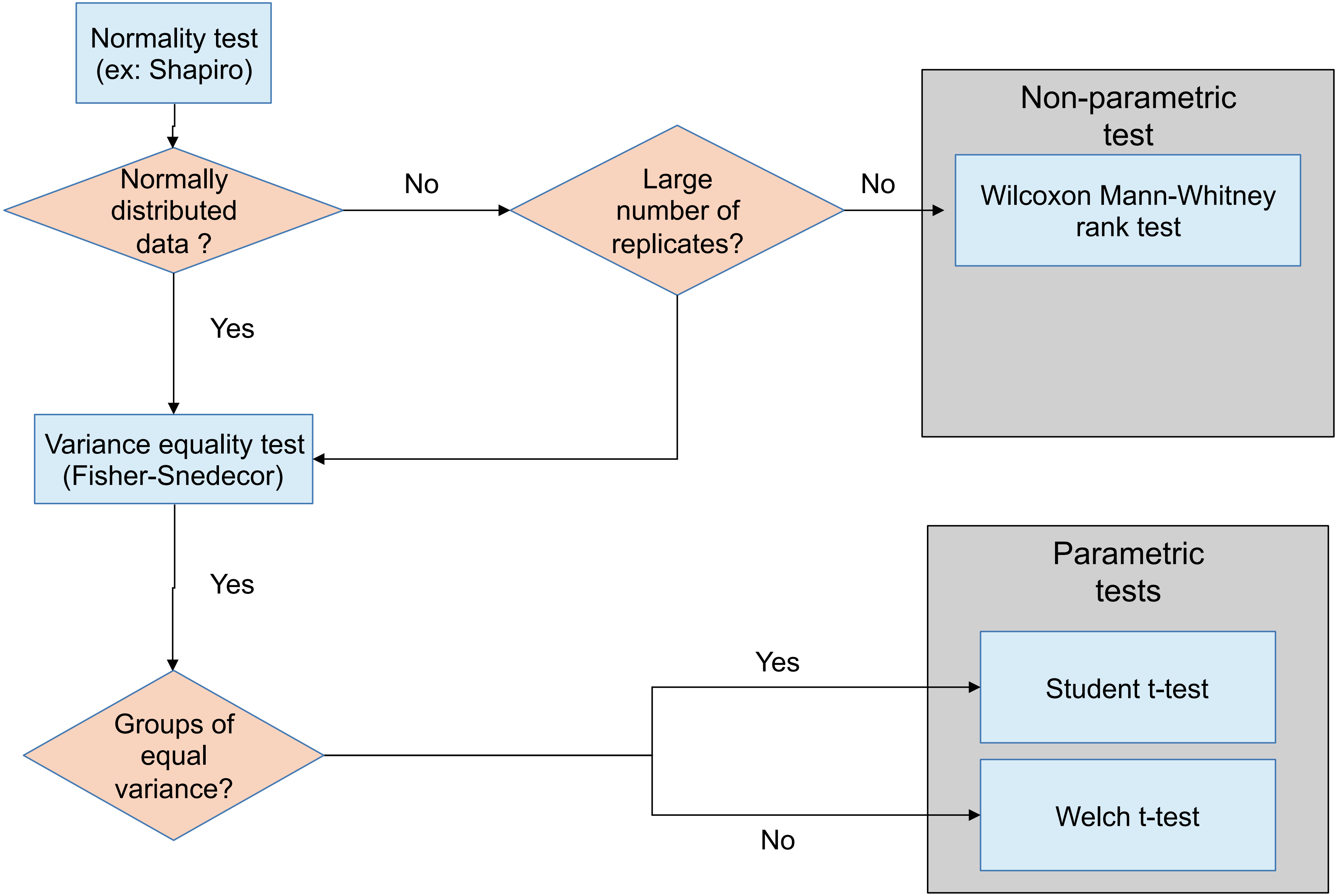
Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univ-amu.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)

Principle of differential analysis

- **Two-groups differential analysis with Welch test**
  - 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 with 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).

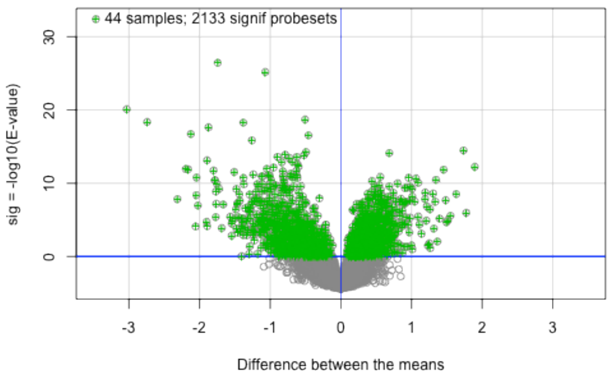
Flow chart for the choice of a two-group mean comparison test



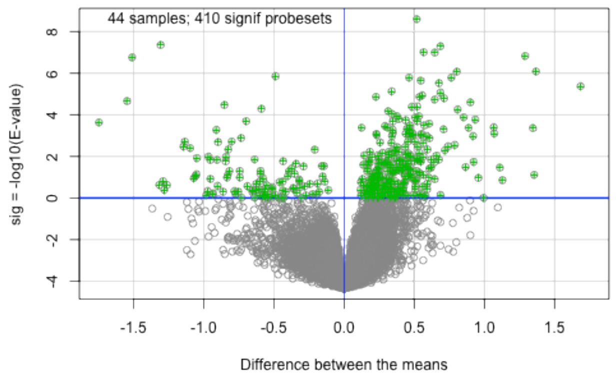
- Adapted from Firas Hammami

Welch test results for two-groups differential analysis

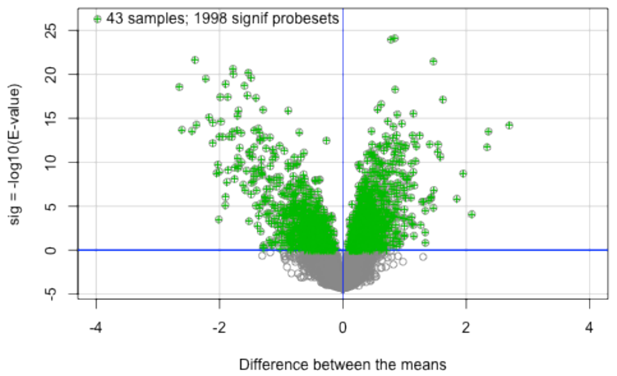
Bh vs others



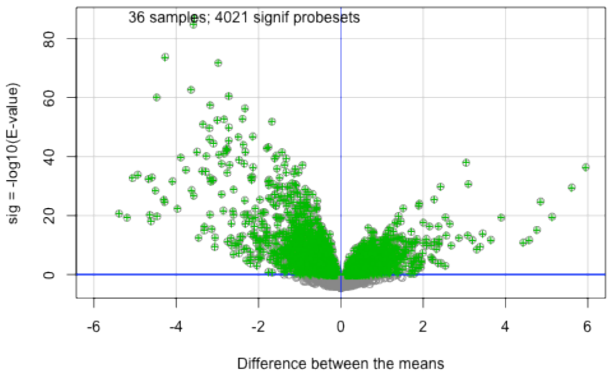
Bo vs others



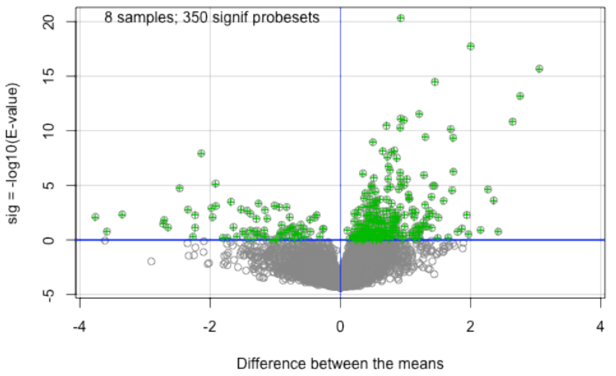
Bt vs others



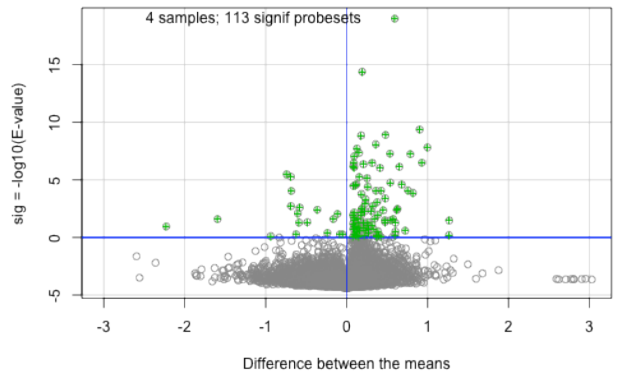
T vs others



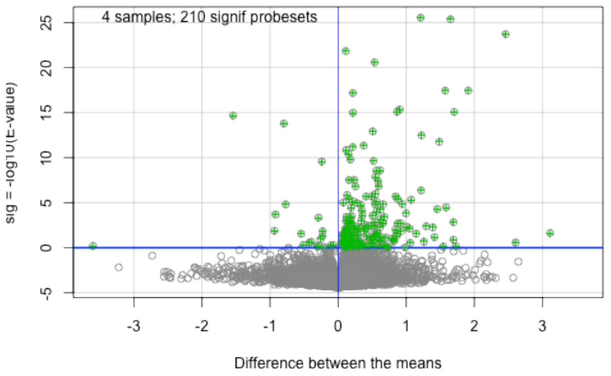
BEp vs others



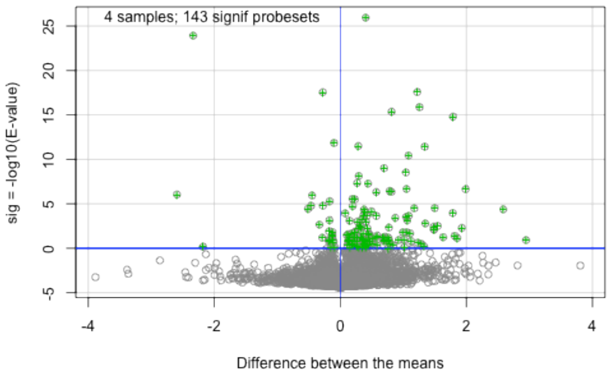
Bc vs others



BEs vs others



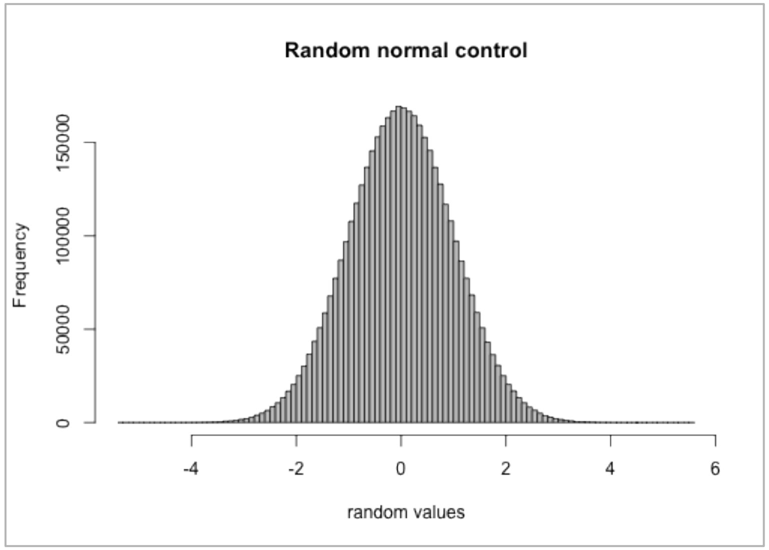
BM vs others

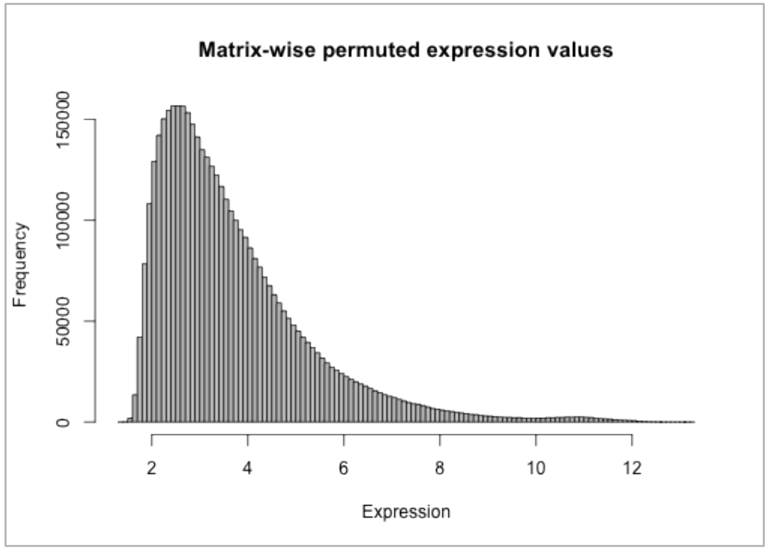


Bh	hyperdiploid	44
Bo	pre-B ALL	44
Bt	TEL-AML1	43
T	T-ALL	36
BEp	E2A-rearranged (EP)	8
Bc	BCR-ABL	4
BEs	E2A-rearranged (E-sub)	4
BM	MLL	4
Bch	BCR-ABL + hyperdiploidy	1
BE	E2A-rearranged (E)	1
Bth	TEL-AML1 + hyperdiploidy	1

Negative controls

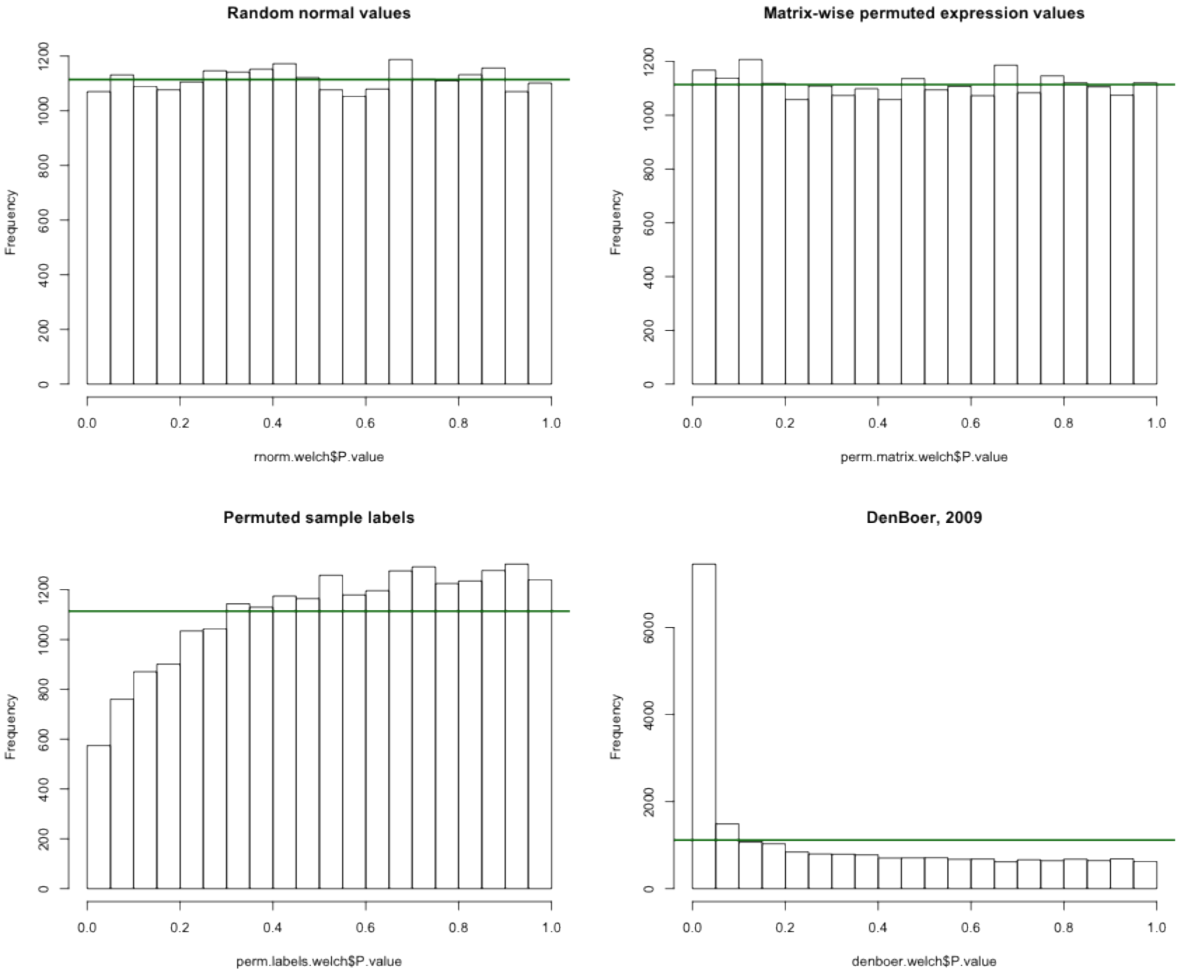
- It is always useful to check empirically the significance of a selection procedure.
- For this, we can build negative controls, i.e. datasets where no difference is expected between groups.
- 3 negative controls
  - **Random normal values**. We build a fake expression matrix by generating random numbers 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

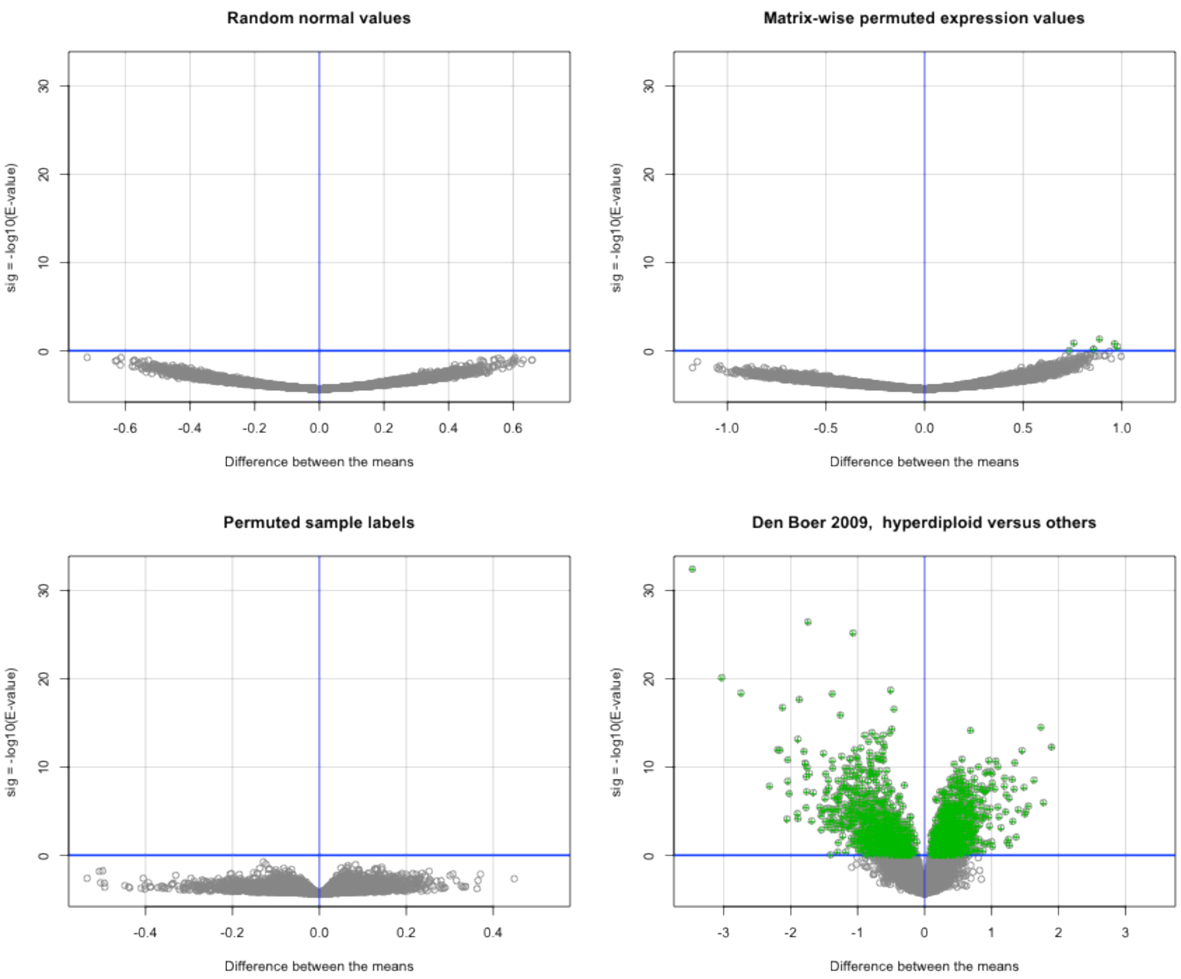
- Data set: Den Boer et al. (2009).
- Welch test: hyperdiploid versus other types of Acute Lymphoblastic Leukemia.
- P-value distribution
  - Abscissa: frequency class of the P-value.
  - Ordinate: number of genes falling in this class.
- 3 negative controls
  - Random normal values.
    - Flat distribution, as expected.
  - Matrix-wise random permutation of expression values.
    - Flat distribution, as expected.
  - Permutation of sample labels, analysis of the original expression matrix.
    - Under-representation of low P-values. Strange.
- Original expression matrix.
  - Striking over-representation of the low P-values. This likely corresponds to differentially expressed genes.



- 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.

Distribution of P-values from Welch test

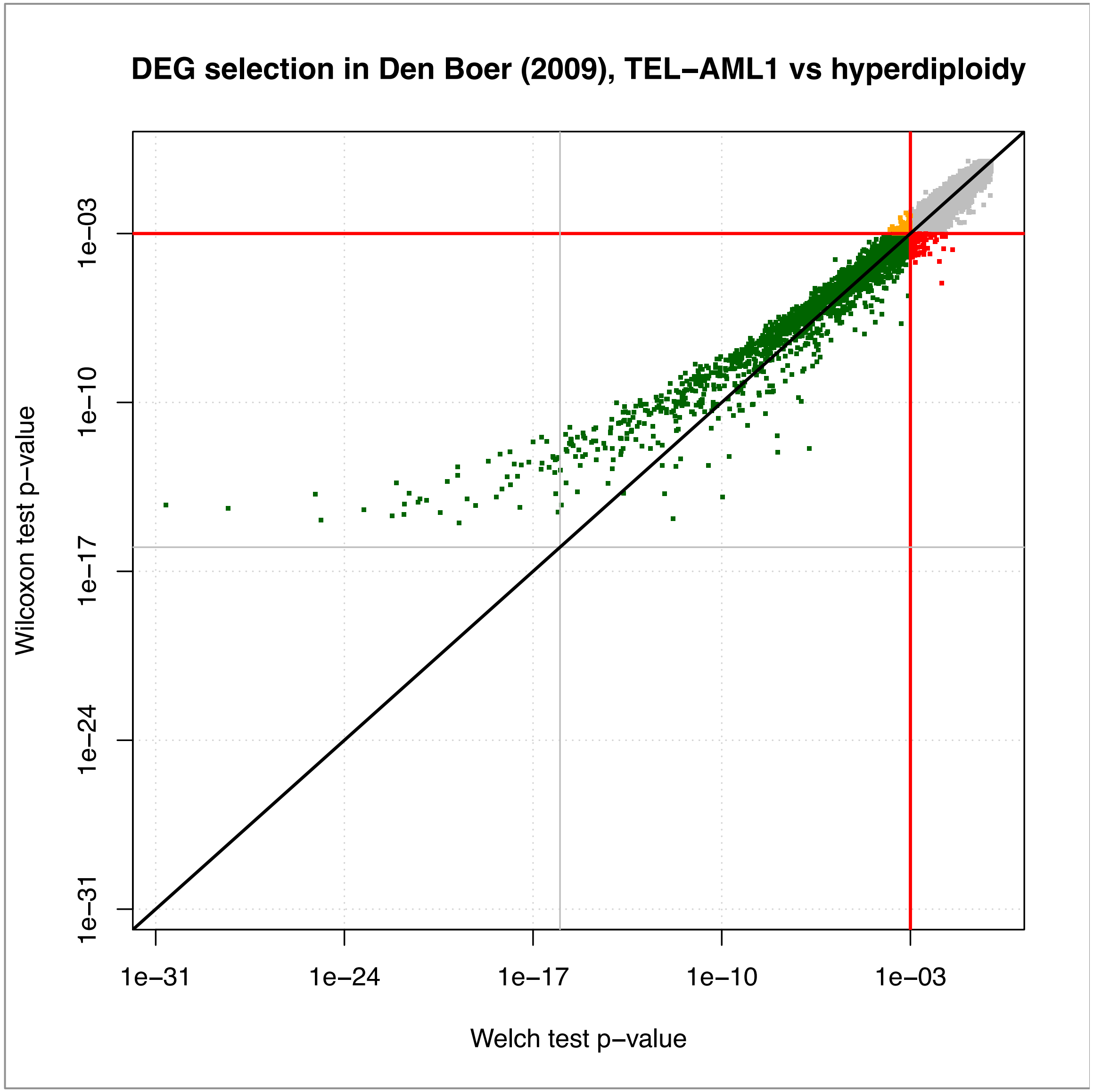
- Data set: Den Boer et al. (2009).
- Welch test: hyperdiploid versus other types of Acute Lymphoblastic Leukemia.
- Volcano plots
  - Abscissa: difference between the means
  - Ordinate: significance of the test.
- 3 negative controls
  - Random normal values.
    - All significances are negative.
  - 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.

Welch versus Wilcoxon test

- Student and Welch tests are called “parametric”, because they rely on the assumption of normality of the data.
- With Affymetrix microarrays, the measured intensities generally strongly discard from normality.
- An alternative way to select differentially expressed genes is to apply a (non-parametric) Wilcoxon test to each gene separately.
- We ran a Welch and Student test on the 22,283 probesets of Den Boer dataset, to detect differentially expressed genes between two cancer types: TEL-AML1 and hyperdiploidy, resp.



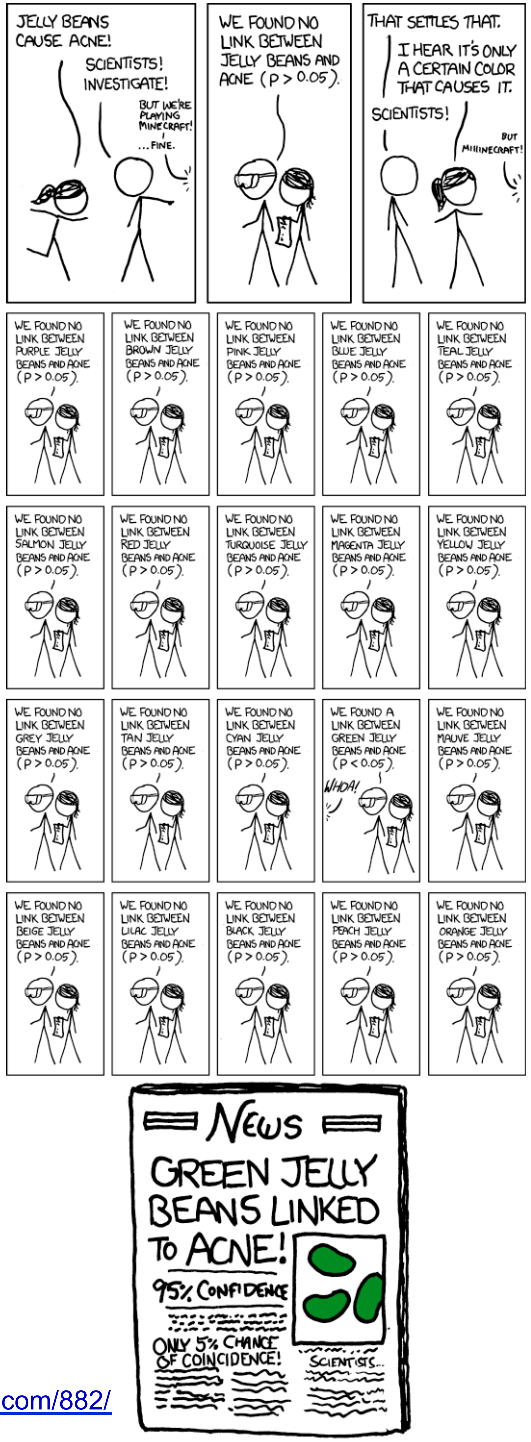
Statistics for bioinformatics

Multiple testing

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.univ-amu.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

http://xkcd.com/882/

What is a P-value ?

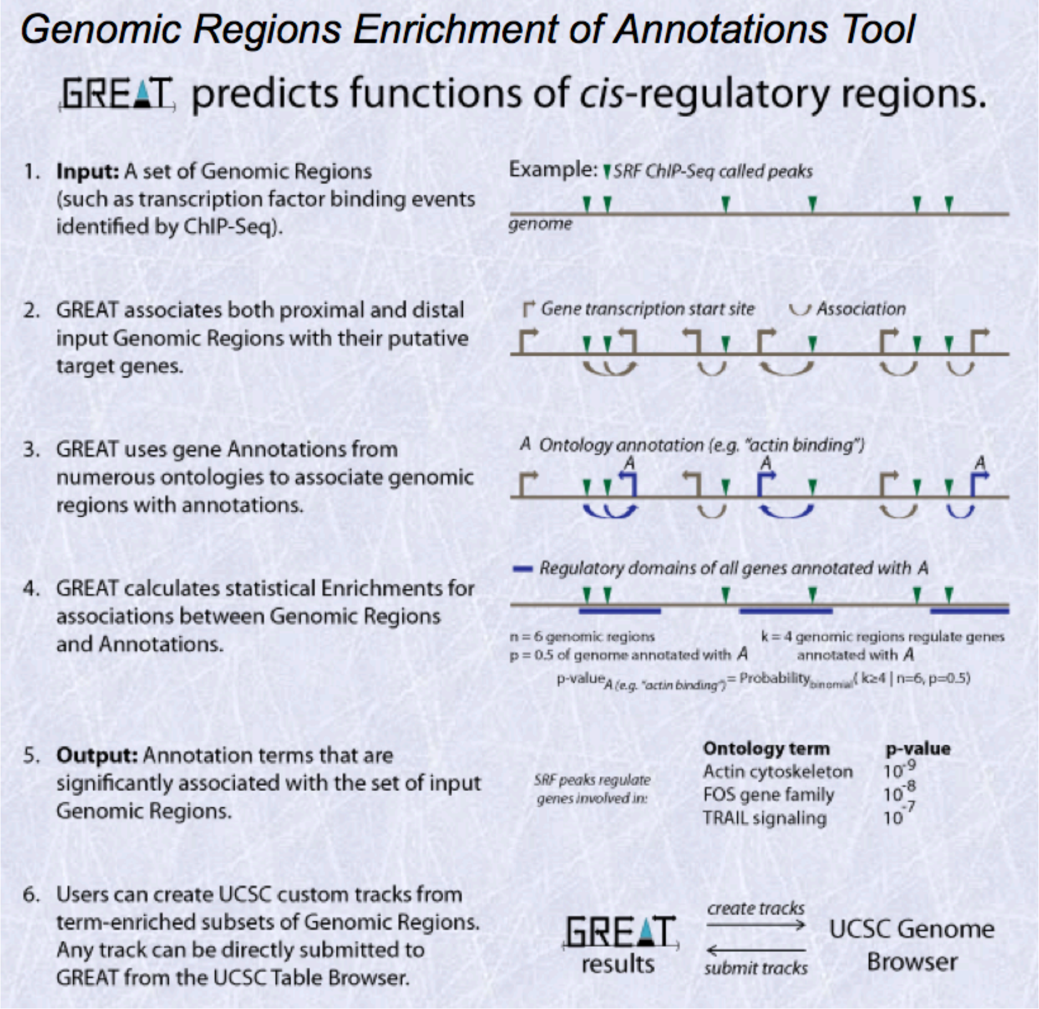
- In the context of significance tests (e.g. detecting over-represented words, or estimating the significance of BLAST matching scores), the P-value represents the probability to generate by chance (under the background model) a value at least as distant from the expectation as the one we observe.
  - Pval = P(X >= obs)

- For the analyst, this P-value indicates the risk to consider something as significant whereas it is not, i.e. the ***False Positive Risk (FPR)***.

- In the context of hypothesis testing, the concept of P-value is associated to the parameter alpha, the risk of first type error. The first type error consists in rejecting the null hypothesis H_0 whereas it is true : $P(RH_0|H_0)$. This alpha risk is estimated by testing the significance of the observed statistics (e.g. $chi2_{obs}$, t_{obs}) according to the theoretical distribution.

Application example:
GREAT - Genomic Regions Enrichment of Annotations Tool

- GREAT takes as input a set of genomic features (e.g. the peaks obtained from a ChIP-seq experiment).
- Identifies the set of genes matched by these features (genes are extended upstream and downstream to include regulatory regions).
- Assesses the enrichment of the set of target genes with each class of the Gene Ontology.
- One analysis involves several thousands of significant tests.



Genomic Regions Enrichment of Annotations Tool

GREAT predicts functions of *cis*-regulatory regions.

1. **Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).
2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.
3. GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.
4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.
5. **Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.
6. Users can create UCSC custom tracks from term-enriched subsets of Genomic Regions. Any track can be directly submitted to GREAT from the UCSC Table Browser.

Example: SRF ChIP-Seq called peaks
genome
Gene transcription start site
Association
A Ontology annotation (e.g. "actin binding")
A
A
A
Regulatory domains of all genes annotated with A
n = 6 genomic regions
p = 0.5 of genome annotated with A
k = 4 genomic regions regulate genes annotated with A
SRF peaks regulate genes involved in:
GREAT results
create tracks
submit tracks
UCSC Genome Browser

$\text{p-value}_{A\ (e.g.\ \text{"actin binding"})} = \text{Probability}_{\text{binomial}}(k \geq 4 \mid n=6, p=0.5)$

Ontology term	**p-value**
Actin cytoskeleton	10^{-9}
FOS gene family	10^{-8}
TRAIL signaling	10^{-7}

▫ http://great.stanford.edu/

Statistics

- Nomenclature
  - F number of false positives (FP)
  - T number of true positives (TP)
  - S number of tests called significant
  - m_0 number of truly null features
  - m_1 number of truly alternative features
  - m total number of features $m = m_0+m_1$
  - p threshold on p-value p = E[F / m0]
  - E[F] expected number of false positives (also called E-value) E[F] = p * m0
  - Pr(F >+ 1) family-wise error rate FWER = 1 – (1 – p)^m0
  - FDR False discovery rate FDR = E[F/S] = E[F / (F + T)]
  - Sp Specificity Sp = (m0 – F) / m0
  - Sn Sensitivity Sn = T / m1
- In practice
  - We never know the values of F, T, m0, m1, or any statistics derived from them.
  - The only observable numbers are the number of tests (*m*), and the number of these declared significant (*S*) or not (*m-S*).
  - Some strategies have however been proposed to **estimate** *m0* and *m1* (see Storey and Tibshirani, 2003).

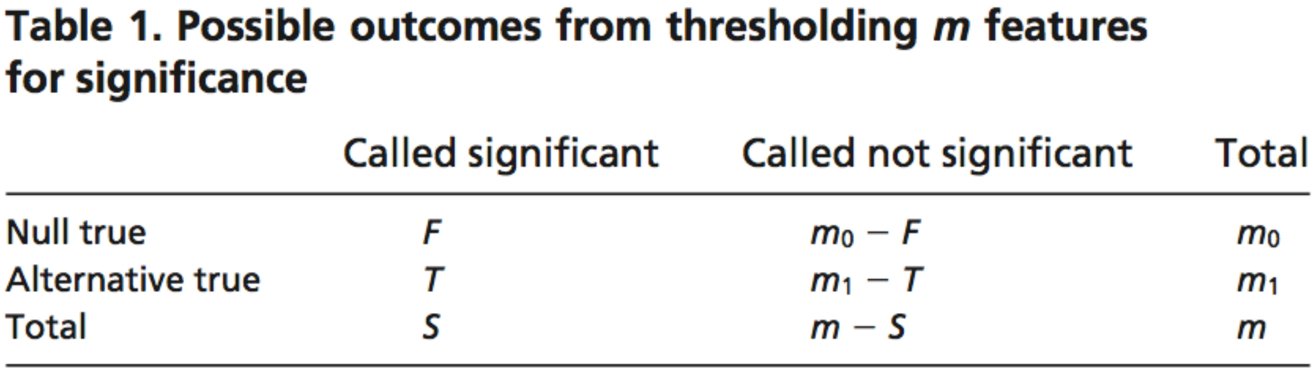


Table 1. Possible outcomes from thresholding *m* features for significance

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

- Storey and Tibshirani. Statistical significance for genomewide studies. Proc Natl Acad Sci USA (2003) vol. 100 (16) pp. 9440-5

Validation statistics

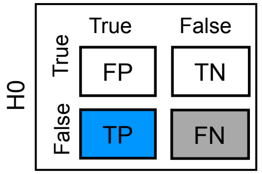
- Various statistics can be derived from the 4 elements of a contingency table.

		Declared significant	
		True	False
H0	True	FP	TN
	False	TP	FN

Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**Sensitivity**	**TP/(TP + FN)**
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - -TP*log(TP)-FP*log(FP)-FN*log(FN)-TN*log(TN)+(TP+FP)*log(TP+FP)+(FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	Geometric accuracy	sqrt(Sn*PPV)
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)

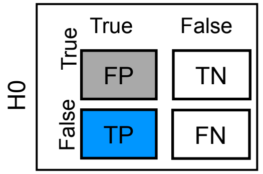
Sn = TP/(TP+FN)

Declared significant



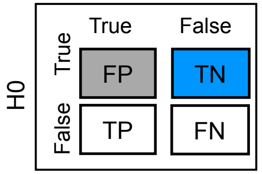
PPV=TP/(TP+FP)

Declared significant



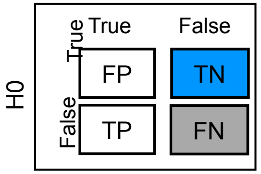
Sp=TN/(FP+TN)

Declared significant



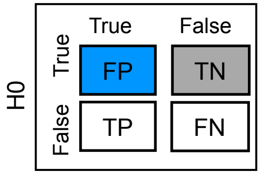
NPV=TN/(FN+TN)

Declared significant



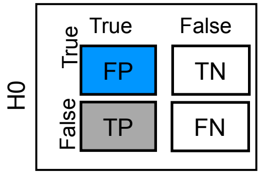
FPR=FP/(FP+TN)

Declared significant



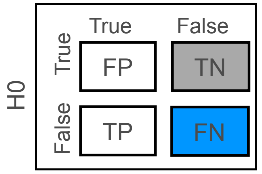
FDR=FP/(FP+TP)

Declared significant



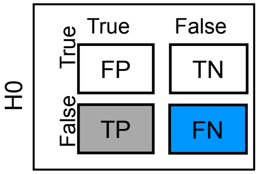
FN/(FN+TN)

Declared significant



FNR=FN/(TP+FN)

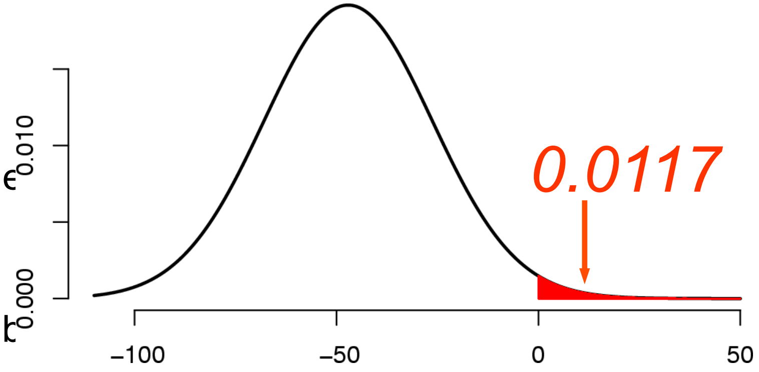
Declared significant



Multi-testing corrections

The problem of multiple testing

- Let us assume that the score of the alignment between two sequences has a p-value
  - P(X > 0) = 0.0117
- What would happen if we consider this score as significant, while scanning a database that contains 200,000 sequences ?
- Let N be the number of tests
  - The risk of error (P-value) associated to each gene will thus b challenged N times.
  - The significance thresholds generally used for single testing (alpha = 0.01, 0.001) are thus likely to return many false positive.
- The situation of multiple testing is very frequent in bioinformatics
  - Assessing the significance of each gene on a chip represents thousands of simultaneous tests.
  - Genome-wide association studies (GWAS) are now routinely performed with SNP chips containing 600.000 SNPs.
  - Sequence similarity searches (e.g. BLAST a sequence against all known proteins) amount to compare a query sequences to billions of database entries.



Multiple testing correction : Bonferroni's rule

- A first approach to correct for multiple tests is to apply Bonferroni's rule
  - Adapt the p-value threshold ("alpha risk") to the number of simultaneous tests.

$$\boxed{\alpha \leq \frac{1}{N}}$$

Multiple testing correction : from P-value to E-value

$$\boxed{Eval = N \cdot Pval}$$

- If p=P(X >0)=0.0117 and the database contains N=200,000 entries, we expect to obtain $N*p = 2340$ false positives !
- We are in a situation of multi-testing : each analysis amounts to test N hypotheses.
- The E-value (expected value) allows to take this effect into account :
  - ***Eval = Pval * N***
  - Instead of setting a threshold on the P-value, we should set a threshold on the E-value.
  - If we want to avoid false positive, this threshold should always be smaller than 1.
    - ***Threshold(Eval) ≤ 1***
- The fact to set a threshold ***≤ 1*** on the E-value is equivalent to Bonferroni's correction, which consists in adapting the threshold on the p-value.
    - ***Threshold(Pval) ≤ 1/N***

Multiple testing correction : Family-wise Error Rate (FWER)

- Another correction for multiple testing consists in estimating the Family-Wise Error Rate (FWER).
- The FWER is the probability to observe at least one false positive in the whole set of tests. This probability can be calculated quite easily from the P-value (*Pval*).

$$FWER = 1 - (1 - Pval)^N$$

False Discovery Rate (FDR)

- Yet another approach is to consider, for a given threshold on P-value, the False Discovery Rate (FDR), i.e. the proportion of false predictions within a set of tests declared significant.
  - FP number of false positives
  - TP number of true positives

$$FDR = FP/(FP + TP)$$

Summary - Multi-testing corrections

$$\boxed{\alpha_{Bonf} \leq \frac{1}{N}}$$

- Bonferroni rule adapt significance threshold

$$\boxed{Eval = N \cdot Pval}$$

- E-value expected number of false positives

$$\boxed{FWER = 1 - (1 - Pval)^{N}}$$

- FWER Family-wise error rate:
  probability to observe
  at least one false positive

$$\boxed{FDR = FP/(FP + TP)}$$

- FDR False discovery rate:
  estimated rate of false positives
  among the predictions

The “q-value” (Storey and Tibshirani, 2003)

How can we estimate the m_0 / m_1 proportions ?

- TO BE COMPLETED
- See the practical about multiple testing correction on the supporting Web site.
  - http://pedagogix-tagc.univ-mrs.fr/courses/statistics_bioinformatics/

Fig 1 from Storey and Tibshirani, 2003)

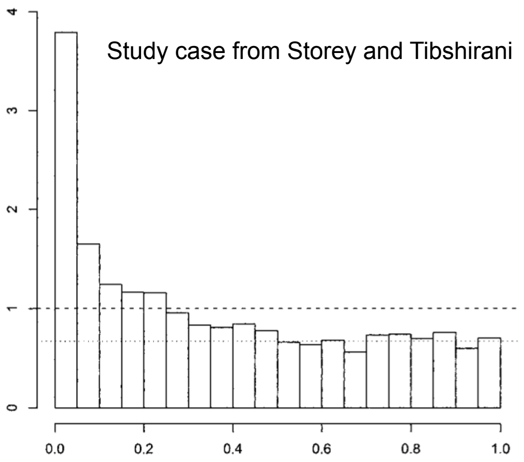
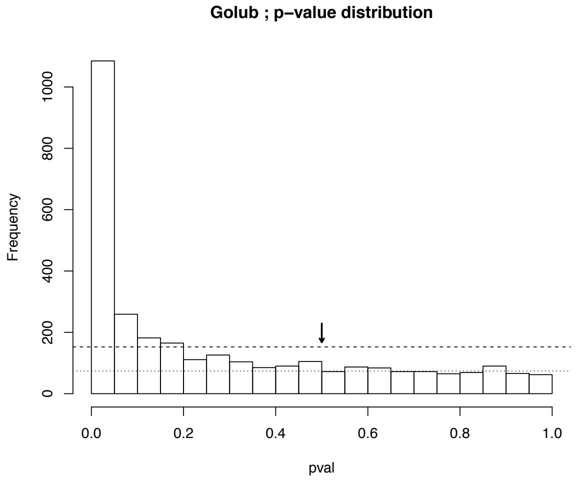
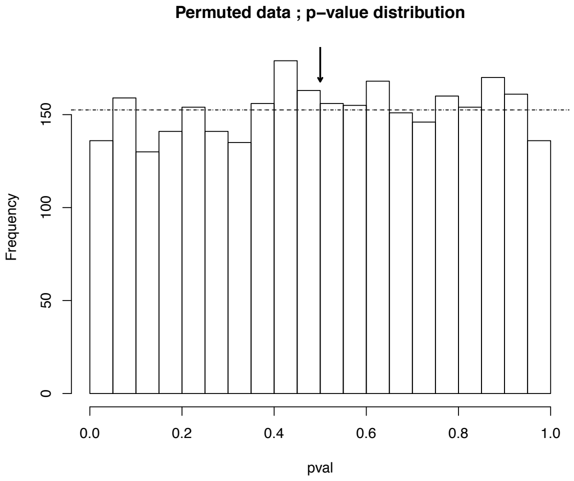


Fig. 1. A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.

Application to another study case (ALL versus AML expression from Goub et al., 1999)



Negative control: permuted data from Golub et al. (1999)



- Storey and Tibshirani. Statistical significance for genomewide studies. Proc Natl Acad Sci USA (2003) vol. 100 (16) pp. 9440-5.

Statistical Analysis of Microarray Data

Clustering

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.univ-amu.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)

Contents

- Data sets
- Distance and similarity metrics
- K-means clustering
- Hierarchical clustering
- Evaluation of clustering results

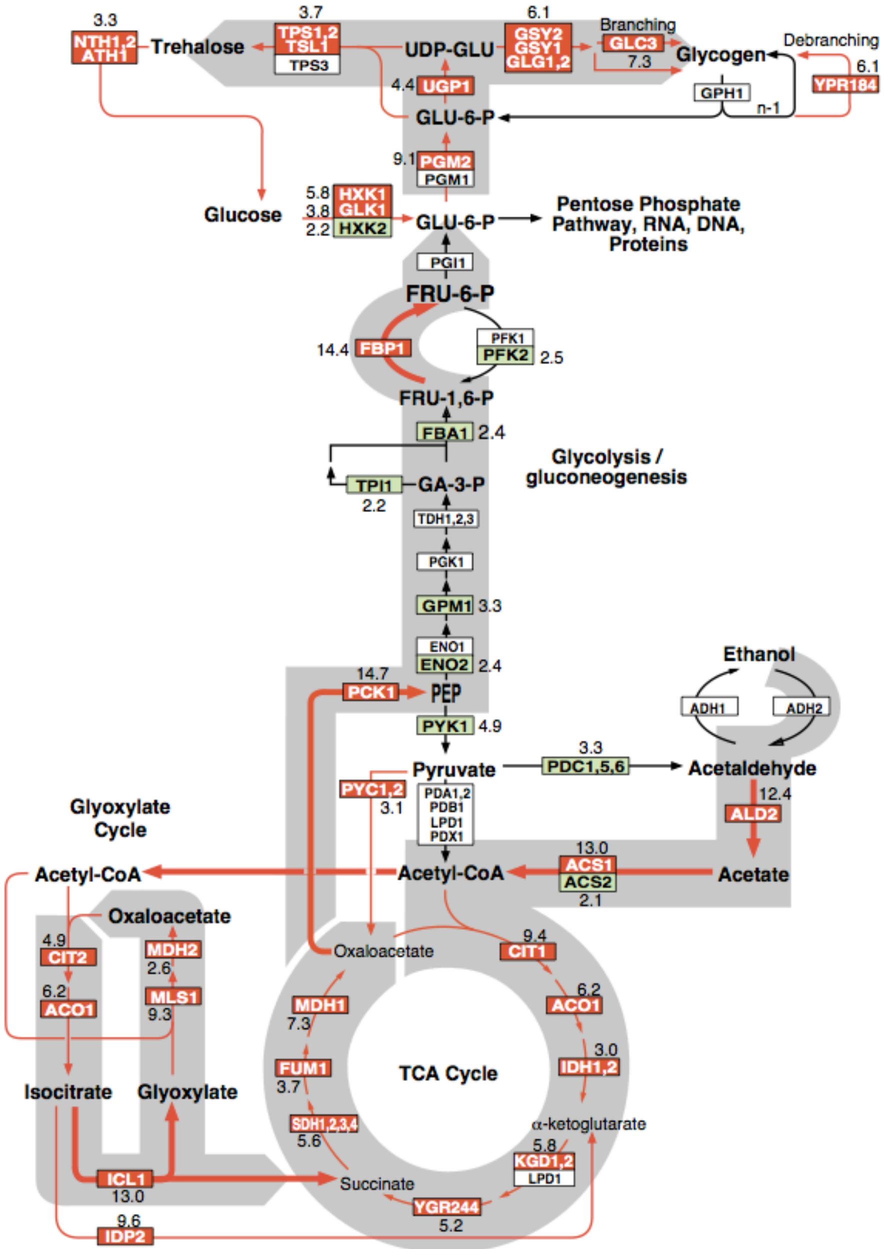
Introduction to clustering

- 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, ...)

Data sets

Diauxic shift

- DeRisi et al published the first article describing a full-genome monitoring of gene expression data.
- This article reported an experiment called “diauxic shift” with with 7 time points.
- Initially, cells are grown in a glucose-rich medium.
- As time progresses, cells
  - Consume glucose -> when glucose becomes limiting
    - Glycolysis stops
    - Gluconeogenesis is activated to produce glucose
  - Produce by-products -> the culture medium becomes polluted/
    - Stress response



- DeRisi et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science (1997) vol. 278 (5338) pp. 680-6

Cell cycle data

- Spellman et al. (1998)
- Time profiles of yeast cells followed during cell cycle.
- Several experiments were regrouped, with various ways of synchronization (elutriation, cdc mutants, …)
- ~800 genes showing a periodic patterns of expression were selected (by Fourier analysis)

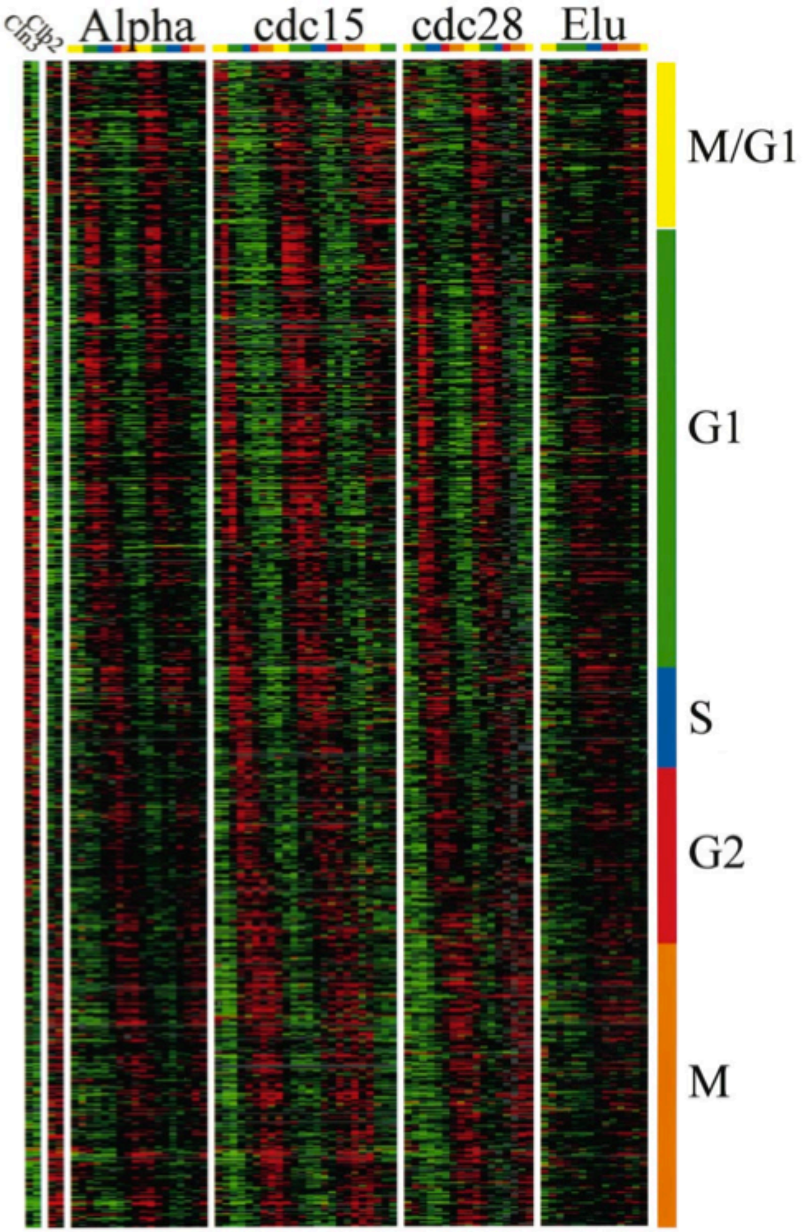
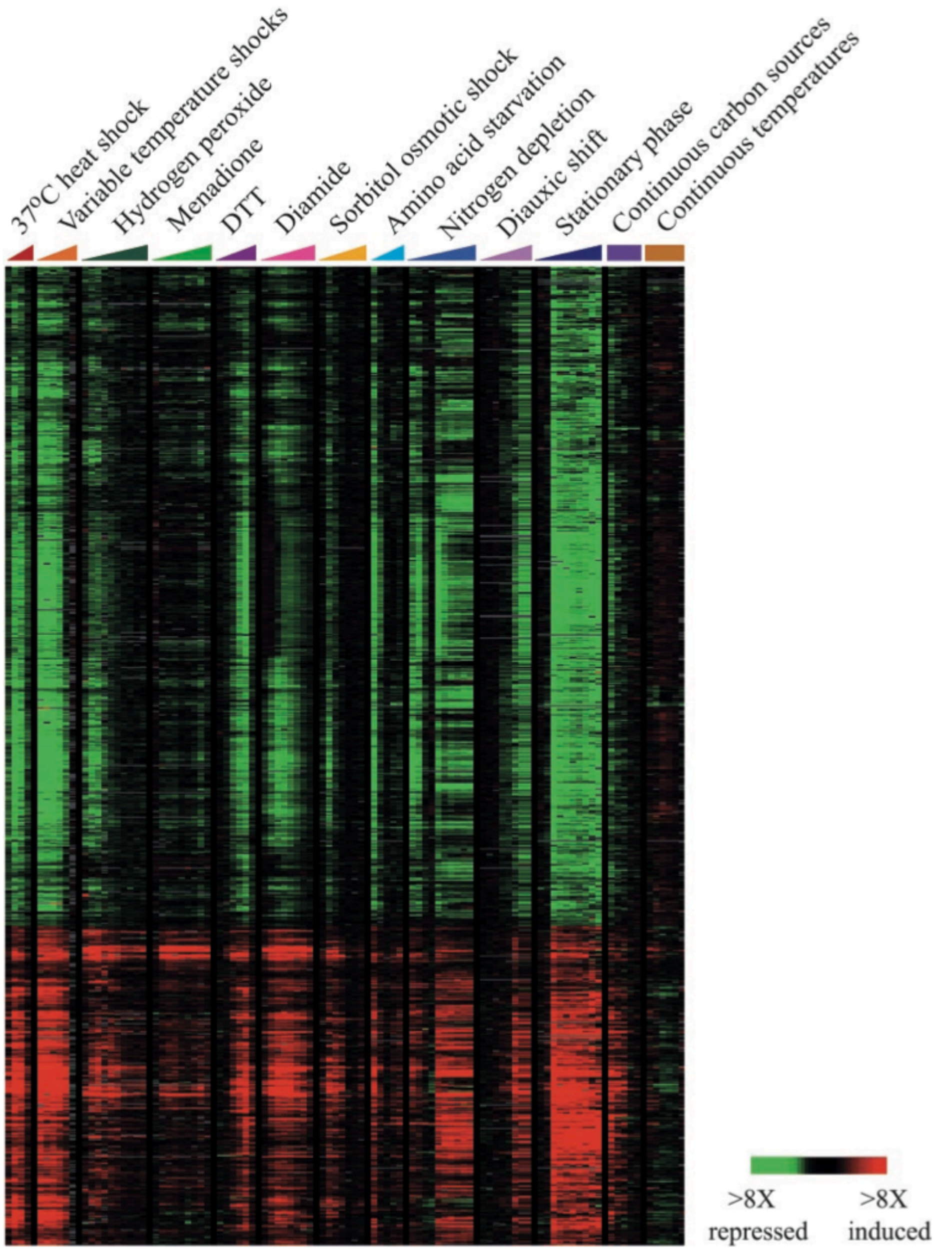


Figure 1.

- Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell (1998) vol. 9 (12) pp. 3273-97

Gene expression data – response to environmental changes

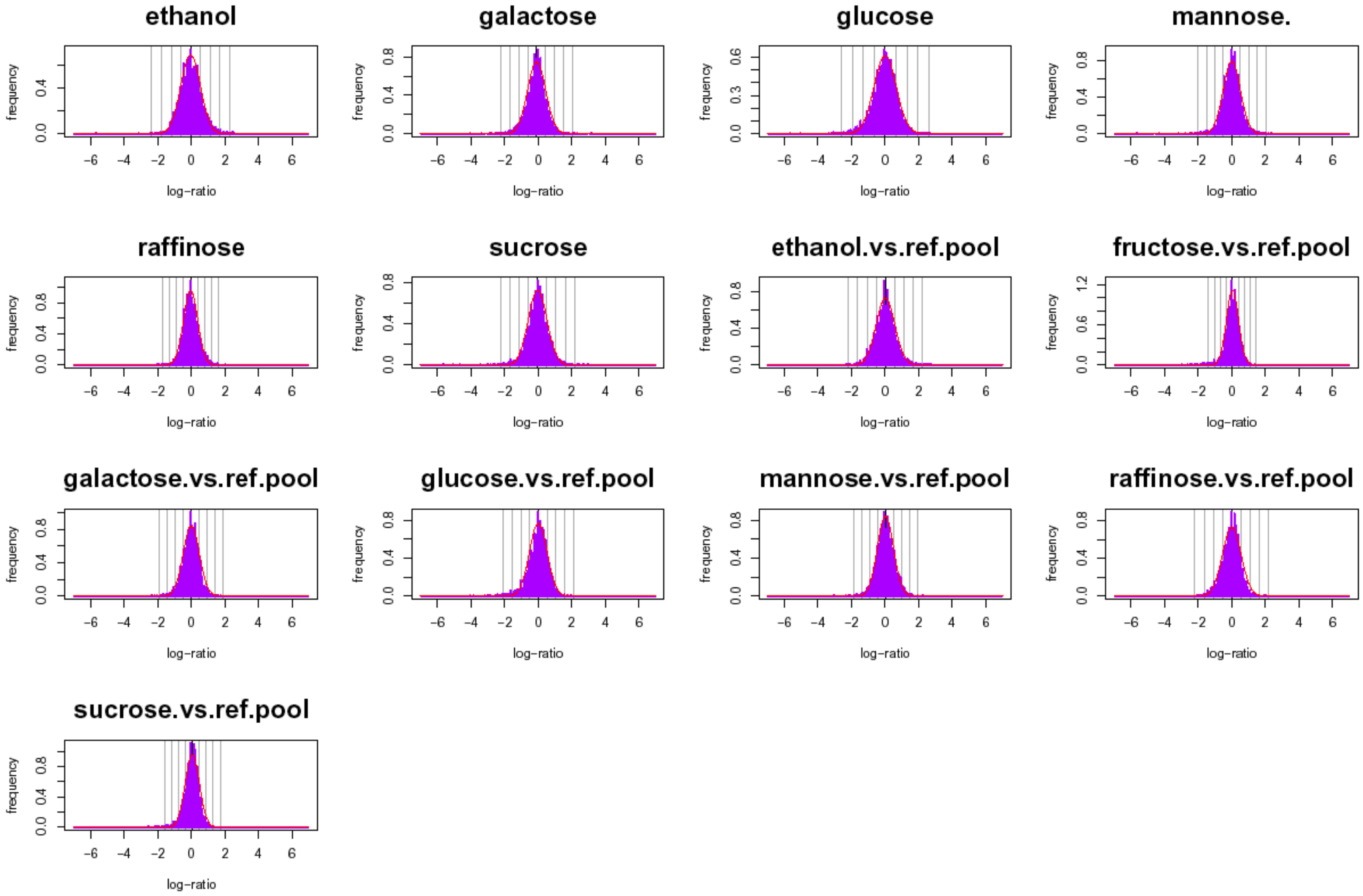
- Gasch et al. (2000), 173 chips (stress response, heat shock, drugs, carbon source, …)



- Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell (2000) vol. 11 (12) pp. 4241-57.

Gene expression data - carbon sources

- Gasch et al. (2000), 173 chips (stress response, heat shock, drugs, carbon source, …)
- We selected the 13 chips with the response to different carbon sources.

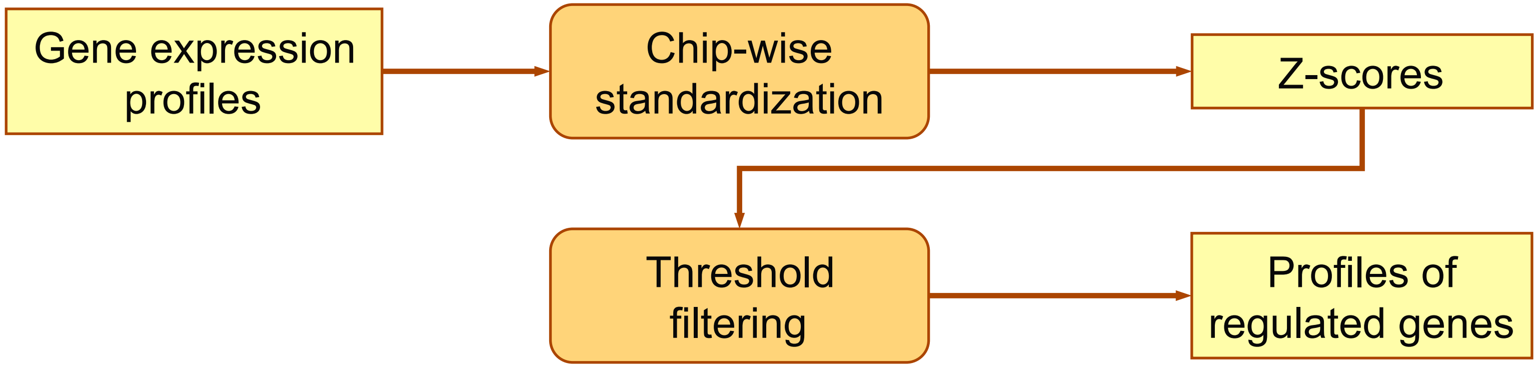


- Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell (2000) vol. 11 (12) pp. 4241-57.

Data standardization and filtering

- For the cell cycle experiments, genes had already been filtered in the original publication. We used the 800 selected genes for the analysis.
- For the diauxic shift and carbon source experiments, each chip contain >6000 genes, most of which are un-regulated.
- Standardization
  - We applied a chip-wise standardization (centring and scaling) with robust estimates (median and IQR) on each chip.
- Filtering
  - Z-scores obtained after standardization were converted
    - to P-value (normal distribution)
    - to E-value (=P-value*N)
  - Only genes with an E-value < 1 were retained for clustering.

Filtering of carbon source data

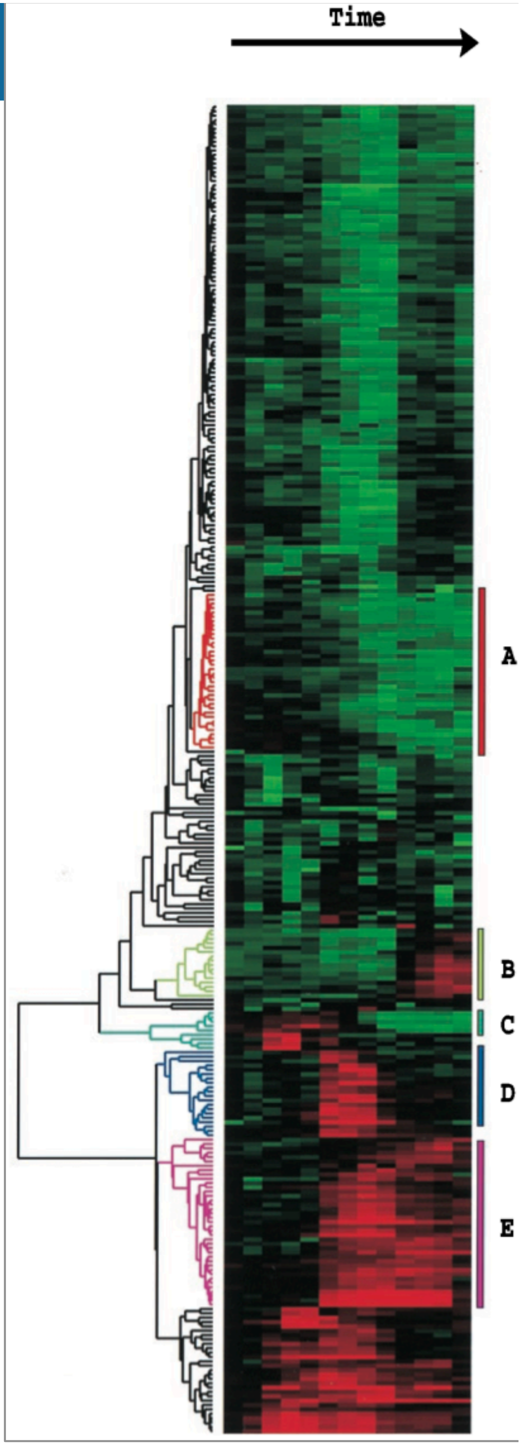


	Carbon sources												
ORF	ethanol	galactose	glucose	mannose.	raffinose	sucrose	ethanol.vs.ref.pool	fructose.vs.ref.pool	galactose.vs.ref.pool	glucose.vs.ref.pool	mannose.vs.ref.pool	raffinose.vs.ref.pool	sucrose.vs.ref.pool
YAL066W	0.71	-1.87	-1.15	-4.90	-1.85	-3.81	-3.34	-0.96	-3.41	-1.04	0.36	-1.55	-0.87
YAR008W	-2.70	0.36	0.03	-4.90	-0.94	-0.53	0.64	-0.53	-2.57	-0.73	0.38	-1.75	-0.55
YAR071W	-5.43	-1.22	2.73	-0.44	-0.24	3.24	-6.69	1.10	-5.21	1.39	-0.70	0.22	2.94
YBL005W	1.40	3.05	3.97	4.92	1.18	5.52	-0.53	0.79	-0.84	-1.00	1.12	-2.26	1.23
YBL015W	4.00	0.28	-3.46	-3.65	-2.38	-4.94	3.26	-4.64	0.59	-3.76	-1.62	1.08	-5.37
YBL043W	3.91	-1.16	-4.89	-4.90	-1.61	-4.76	4.47	-6.97	-0.61	-6.67	-7.12	0.78	-9.73
YBR018C	-9.68	5.53	-8.66	-11.19	-13.49	-10.23	-9.81	-15.15	6.32	-10.89	-13.01	-12.10	-13.73
YBR019C	-9.68	6.16	-7.77	-11.19	-12.09	-9.17	-9.42	-12.93	6.07	-10.58	-10.90	-9.08	-11.97
YBR020W	-9.68	6.05	-8.66	-11.19	-13.49	-10.23	-10.04	-12.70	6.83	-12.82	-13.01	-8.95	-14.94
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Hierarchical clustering

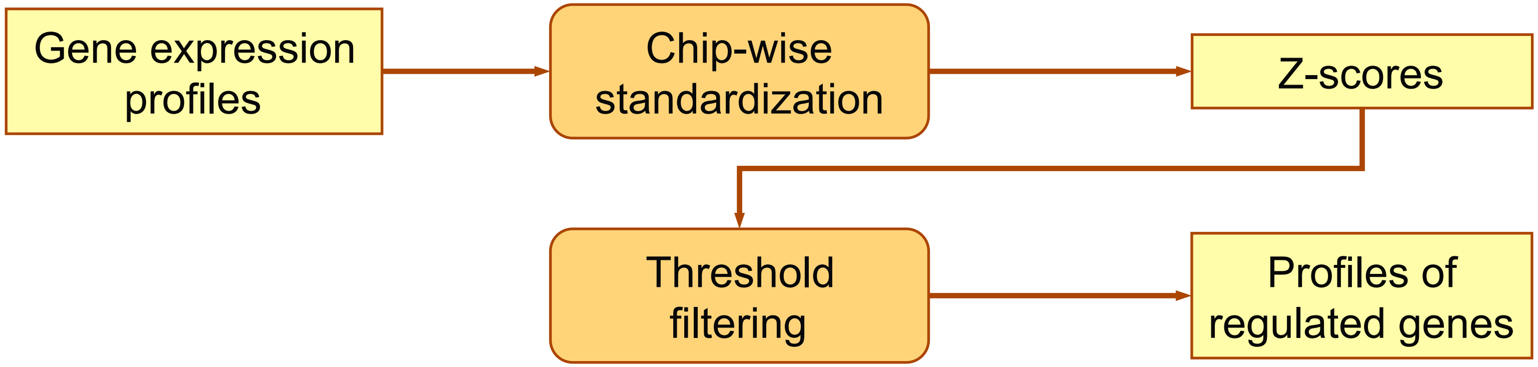
Hierarchical clustering of expression profiles

- 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

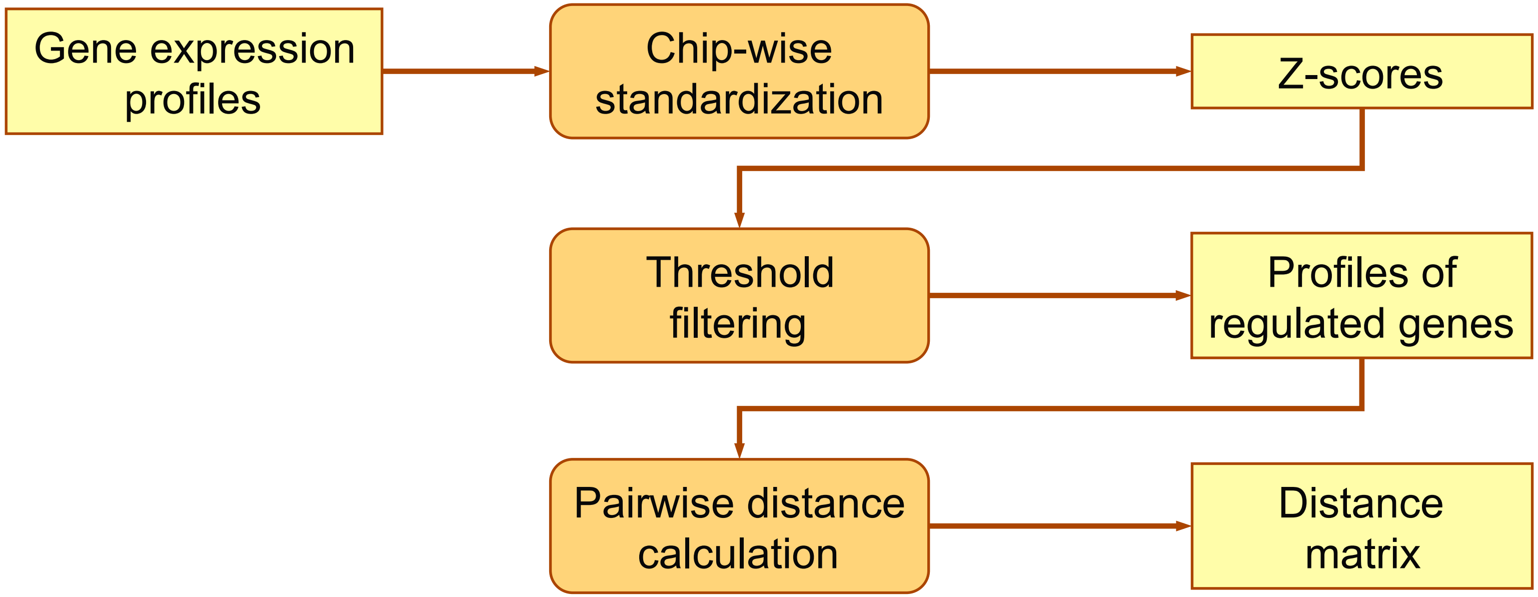
Clustering with gene expression data



Carbon sources

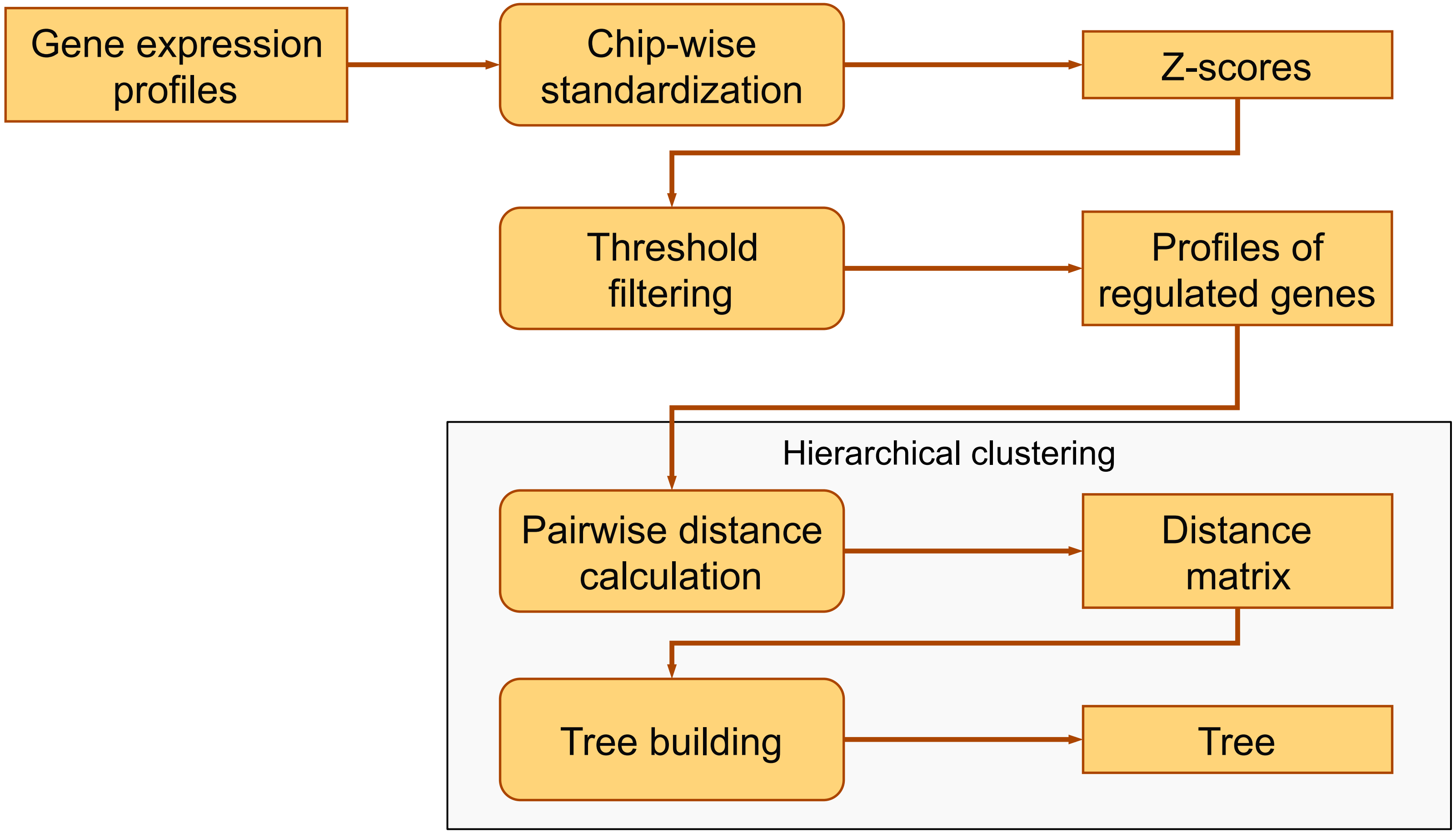
ORF	ethanol	galactose	glucose	mannose.	raffinose	sucrose	ethanol.vs.ref.pool	fructose.vs.ref.pool	galactose.vs.ref.pool	glucose.vs.ref.pool	mannose.vs.ref.pool	raffinose.vs.ref.pool	sucrose.vs.ref.pool
YAL066W	0.71	-1.87	-1.15	-4.90	-1.85	-3.81	-3.34	-0.96	-3.41	-1.04	0.36	-1.55	-0.87
YAR008W	-2.70	0.36	0.03	-4.90	-0.94	-0.53	0.64	-0.53	-2.57	-0.73	0.38	-1.75	-0.55
YAR071W	-5.43	-1.22	2.73	-0.44	-0.24	3.24	-6.69	1.10	-5.21	1.39	-0.70	0.22	2.94
YBL005W	1.40	3.05	3.97	4.92	1.18	5.52	-0.53	0.79	-0.84	-1.00	1.12	-2.26	1.23
YBL015W	4.00	0.28	-3.46	-3.65	-2.38	-4.94	3.26	-4.64	0.59	-3.76	-1.62	1.08	-5.37
YBL043W	3.91	-1.16	-4.89	-4.90	-1.61	-4.76	4.47	-6.97	-0.61	-6.67	-7.12	0.78	-9.73
YBR018C	-9.68	5.53	-8.66	-11.19	-13.49	-10.23	-9.81	-15.15	6.32	-10.89	-13.01	-12.10	-13.73
YBR019C	-9.68	6.16	-7.77	-11.19	-12.09	-9.17	-9.42	-12.93	6.07	-10.58	-10.90	-9.08	-11.97
YBR020W	-9.68	6.05	-8.66	-11.19	-13.49	-10.23	-10.04	-12.70	6.83	-12.82	-13.01	-8.95	-14.94
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Hierarchical clustering on gene expression data


Gene expression profiles
Chip-wise standardization
Z-scores
Threshold filtering
Profiles of regulated genes
Pairwise distance calculation
Distance matrix

	YAL066W	YAR008W	YAR071W	YBL005W	YBL015W	YBL043W	YBR018C	YBR019C	YBR020W	YBR054W	…
YAL066W	0.00	6.82	12.99	16.33	11.64	17.39	36.41	32.52	36.07	12.00	…
YAR008W	6.82	0.00	11.70	13.69	12.58	18.18	37.51	33.46	37.18	12.36	…
YAR071W	12.99	11.70	0.00	13.32	21.77	26.62	42.48	38.48	42.15	21.09	…
YBL005W	16.33	13.69	13.32	0.00	19.52	25.04	44.95	41.16	44.62	17.86	…
YBL015W	11.64	12.58	21.77	19.52	0.00	8.51	34.47	30.79	33.77	6.46	…
YBL043W	17.39	18.18	26.62	25.04	8.51	0.00	31.74	28.64	30.90	11.13	…
YBR018C	36.41	37.51	42.48	44.95	34.47	31.74	0.00	5.12	4.66	35.84	…
YBR019C	32.52	33.46	38.48	41.16	30.79	28.64	5.12	0.00	4.81	32.58	…
YBR020W	36.07	37.18	42.15	44.62	33.77	30.90	4.66	4.81	0.00	35.63	…
YBR054W	12.00	12.36	21.09	17.86	6.46	11.13	35.84	32.58	35.63	0.00	…
…	…	…	…	…	…	…	…	…	…	…	…

Hierarchical clustering on gene expression data



Gene expression profiles

Chip-wise standardization

Z-scores

Threshold filtering

Profiles of regulated genes

Hierarchical clustering

Pairwise distance calculation

Distance matrix

Tree building

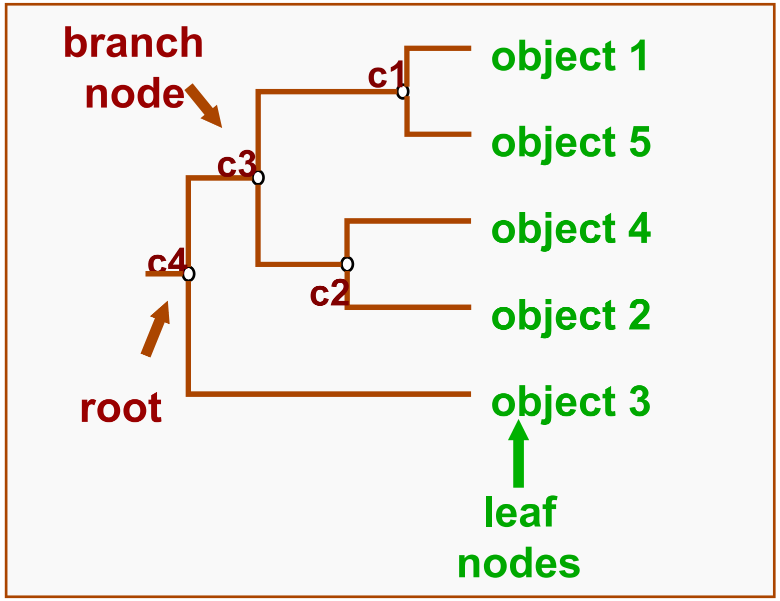
Tree

Principle of tree building

Distance matrix

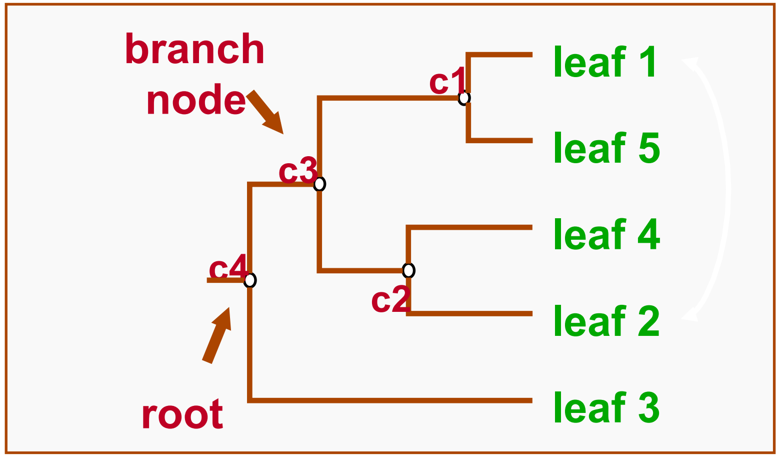
	object 1	object 2	object 3	object 4	object 5
object 1	0.00	4.00	6.00	3.50	1.00
object 2	4.00	0.00	6.00	2.00	4.50
object 3	6.00	6.00	0.00	5.50	6.50
object 4	3.50	2.00	5.50	0.00	4.00
object 5	1.00	4.50	6.50	4.00	0.00

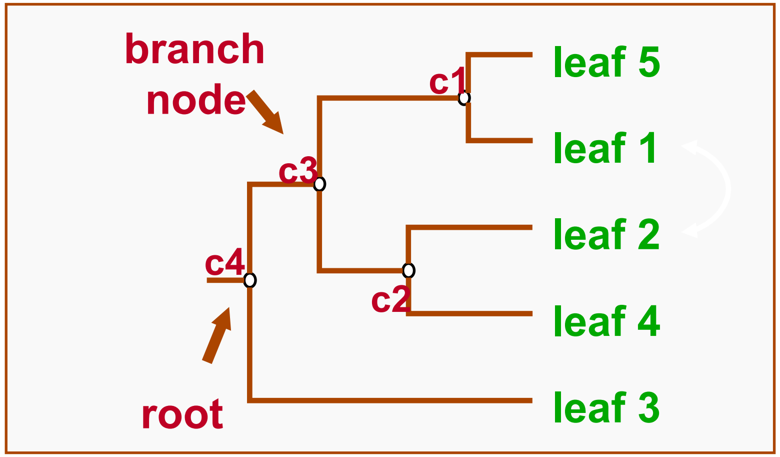
Tree representation



- 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

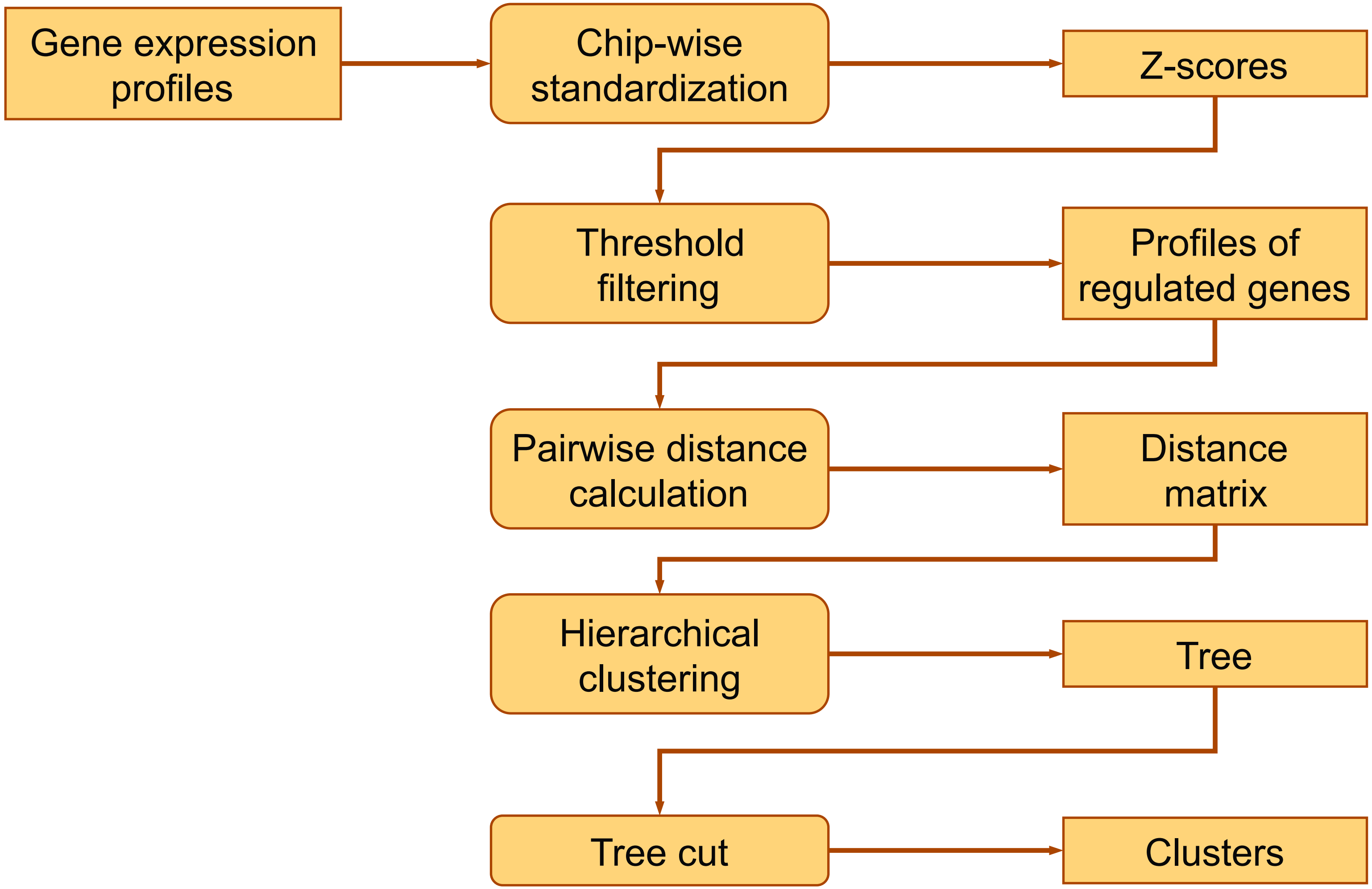
Isomorphism on a tree





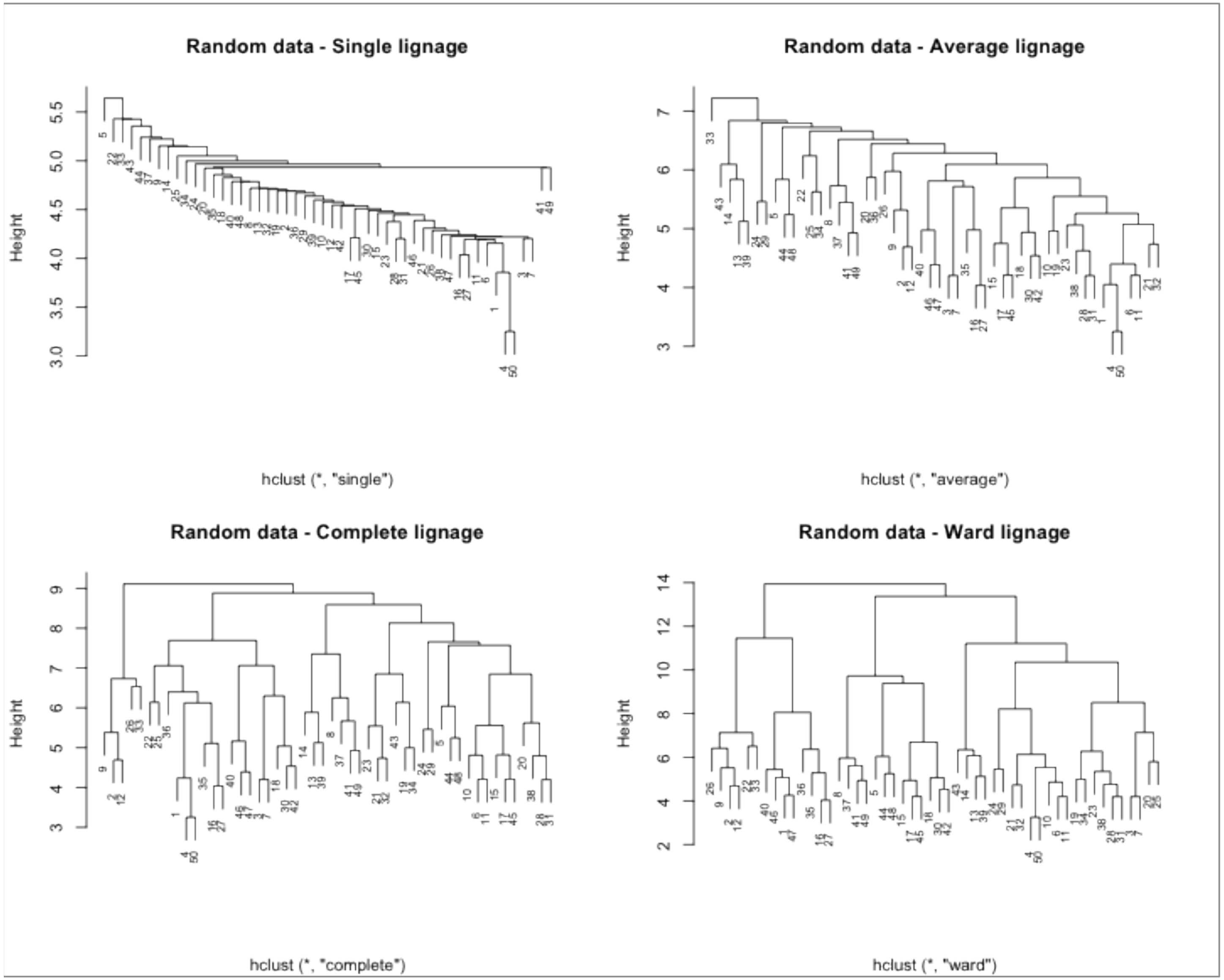
- In a tree, the two children of any branch node can be swapped. The result is an ***isomorphic tree***, considered as equivalent to the intial one.
- The two trees shown here are equivalent, however
  - Top tree: leaf 1 is far away from leaf 2
  - Bottom tree: leaf 1 is neighbour from leaf 2
- The vertical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

Hierarchical clustering on gene expression data



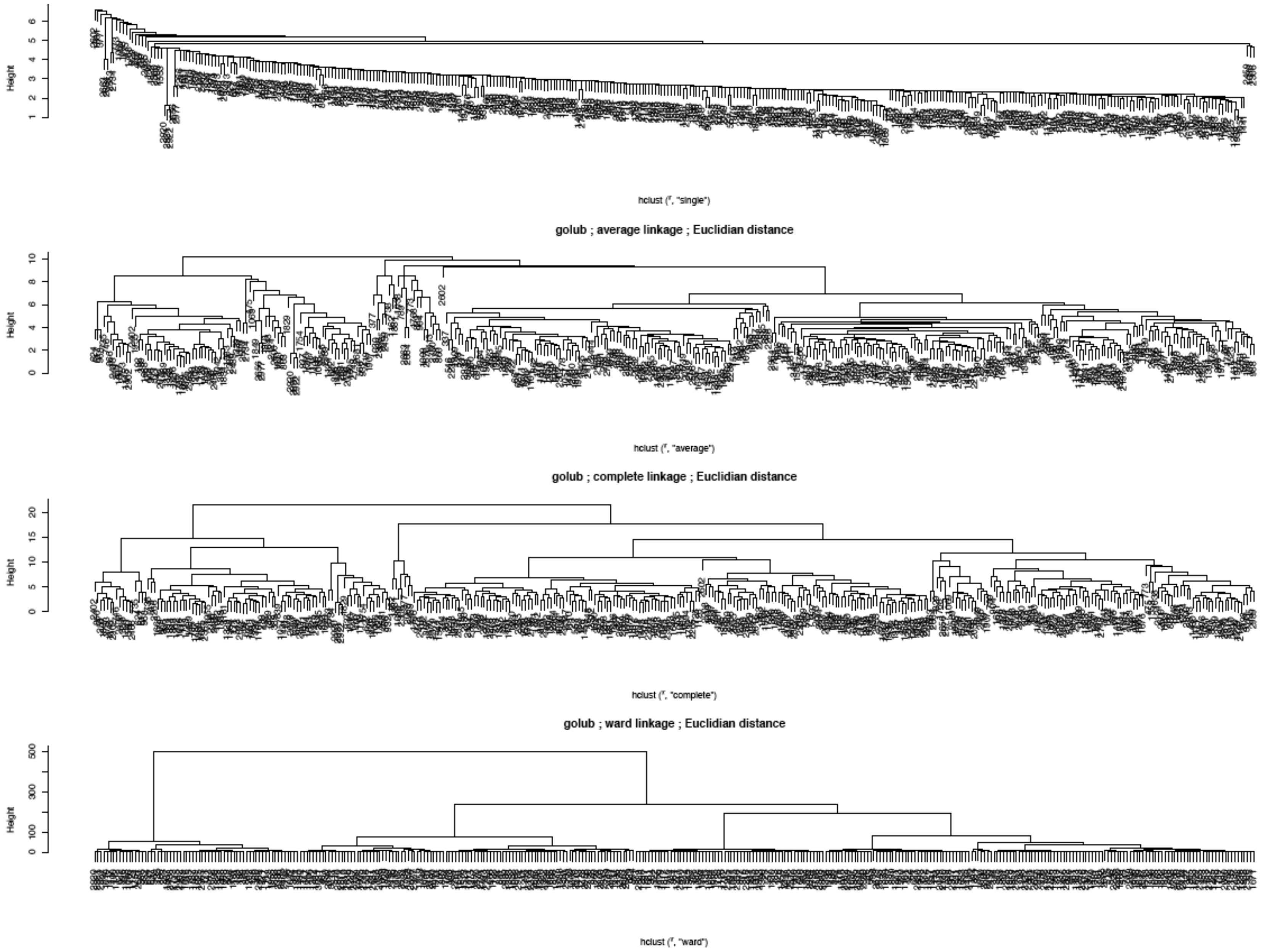
Impact of the agglomeration rule

- 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)

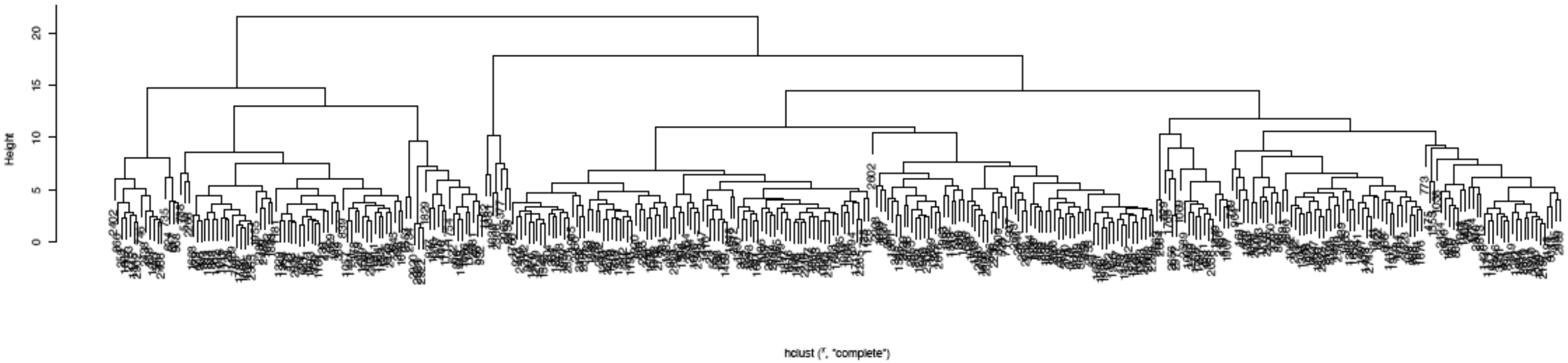


golub ; average linkage ; Euclidian distance

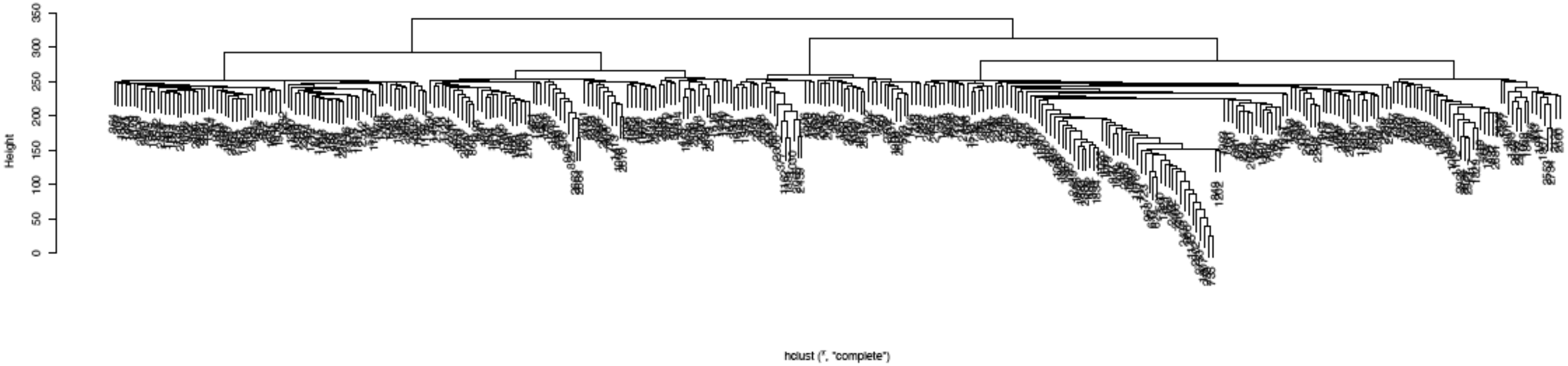
golub ; complete linkage ; Euclidian distance

golub ; ward linkage ; Euclidian distance

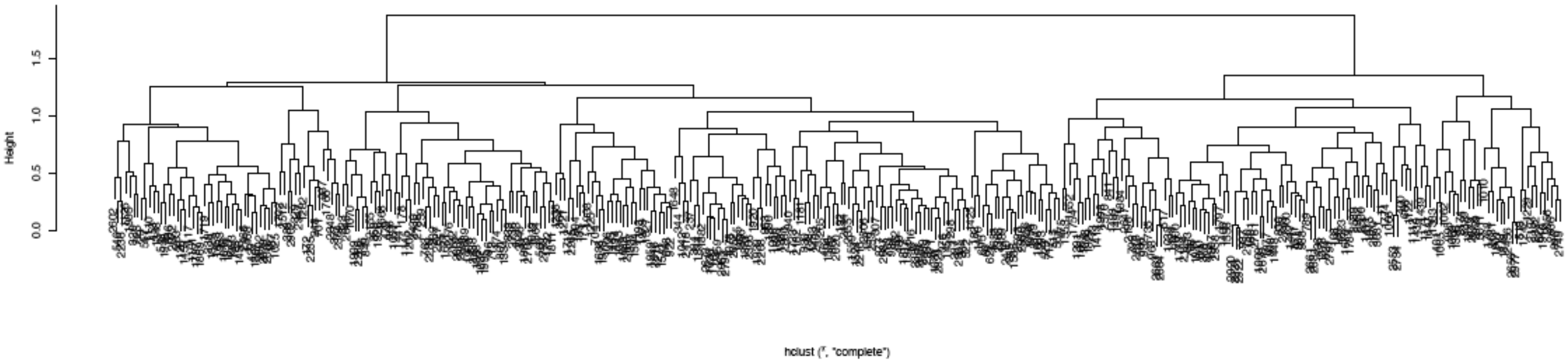
Golub 1999 - Effect of the distance metrics (complete linkage for all the trees)



golub ; complete linkage ; Dot product

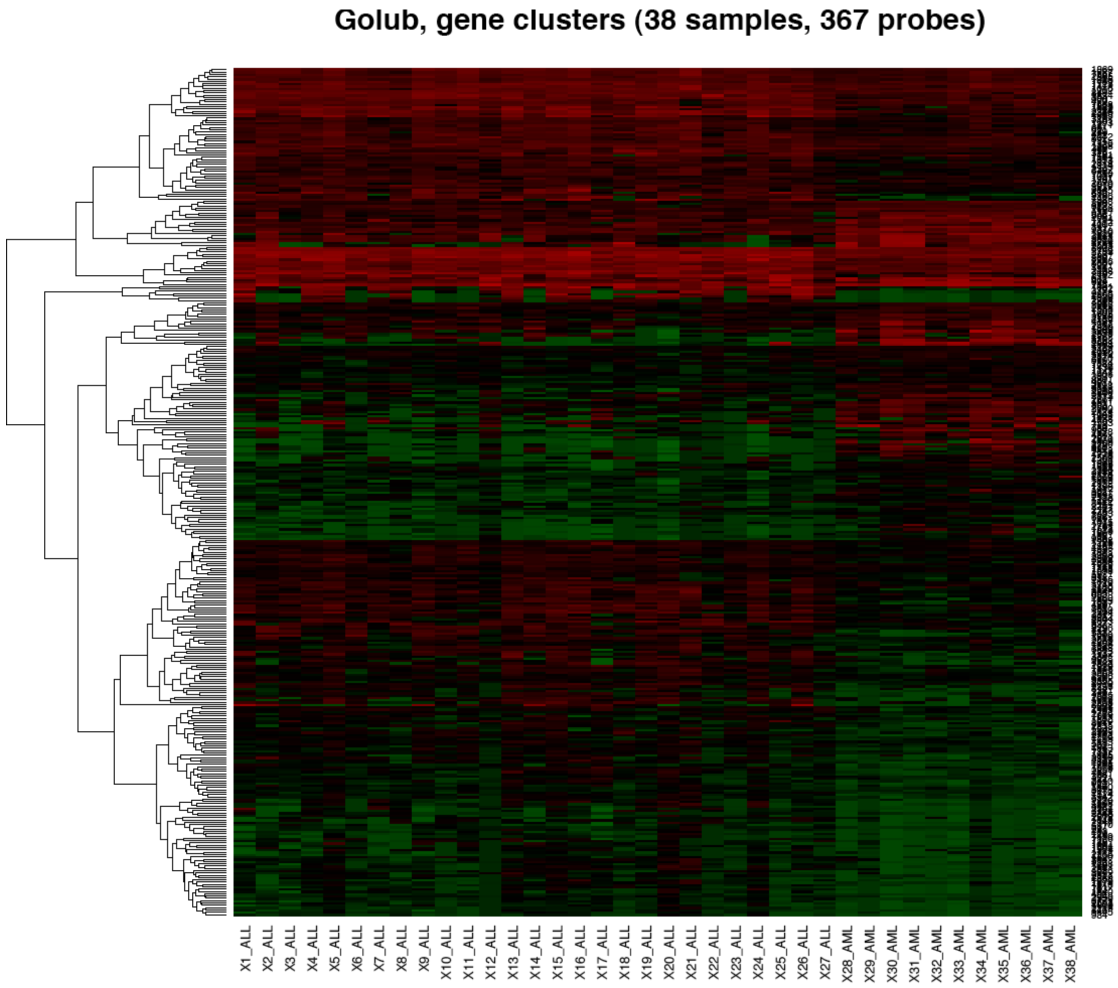


golub ; complete linkage ; Correlation

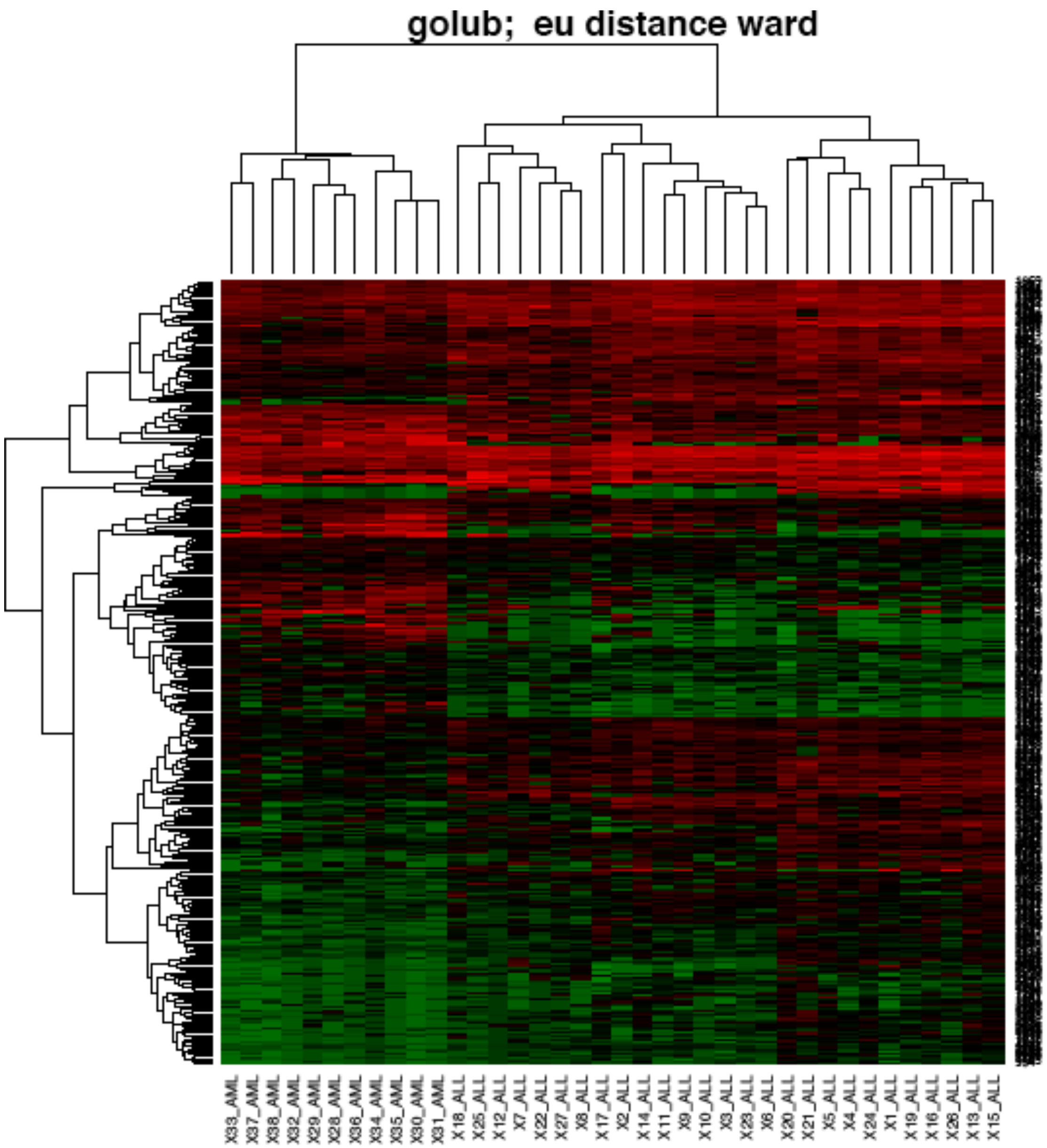


Golub 1999 - Gene clustering

- Gene clustering highlights groups of genes with similar expression profiles.

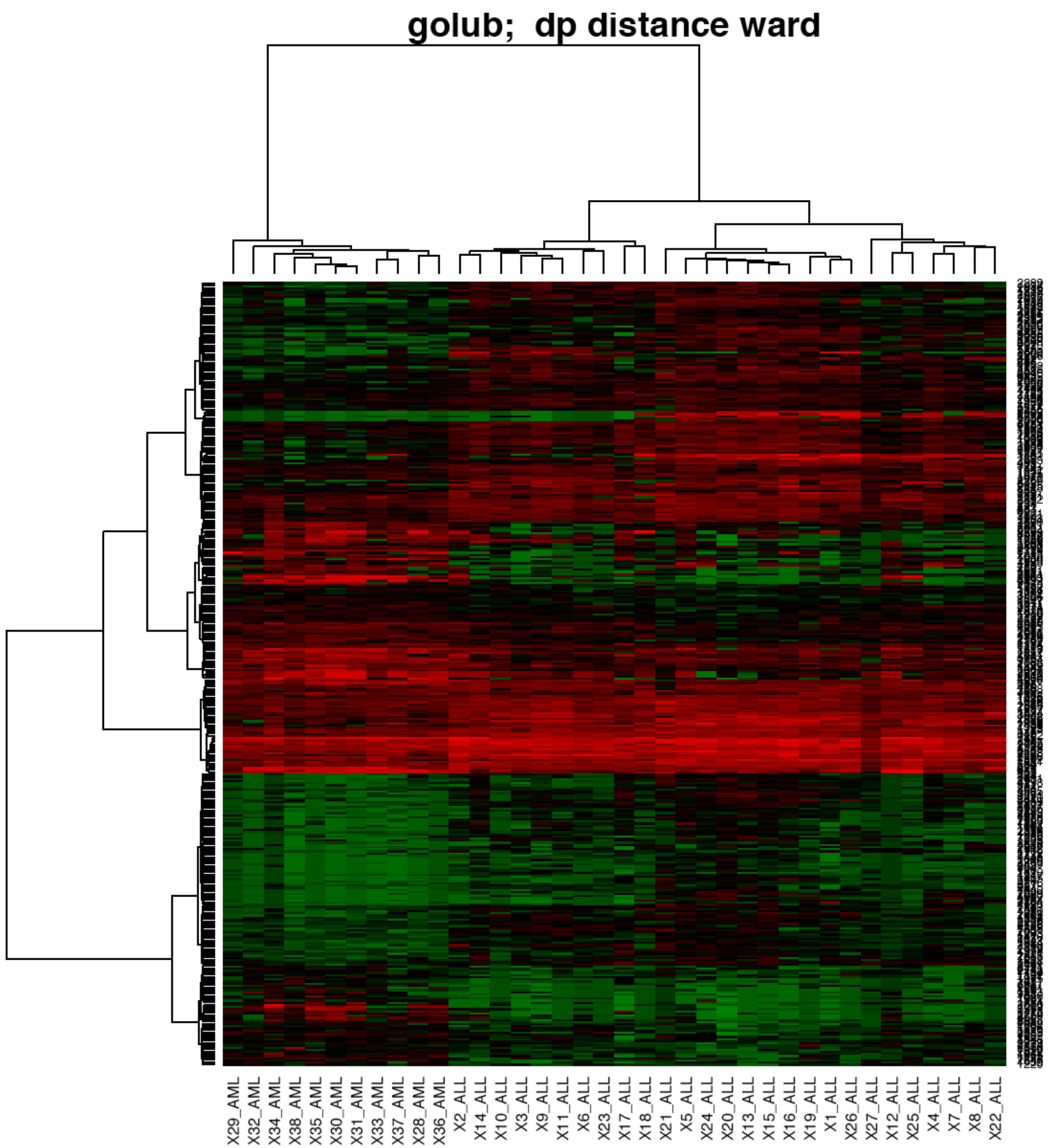


Golub 1999 - Ward Biclustering - Euclidian distance



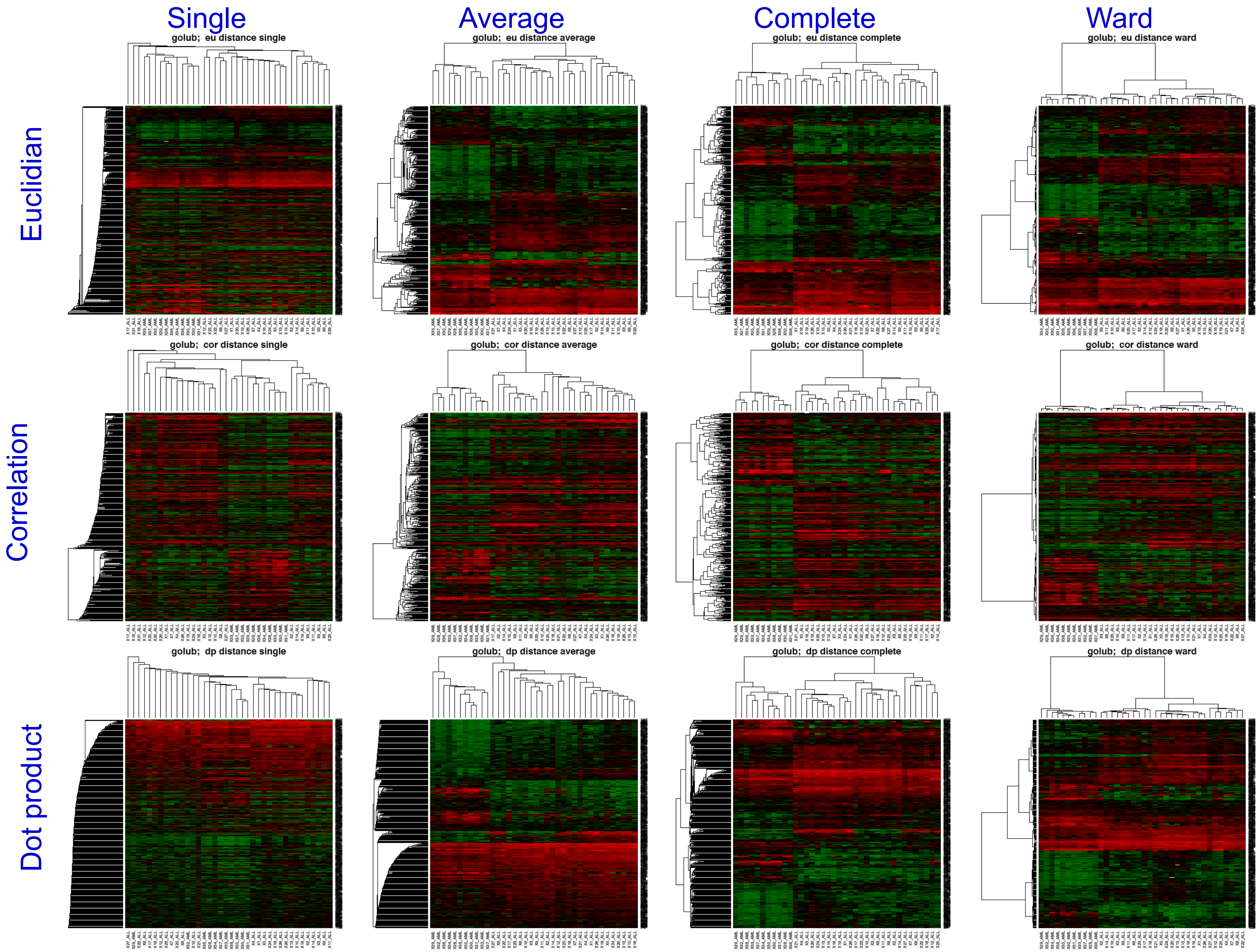
- Biclustering consists in clustering the rows (genes) and the columns (samples) of the data set.
- This reveals some subgroups of samples.
- With the golub 1999 data set
  - The AML and ALL patients are clearly separated at the top level of the tree
  - There are apparently two clusters among the AML samples.

Golub 1999 - Ward Biclustering - Dot product distance

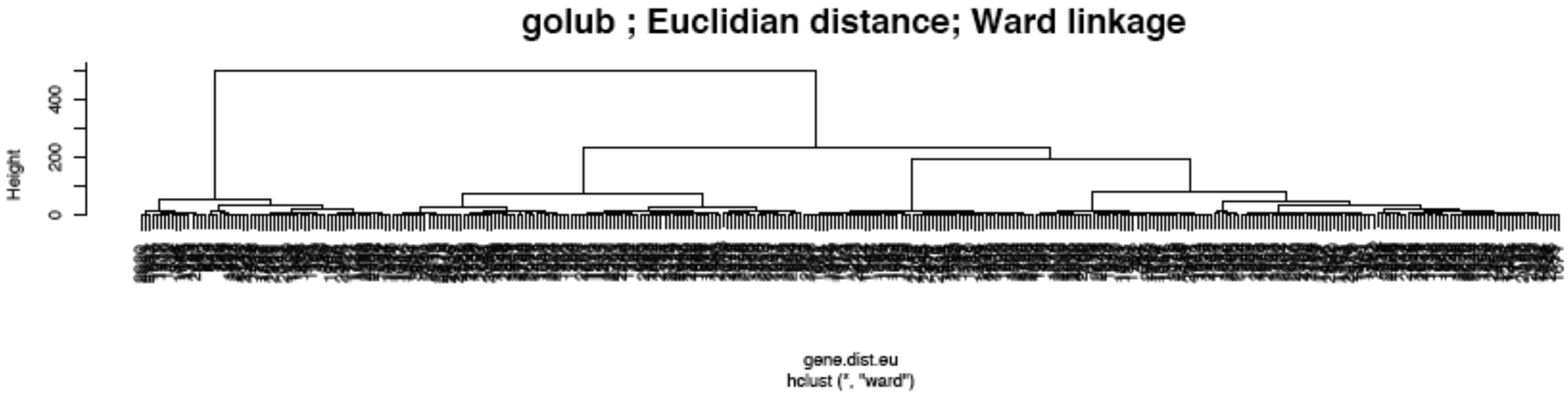


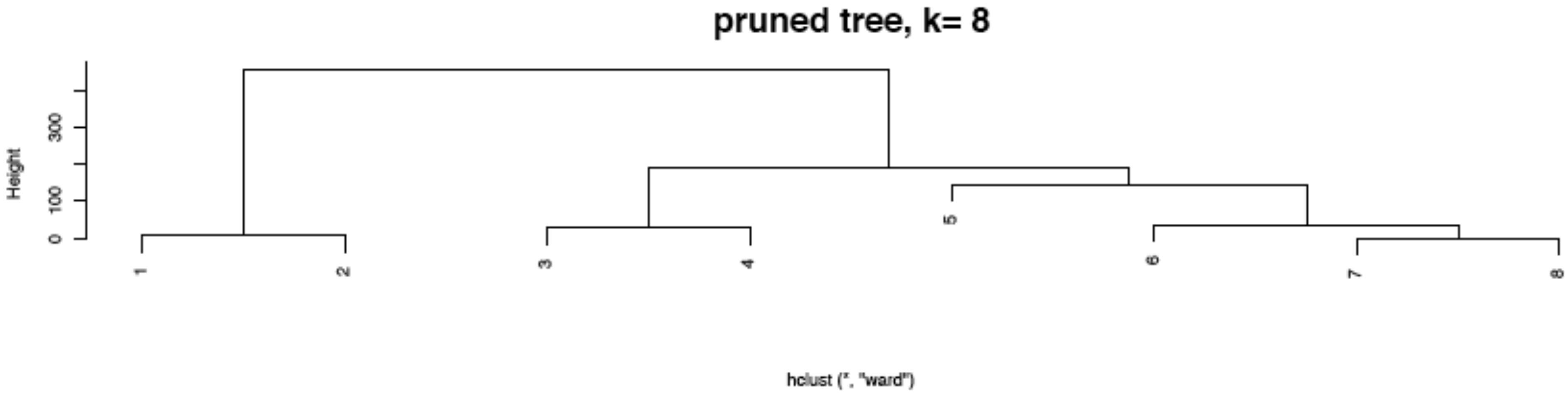
- Biclustering consists in clustering the rows (genes) and the columns (samples) of the data set.
- This reveals some subgroups of samples.
- With the golub 1999 data set
  - The AML and ALL patients are clearly separated at the top level of the tree
  - There are apparently two clusters among the ALL samples. Actually these two clusters correspond to distinct cell subtypes: T and B cells, respectively.

Impact of distance metrics and agglomeration rules

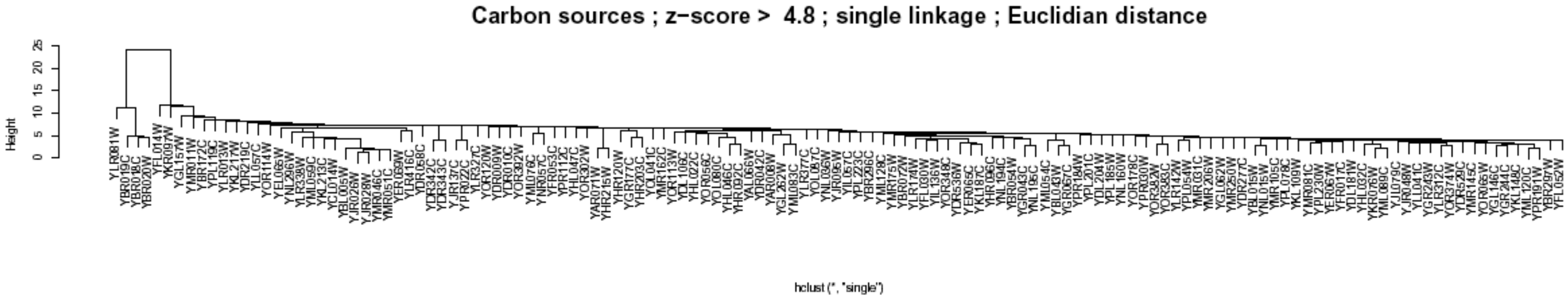


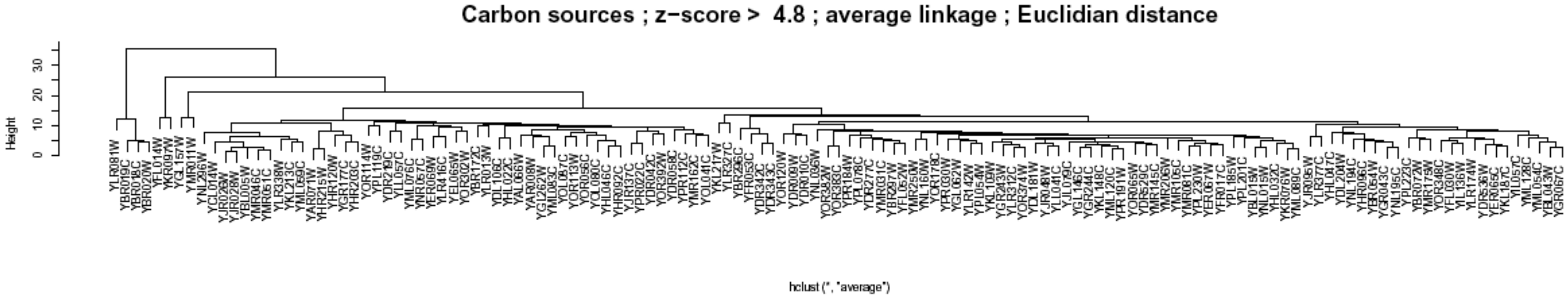
Golub 1999 - Pruning the tree

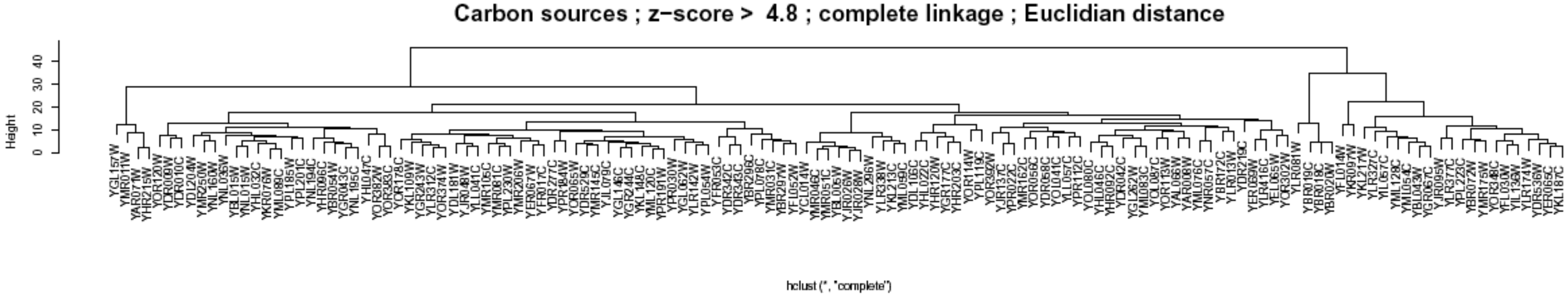


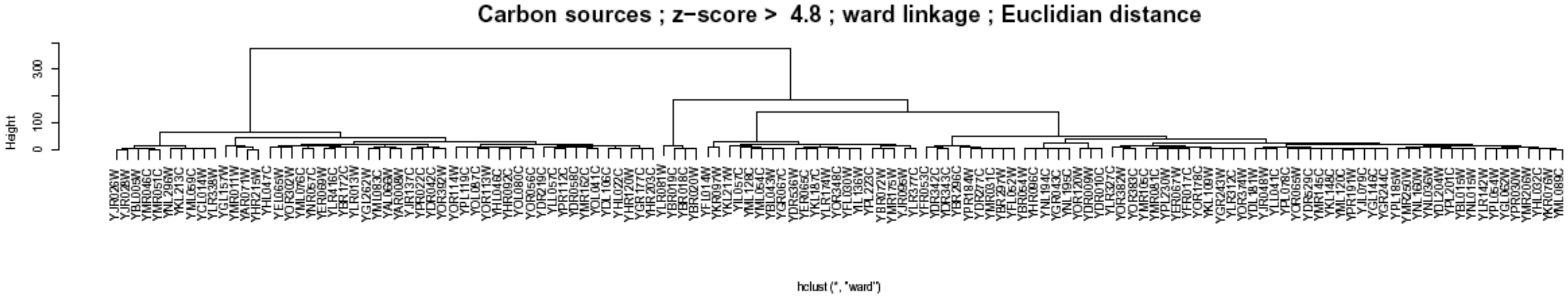


Impact of the linkage method

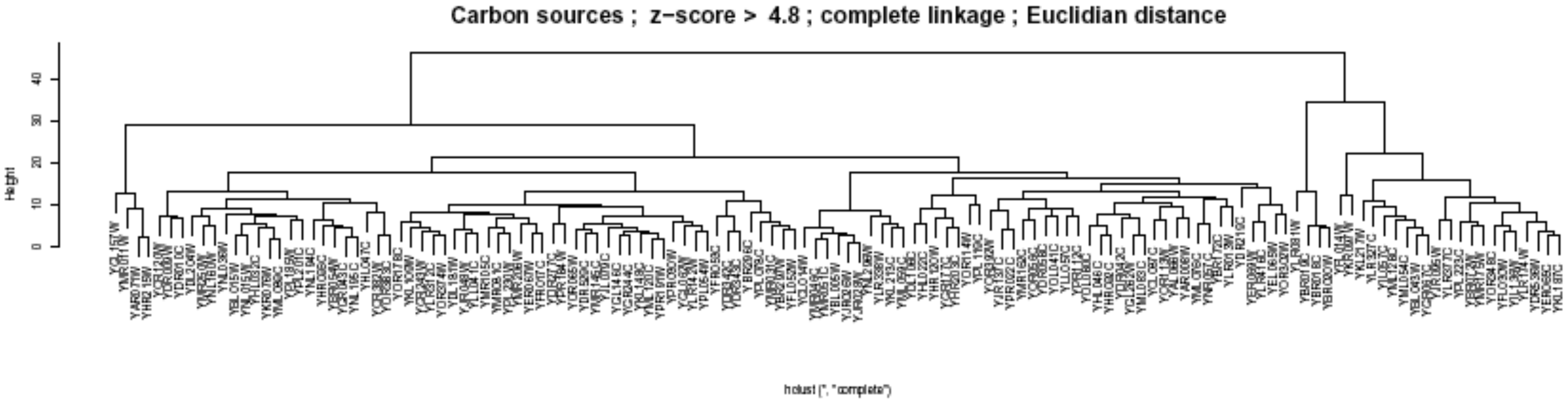

Carbon sources ; z−score > 4.8 ; single linkage ; Euclidian distance


Carbon sources ; z−score > 4.8 ; average linkage ; Euclidian distance

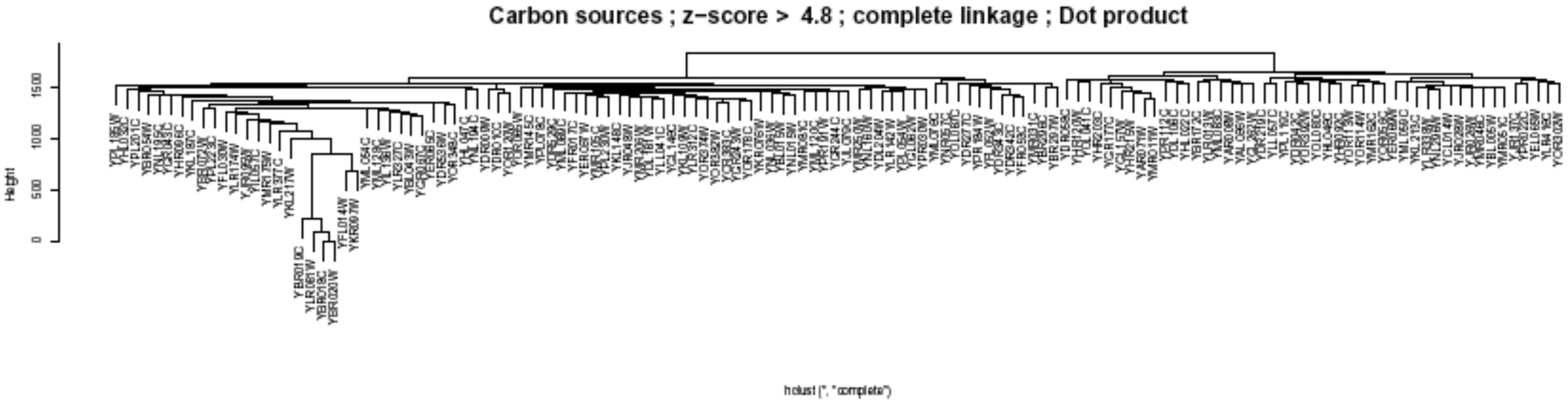

Carbon sources ; z−score > 4.8 ; complete linkage ; Euclidian distance


Carbon sources ; z−score > 4.8 ; ward linkage ; Euclidian distance

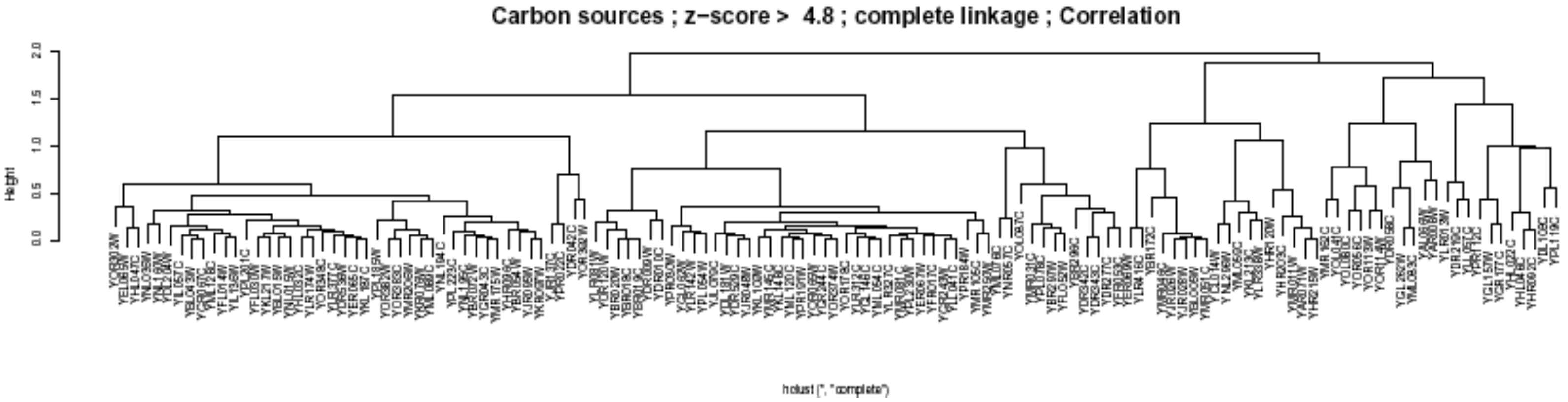
Impact of the distance metric - complete linkage



Carbon sources ; z-score > 4.8 ; complete linkage ; Euclidian distance

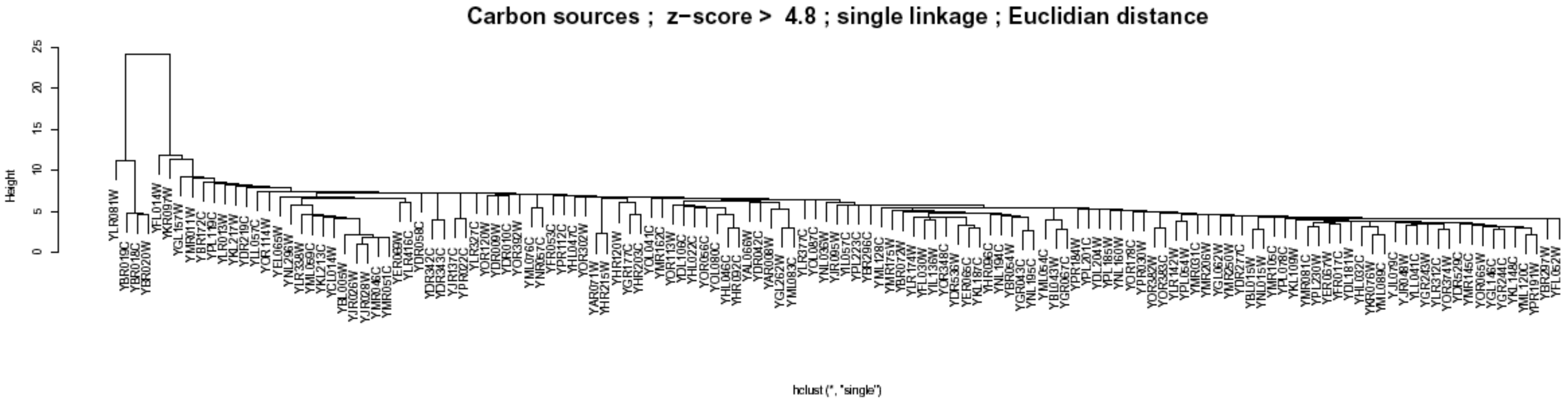


Carbon sources ; z-score > 4.8 ; complete linkage ; Dot product

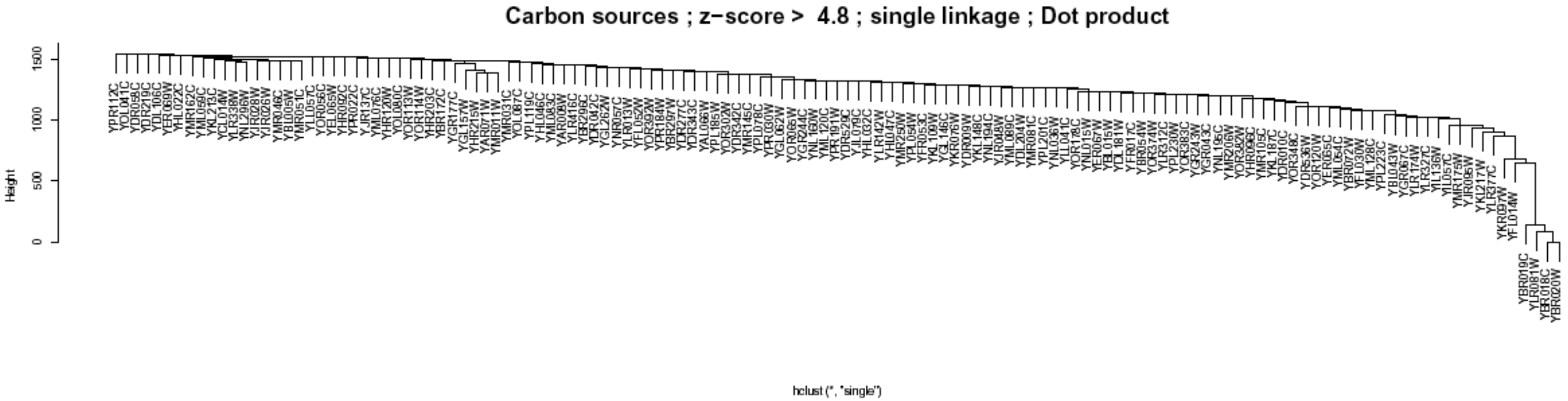


Carbon sources ; z-score > 4.8 ; complete linkage ; Correlation

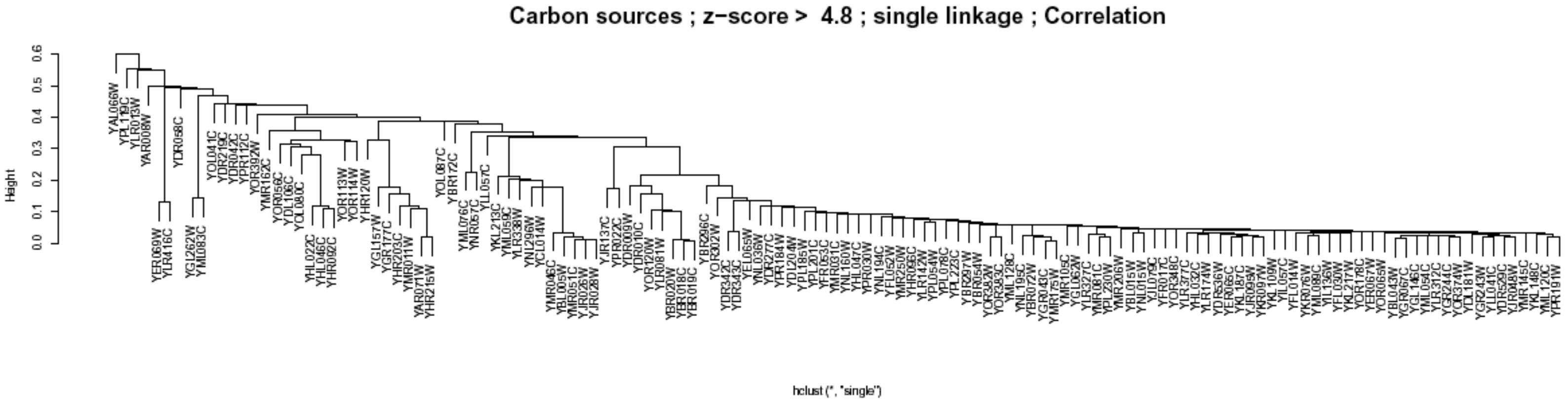
Ipact of the distance metric - single linkage



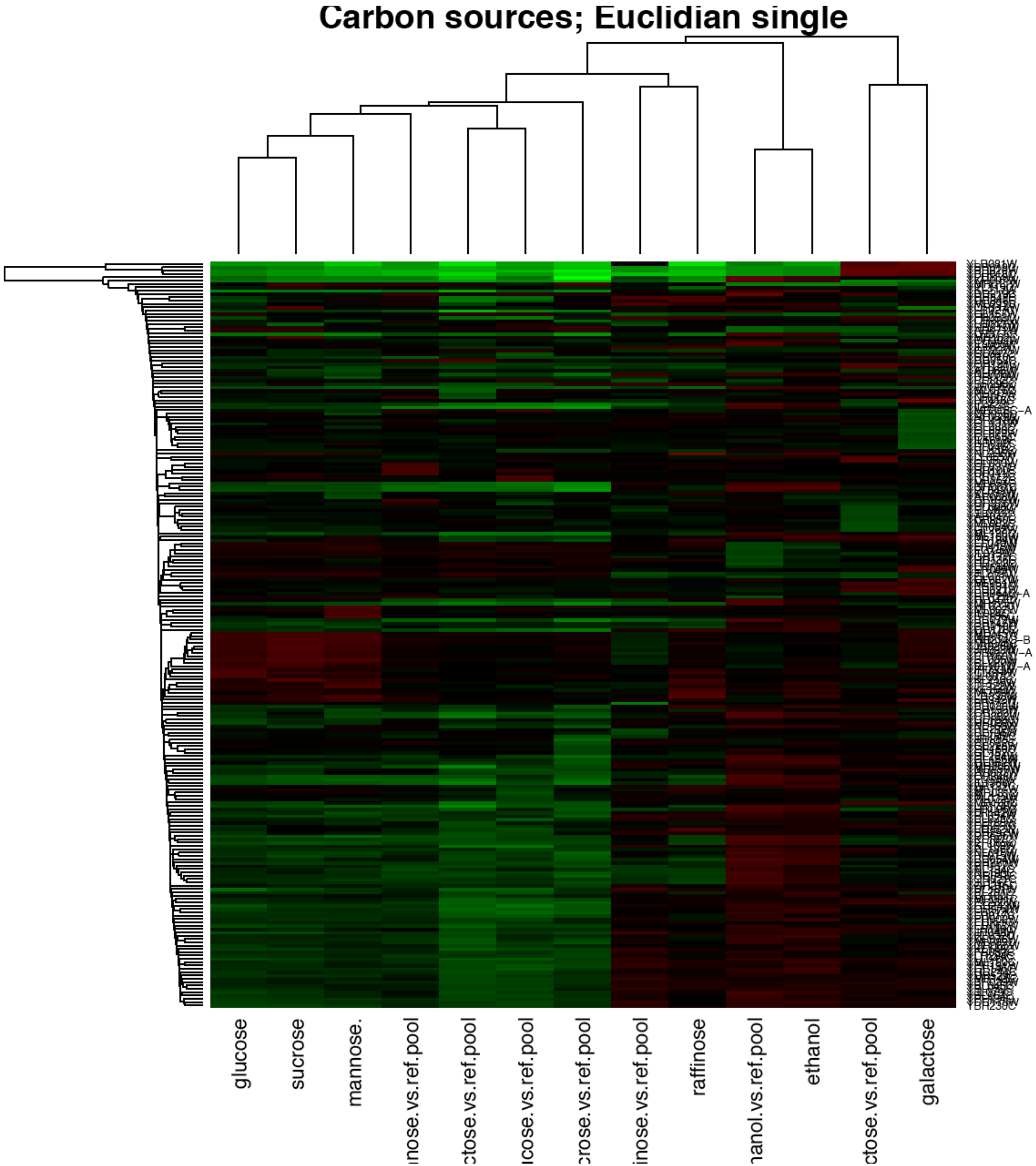
Carbon sources ; z−score > 4.8 ; single linkage ; Euclidian distance

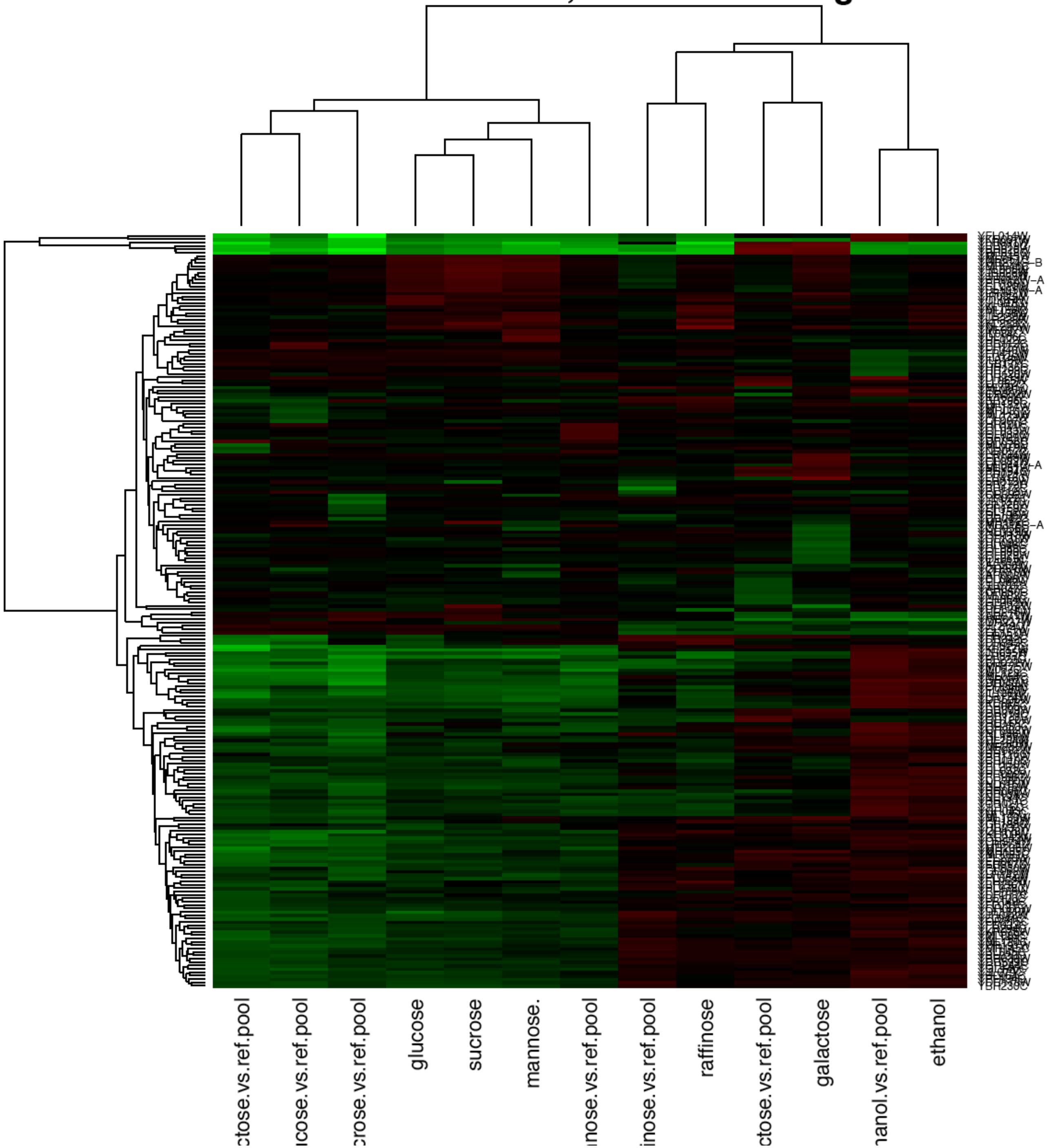


Carbon sources ; z−score > 4.8 ; single linkage ; Dot product

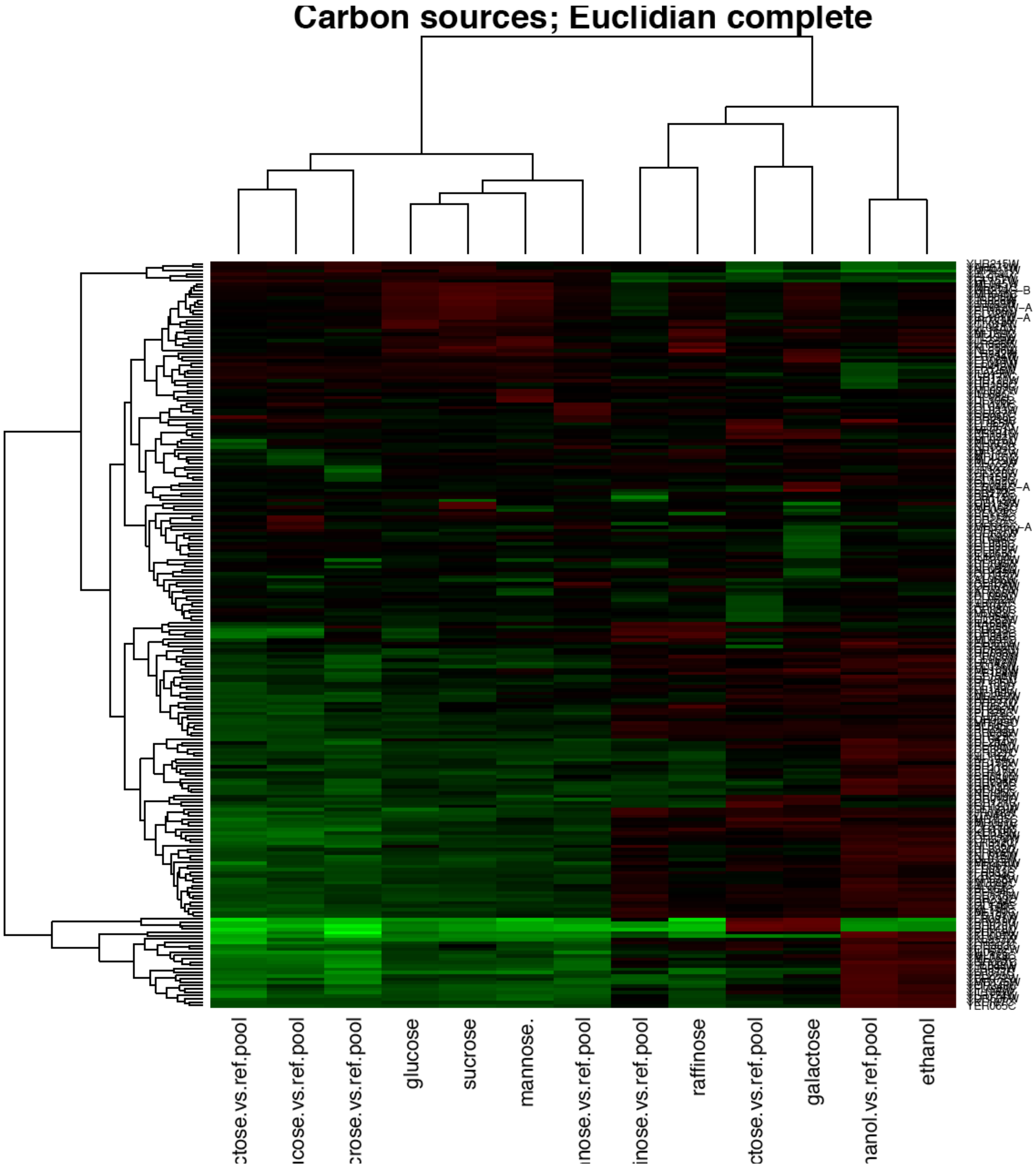


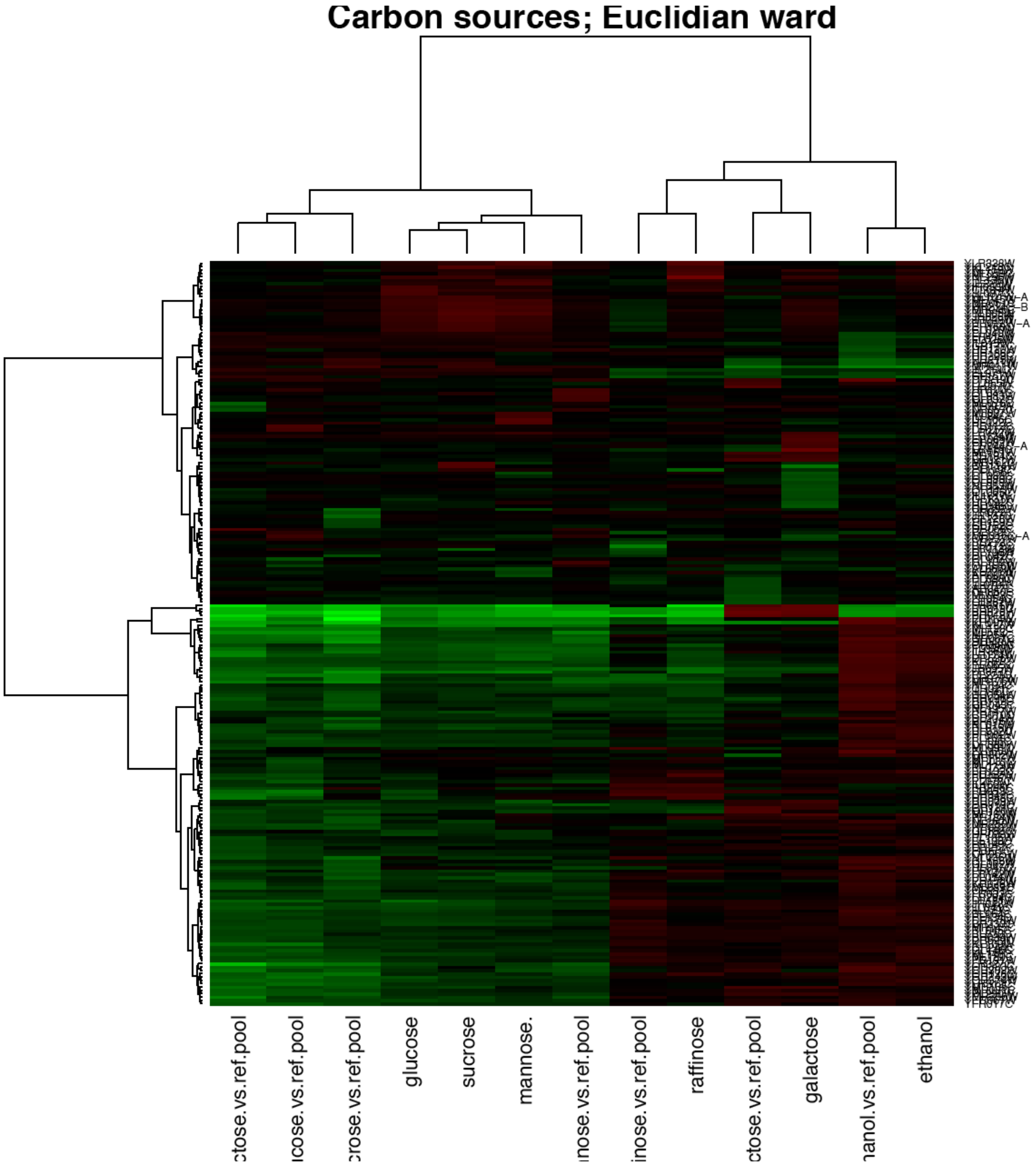
Carbon sources ; z−score > 4.8 ; single linkage ; Correlation


Carbon sources; Euclidian single
YLR081W
YBR230C
glucose
sucrose
mannose.
inose.vs.ref.pool
ctose.vs.ref.pool
icose.vs.ref.pool
crose.vs.ref.pool
inose.vs.ref.pool
raffinose
hanol.vs.ref.pool
ethanol
ctose.vs.ref.pool
galactose



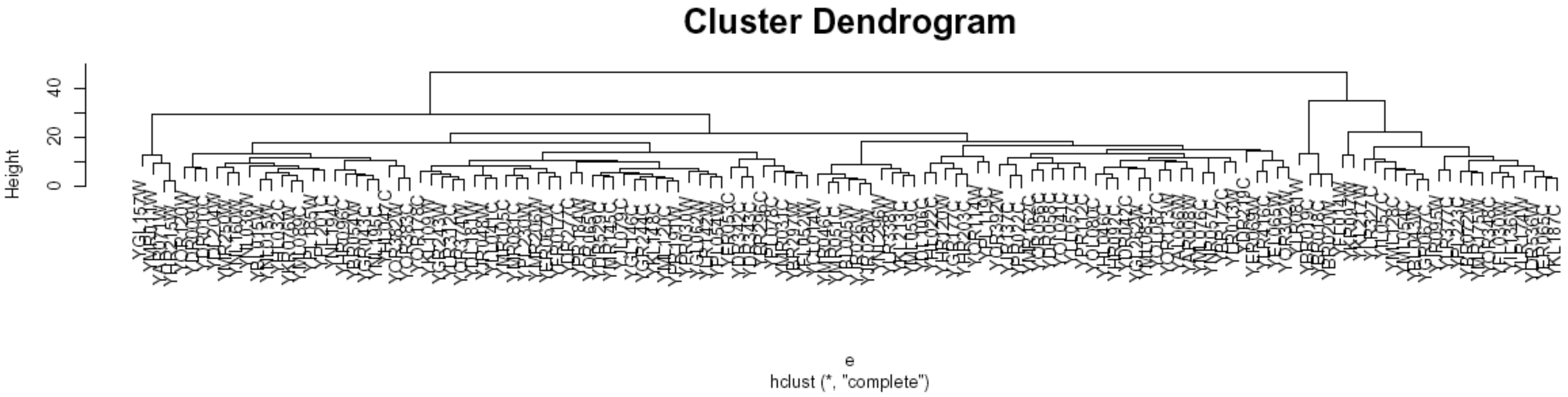
Carbon sources; Euclidian average

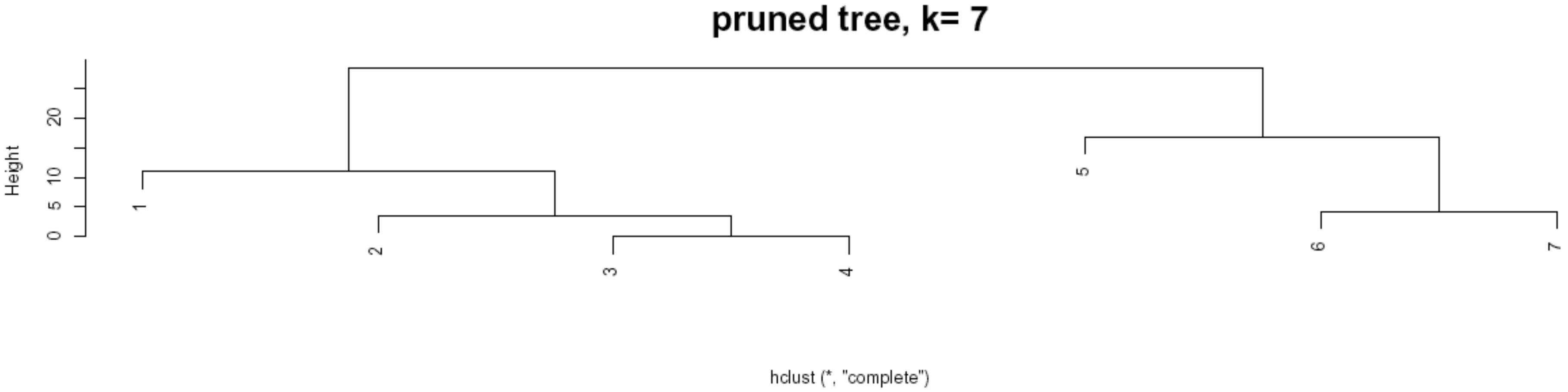

Carbon sources; Euclidian complete
ctose.vs.ref.pool
icose.vs.ref.pool
crose.vs.ref.pool
glucose
sucrose
mannose.
inose.vs.ref.pool
inose.vs.ref.pool
raffinose
ctose.vs.ref.pool
galactose
nanol.vs.ref.pool
ethanol
YHR215W
YER065C


Carbon sources; Euclidian ward
ctose.vs.ref.pool
icose.vs.ref.pool
crose.vs.ref.pool
glucose
sucrose
mannose.
nose.vs.ref.pool
inose.vs.ref.pool
raffinose
ctose.vs.ref.pool
galactose
anol.vs.ref.pool
ethanol
YLR338W
YFR017C

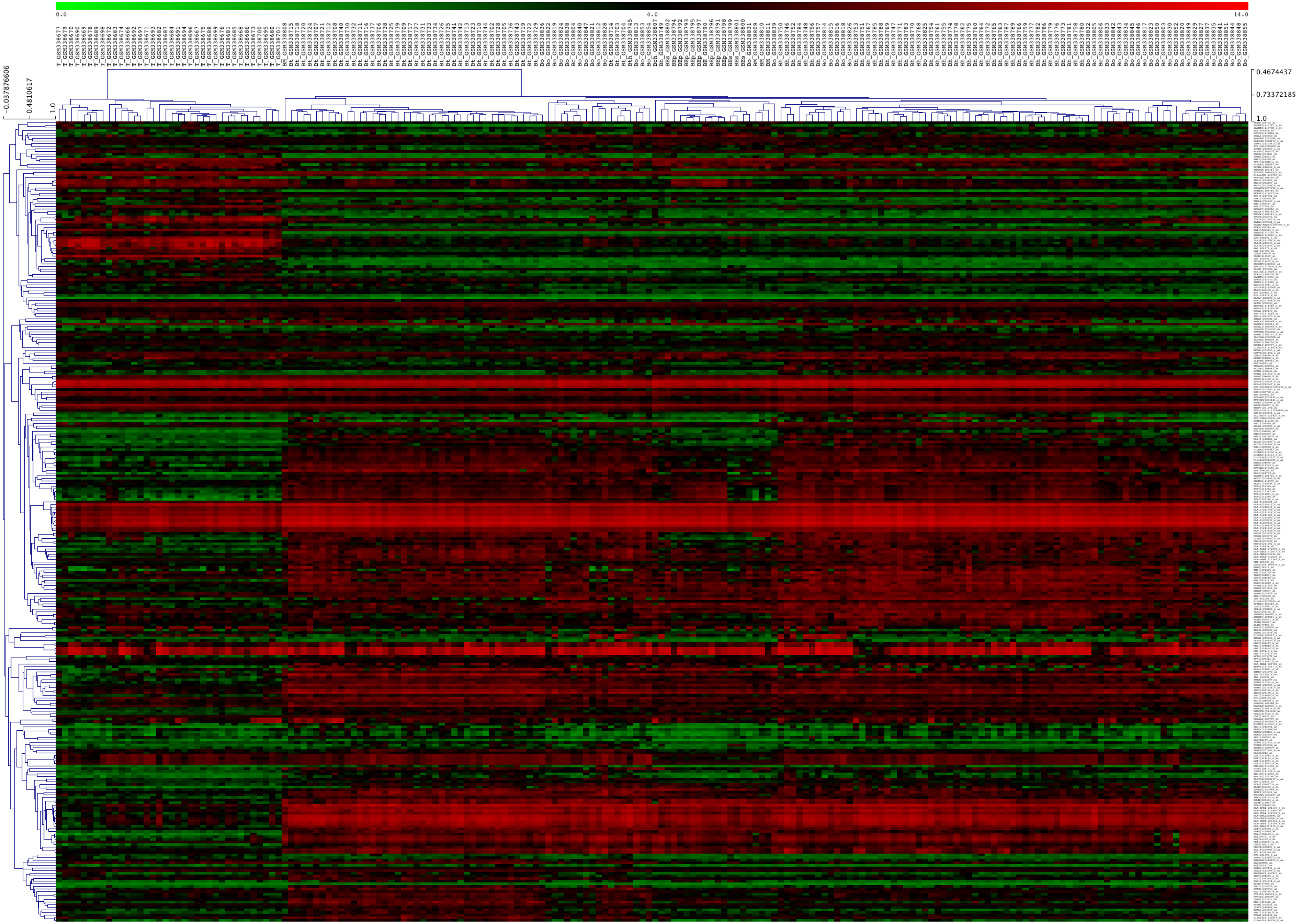
Pruning and cutting the tree

- The tree can be cut at level k (starting from the root), which creates *k* clusters
- A k-group partitioning is obtained by collecting the leaves below each branch of the pruned tree





Den Boer 2009 – Hierarchical clustering

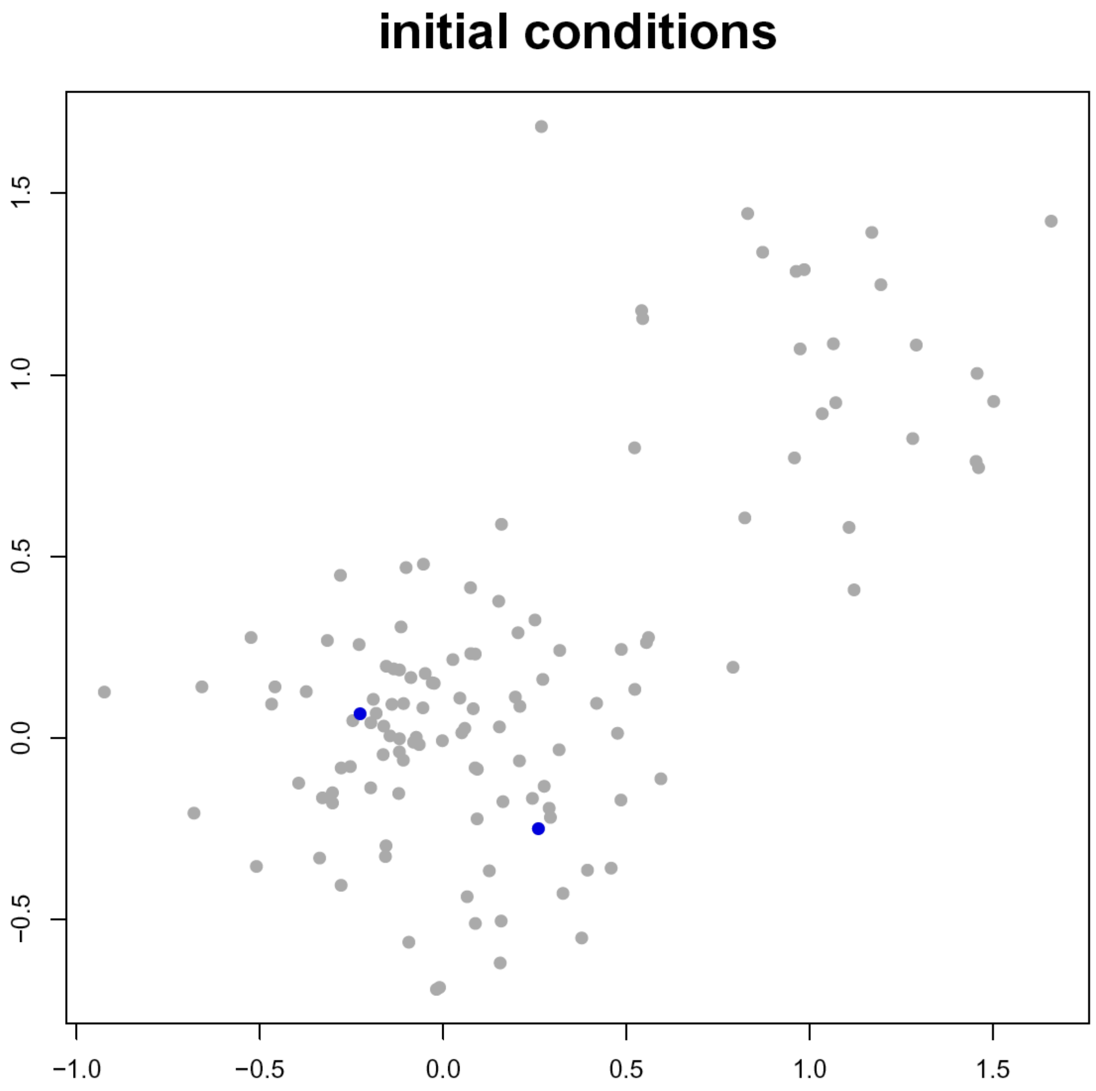


K-means clustering

Clustering around mobile centres

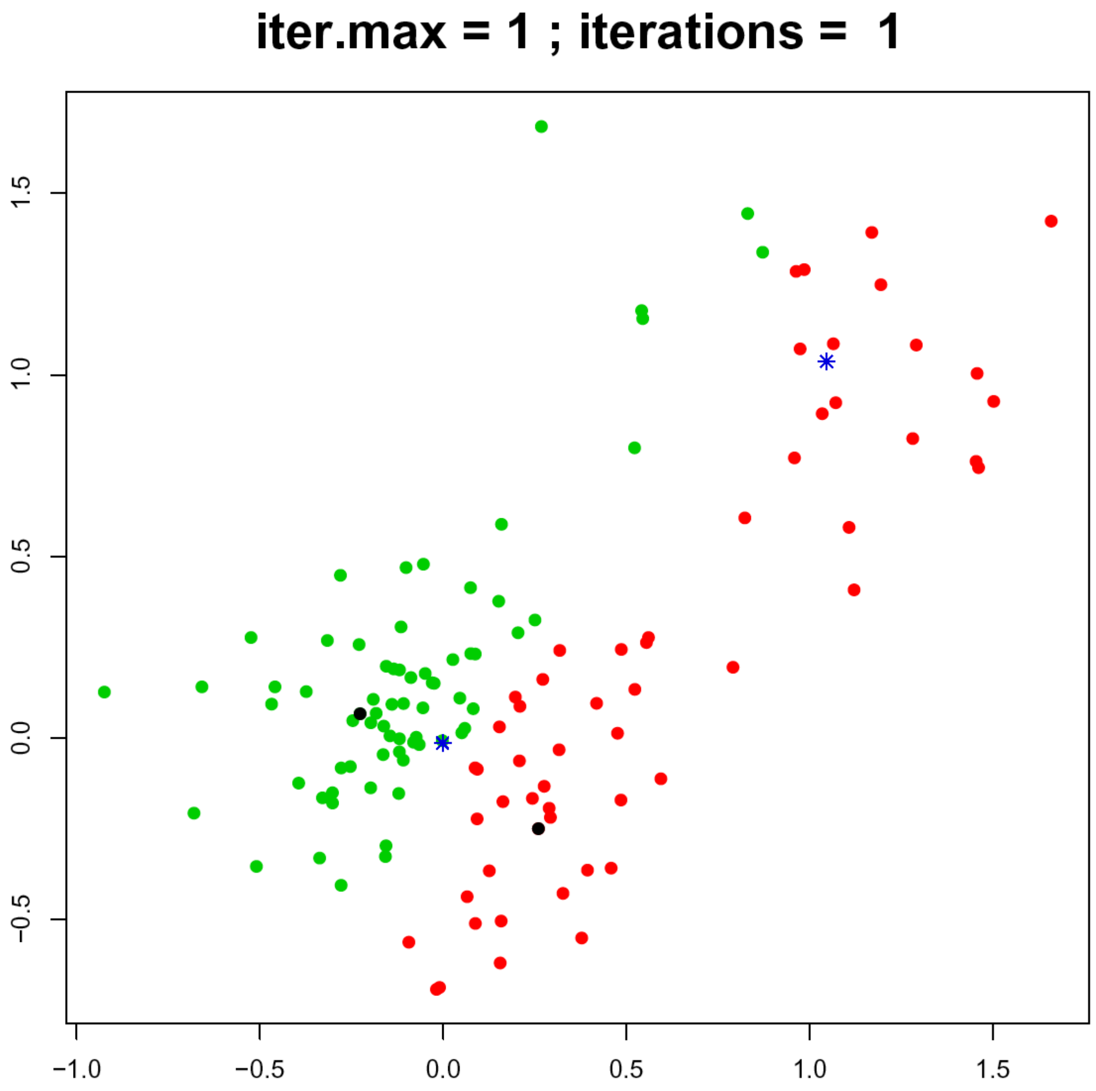
- The number of centres (k) has to be specified a priori
- Algorithm
  - ❑ (1) Arbitrarily select k initial centres
  - ❑ (2) Assign each element to the closest centre
  - ❑ (3) Re-calculate centres (mean position of the assigned elements)
  - ❑ (4) Repeat (2) and (3) until one of the stopping conditions 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



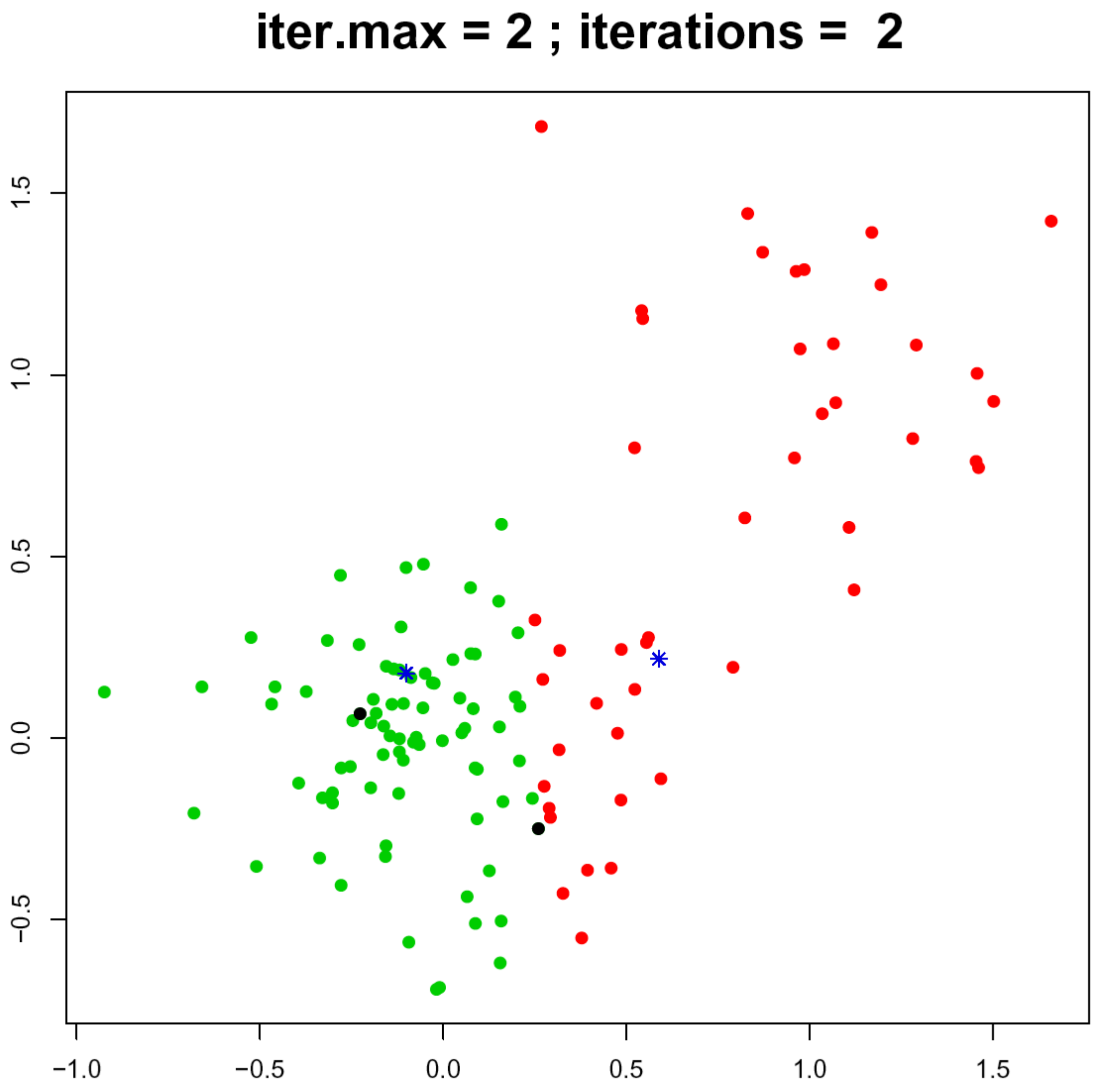
- Two sets of random points are randomly generated
  - 200 points centred on (0,0)
  - 50 points centred on (1,1)
- Two points are randomly chosen as seeds (blue dots)

Mobile centres example - first iteration



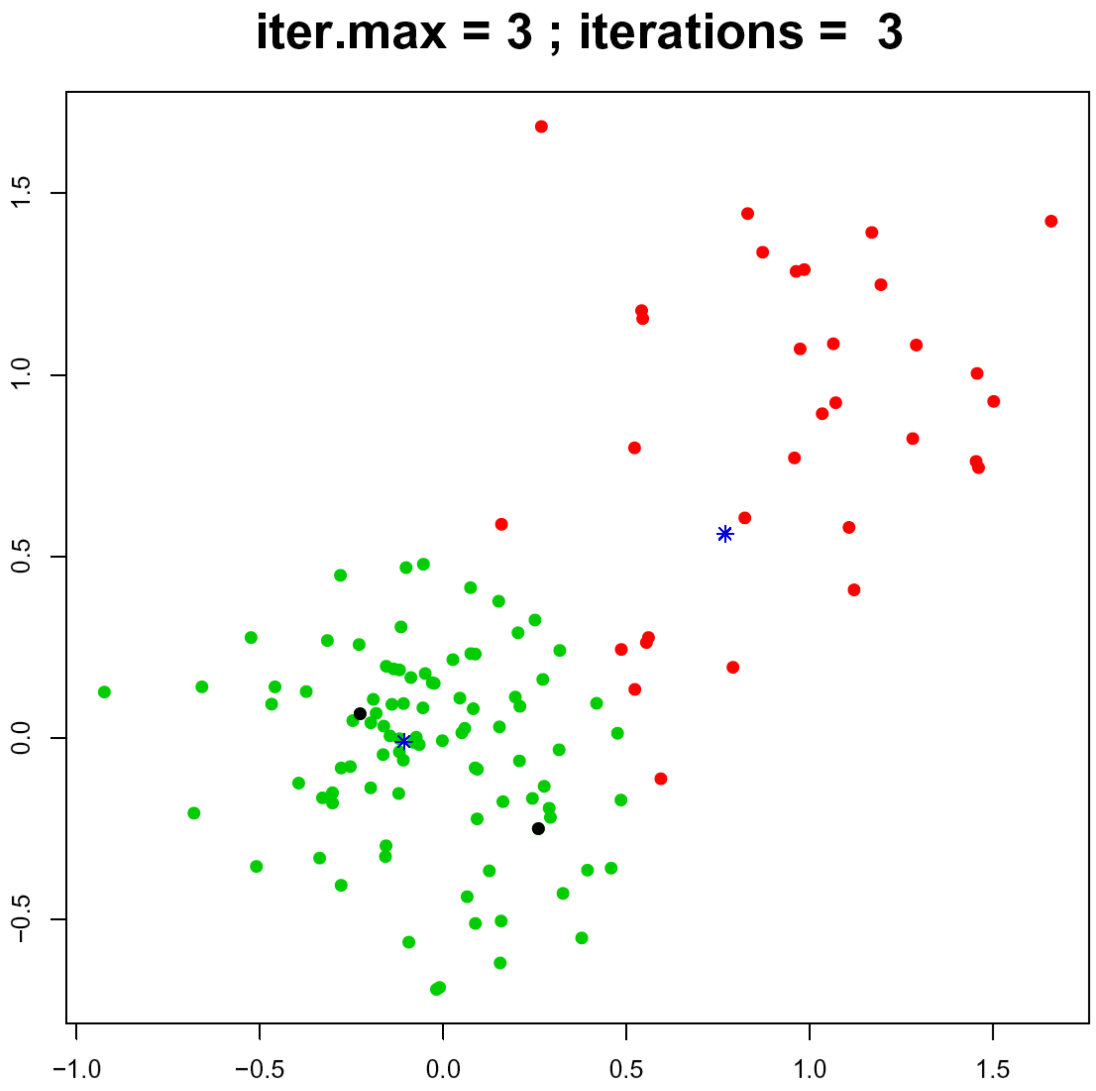
- Step 1
  - Each dot is assigned to the cluster with the closest centre
  - Centres are re-calculated (blue star) on the basis of the new clusters

Mobile centres example - second iteration



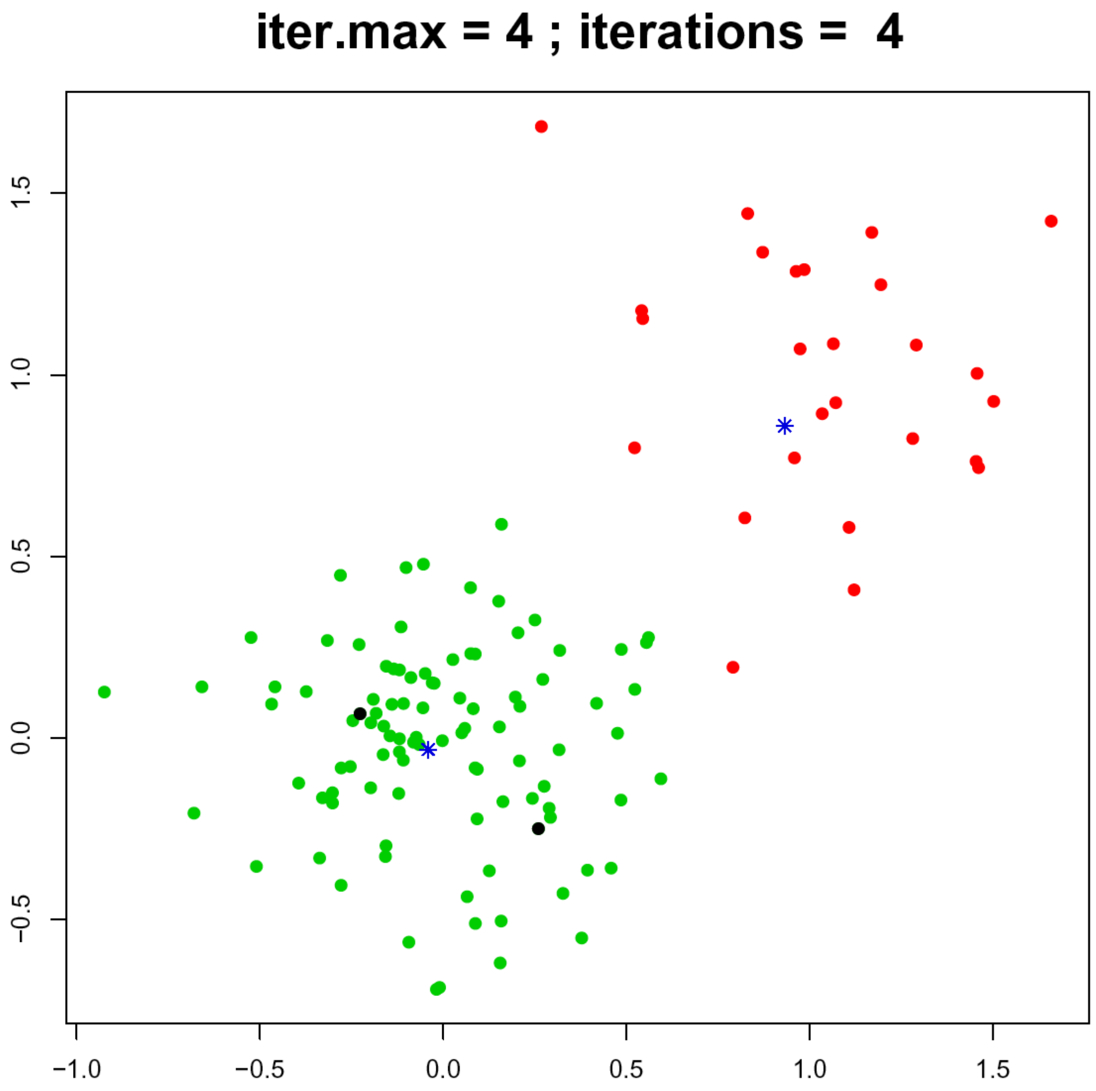
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 3 iterations



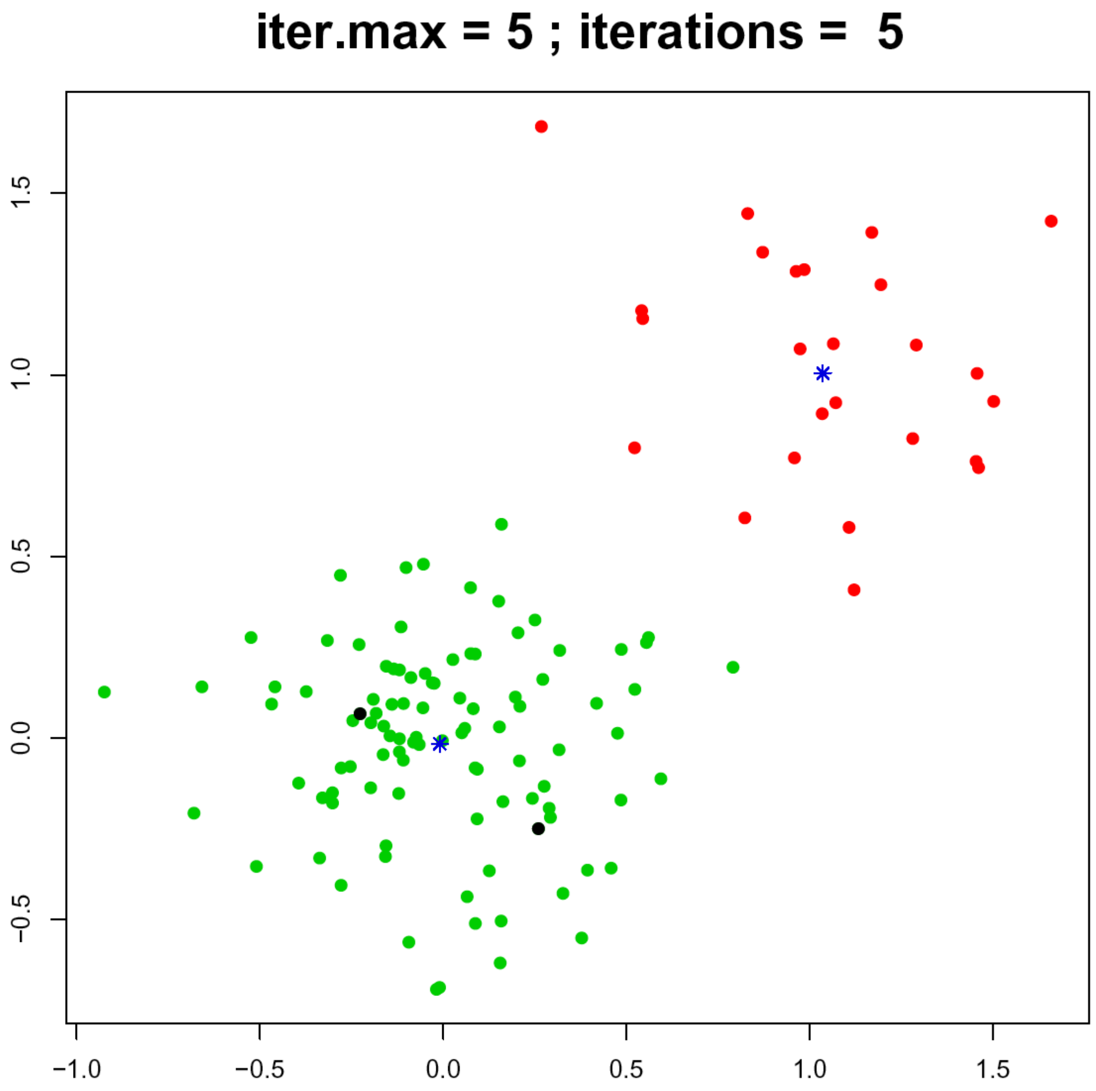
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 4 iterations



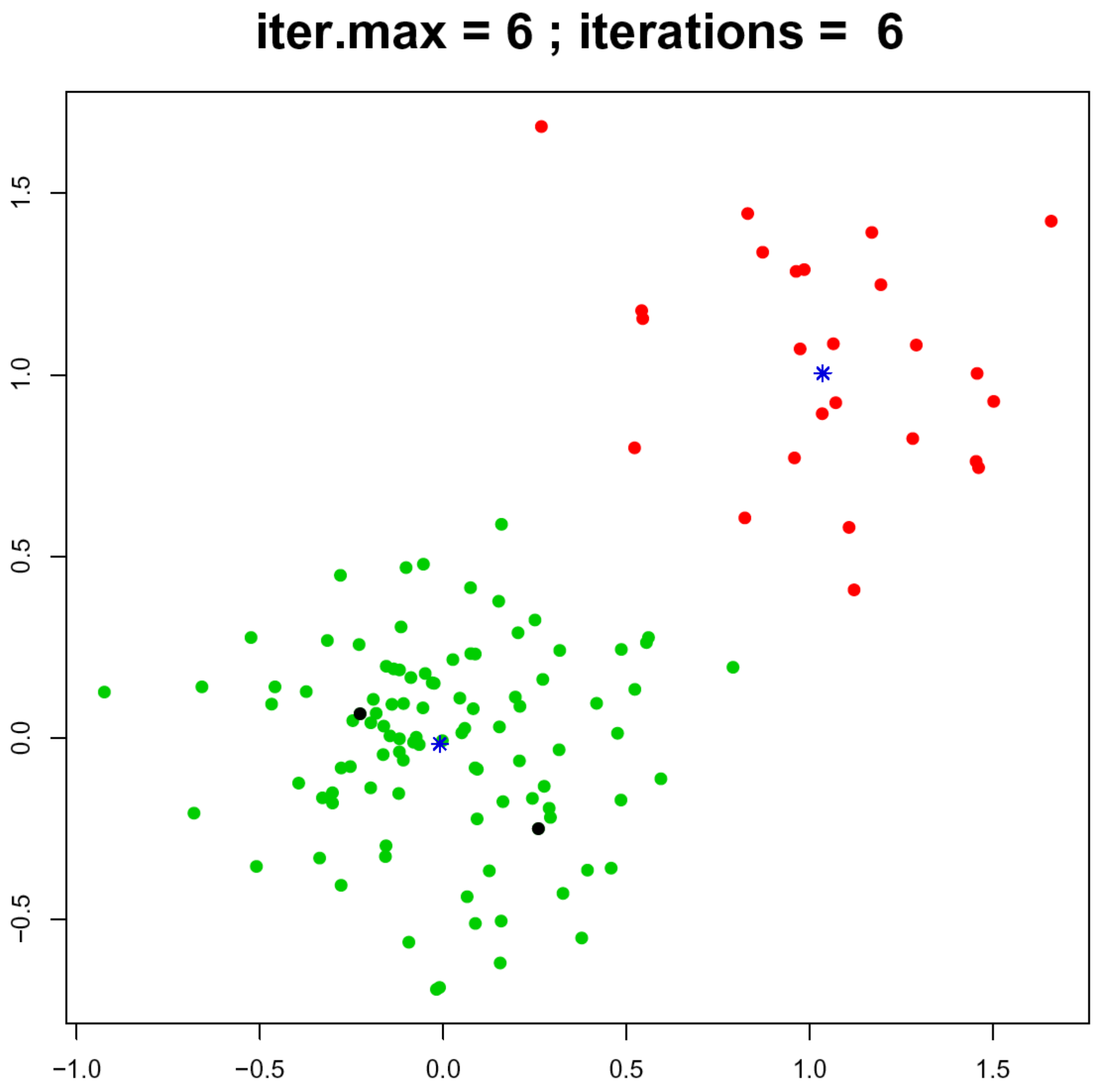
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 5 iterations



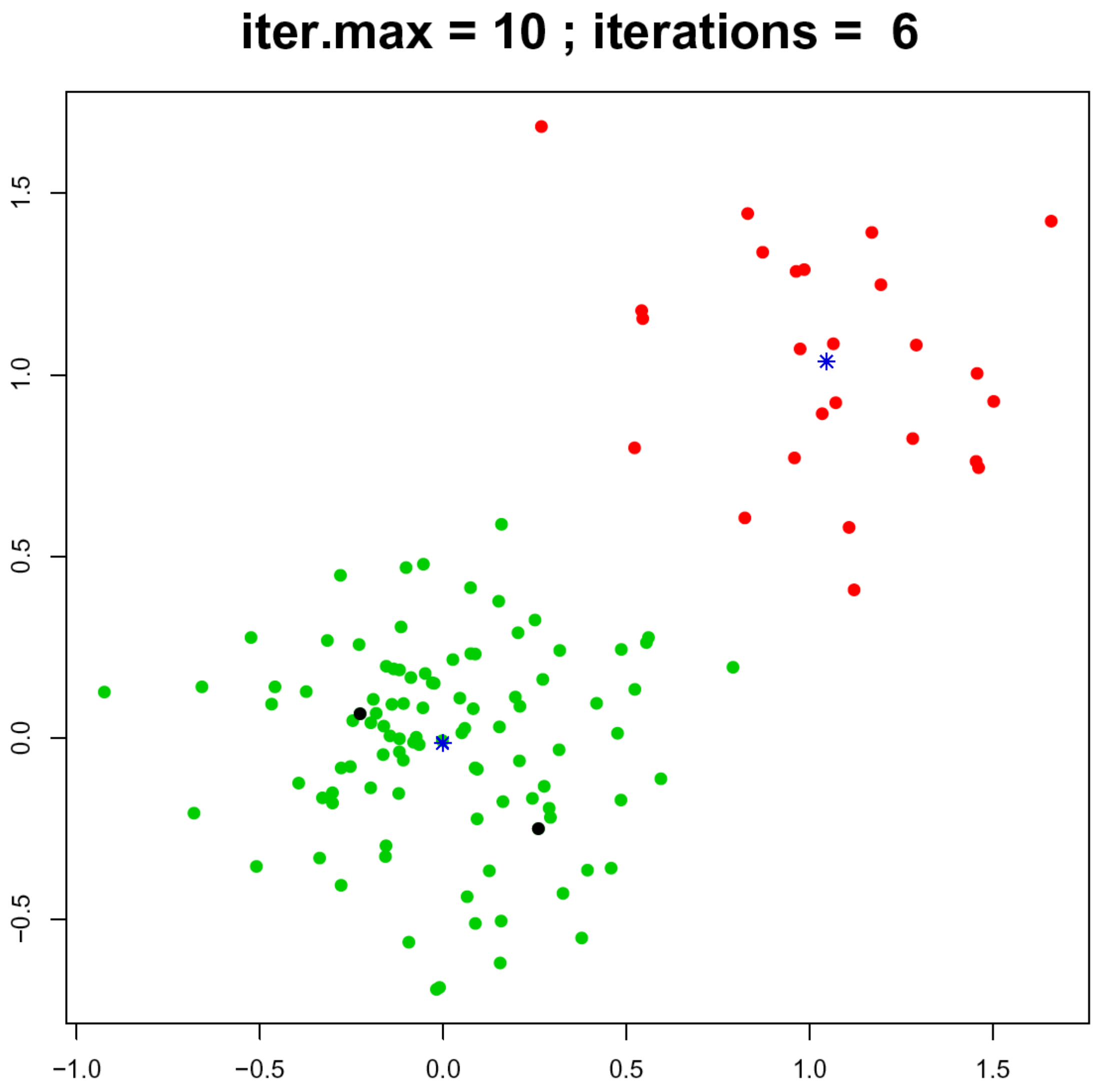
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

Mobile centres example - after 6 iterations



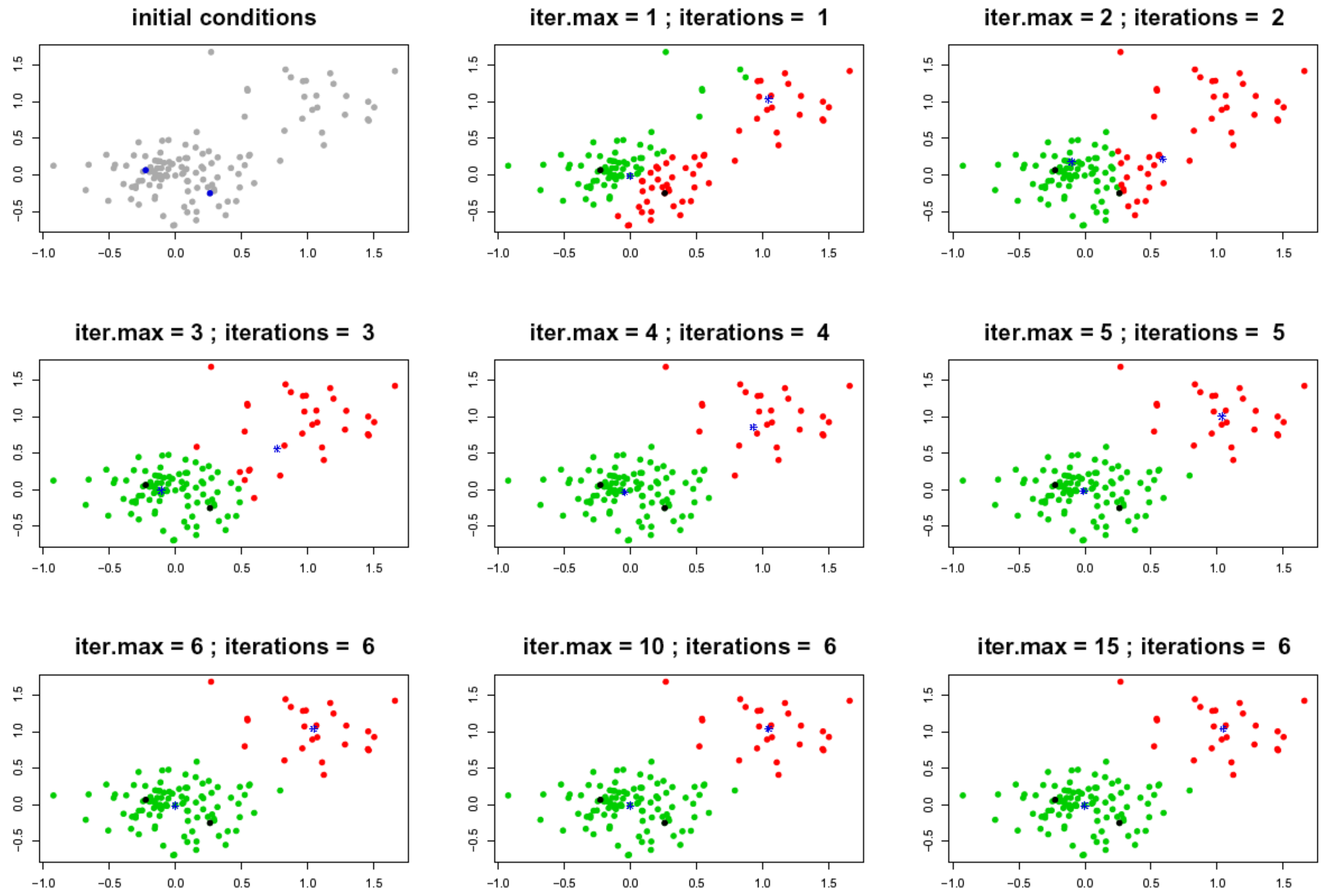
- At each step,
  - points are re-assigned to clusters
  - centres are re-calculated
- Cluster boundaries and centre positions evolve at each iteration

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



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)

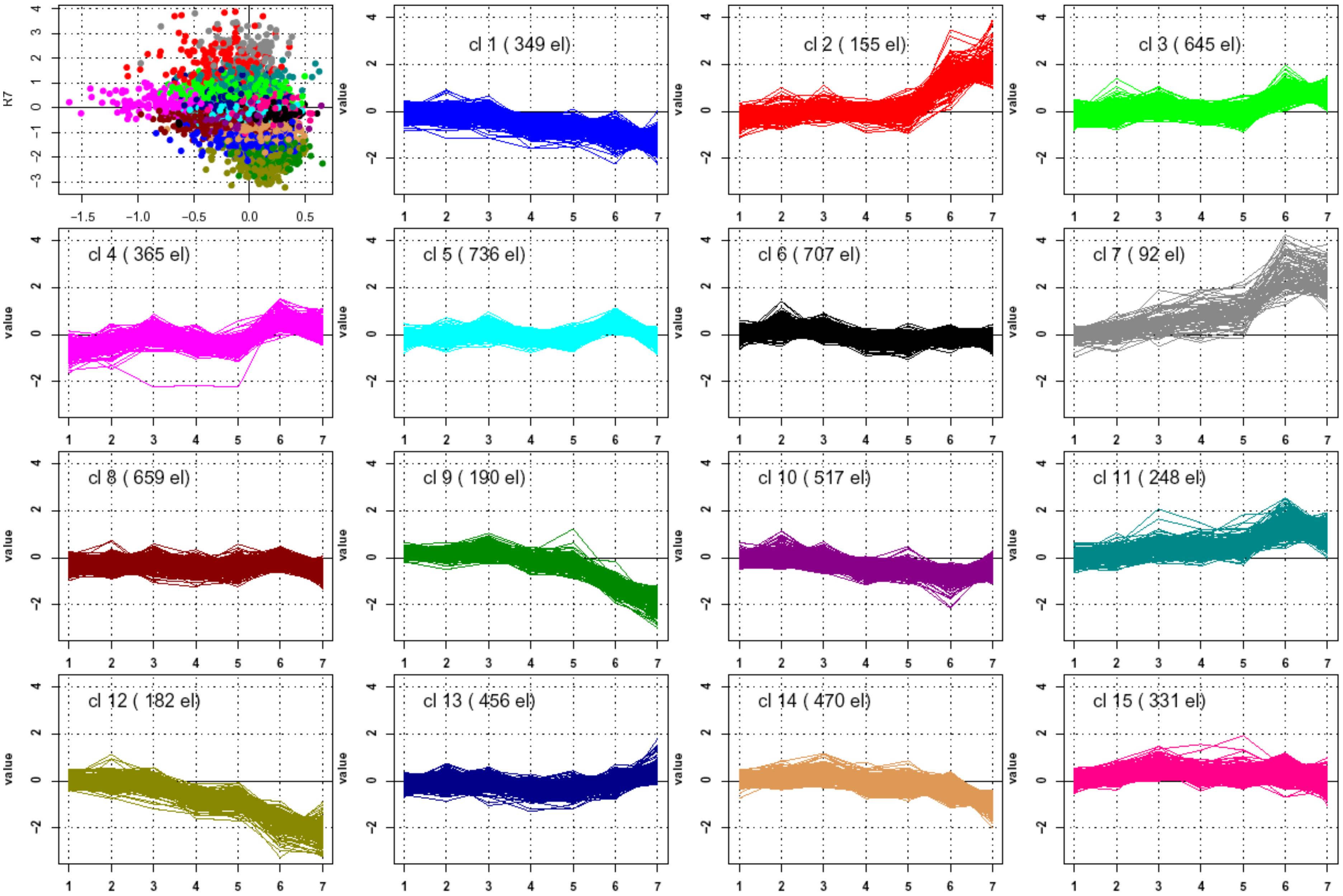
Clustering with gene expression data

- Clustering can be performed in two ways
  - Taking genes as objects and conditions/cell types as variables
  - Taking conditions/cell types as objects and genes as variables
- Problem of dimensionality
  - When genes are considered as variables, there are many more variables than objects
  - Generally, only a very small fraction of the genes are regulated (e.g. 30 genes among 6,000)
  - However, all genes will contribute equally to the distance metrics
  - The noise will thus affect the calculated distances between conditions
- Solution
  - Selection of a subset of strongly regulated genes before applying clustering to conditions/cell types

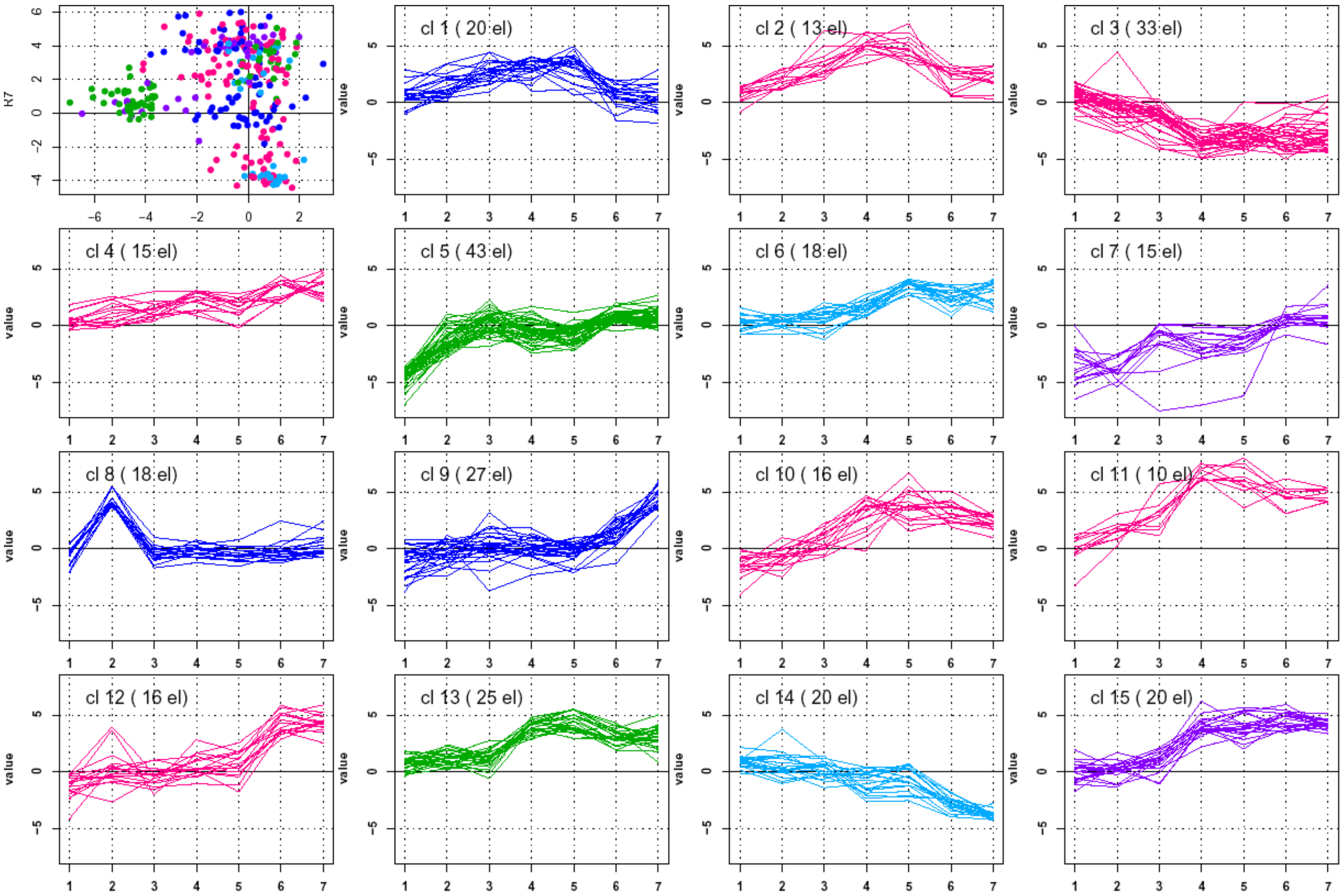
K-means clustering

- 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)

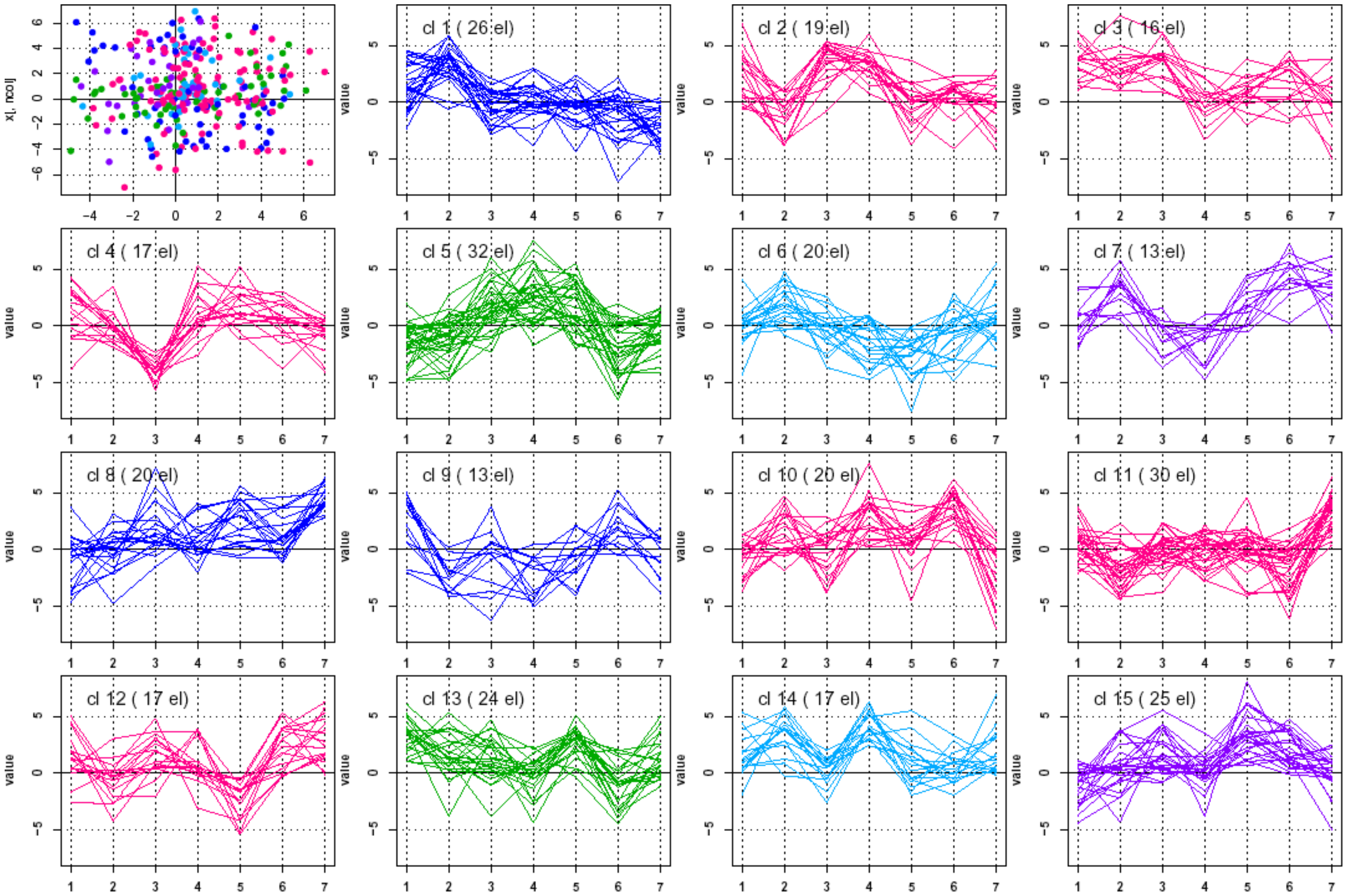
Diauxic shift: k-means clustering on all genes



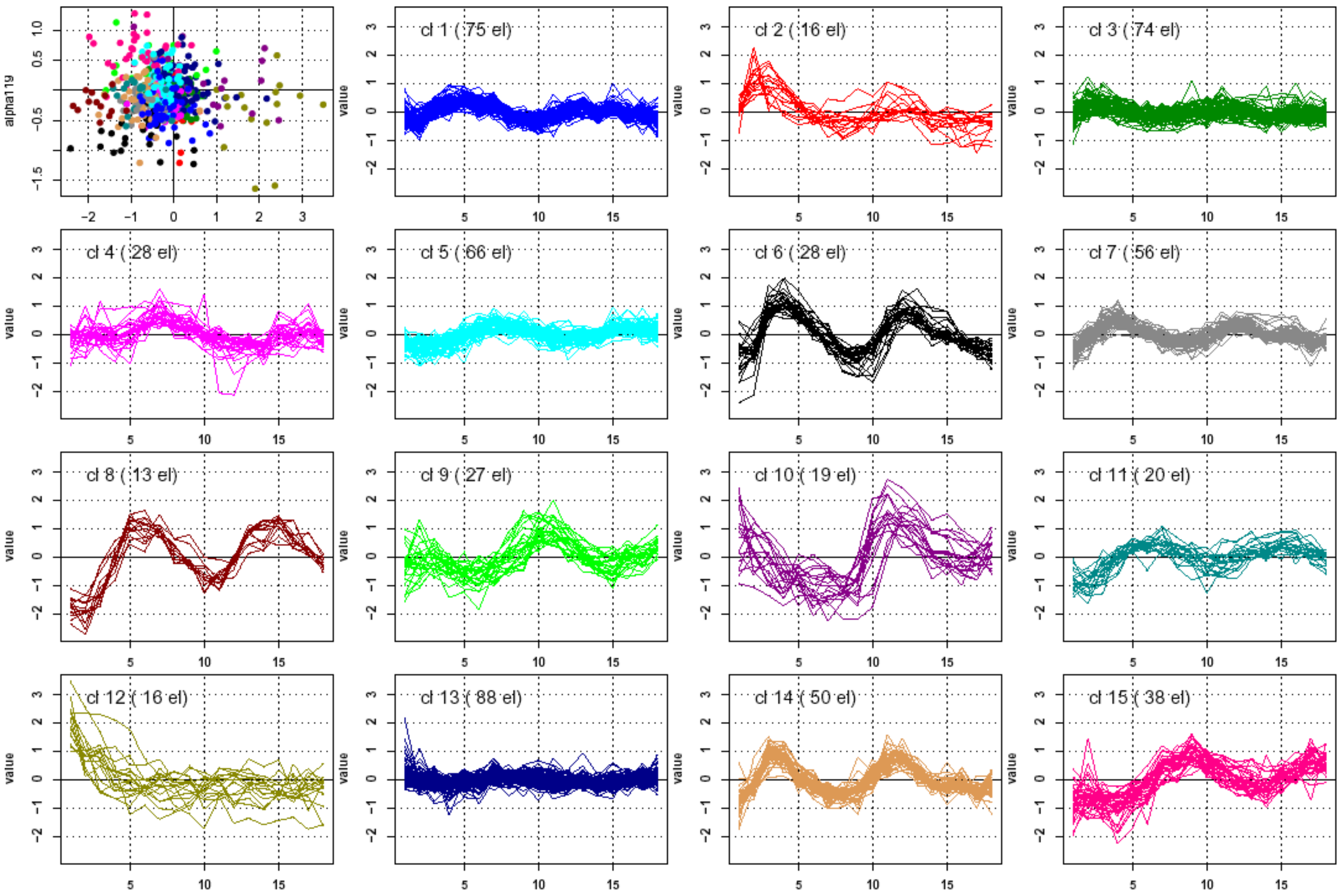
Diauxic shift: k-means clustering on filtered genes



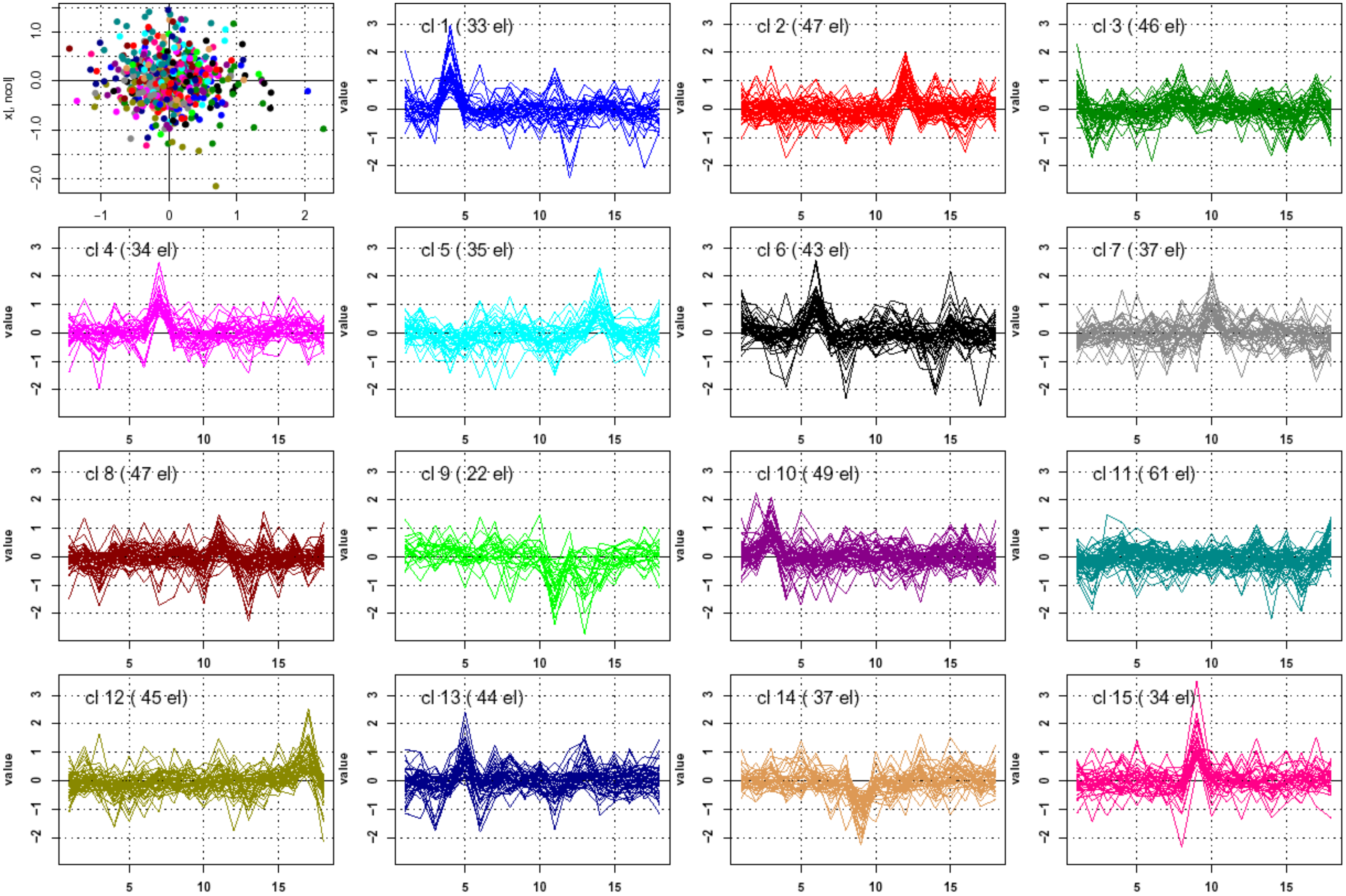
Diauxic shift: k-means clustering on permuted filtered genes



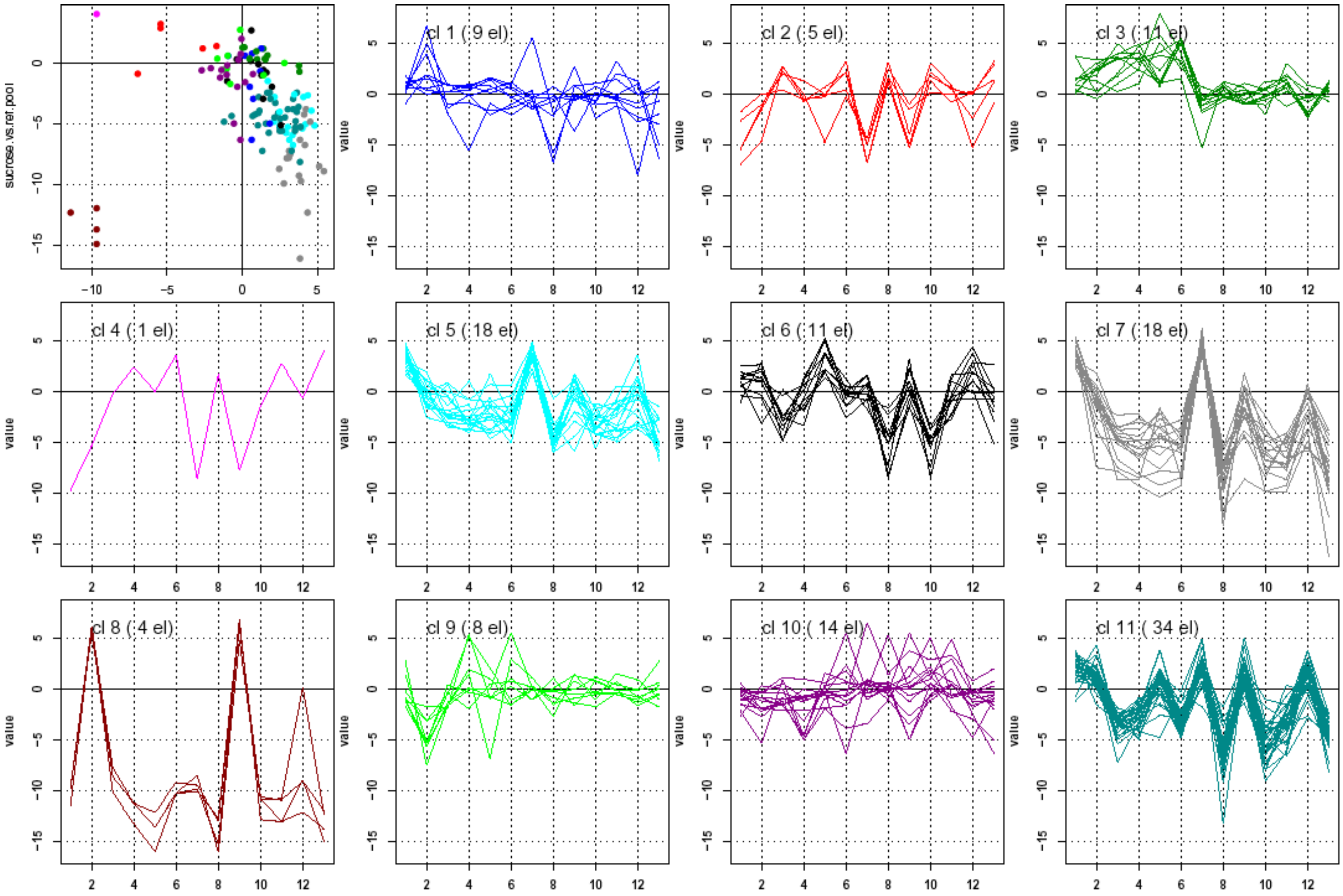
Cell cycle data: K-means clustering



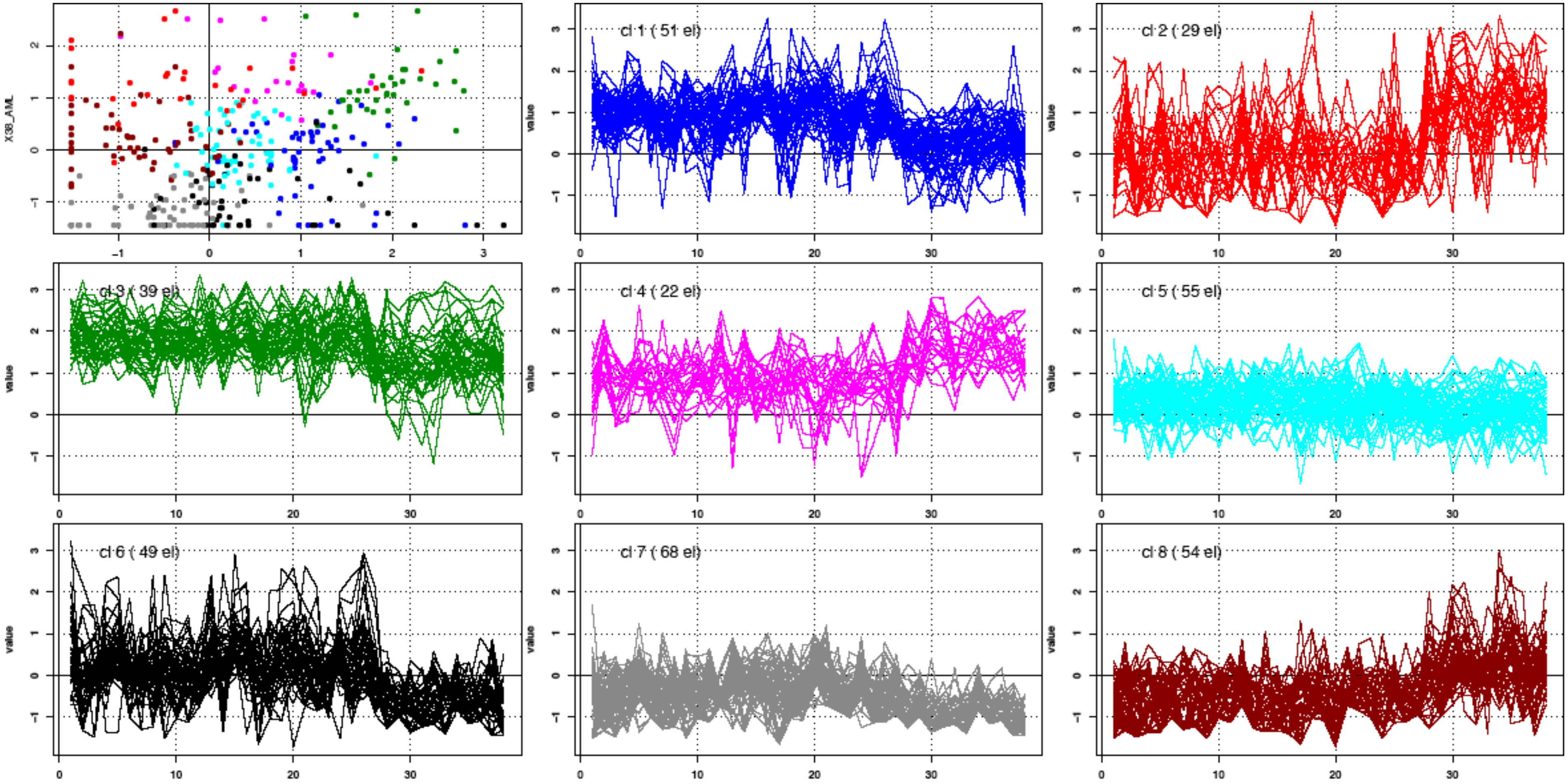
Cell cycle data: K-means clustering, permuted data



Carbon sources: K-means clustering



Golub - K-means clustering



K-means clustering - summary

- Strengths
  - Simple to use
  - Fast
  - Can be used with very large data sets
- Weaknesses
  - The choice of the number of groups is arbitrary
  - The results vary depending on the initial positions of centres
  - The R implementation is based on Euclidian distance, no other metrics are proposed
- Solutions
  - Try different values for k and compare the result
  - For each value of k, run repeatedly to sample different initial conditions
- Weakness of the solution
  - Instead of one clustering, you obtain hundreds of different clustering results, totaling thousands of clusters, how to decide among them

Evaluation of clustering results

How to evaluate the result ?

- It is very hard to make a choice between the multiple possibilities of distance metrics, clustering algorithms and parameters.
- Several criteria can be used to evaluate the clustering results
  - **Consensus**: using different methods, comparing the results and extracting a consensus
  - **Robustness**: running the same algorithm multiple times, with different initial conditions
    - Bootstrap
    - Jack-knife
    - Test different initial positions for the k-means
  - **Biological relevance:** compare the clustering result to functional annotations (functional catalogs, metabolic pathways, ...)

Comparing two clustering results

- If two methods return partitions of the same size, their clusters can be compared in a confusion table
- Optimal correspondences between clusters can be established (permuting columns to maximize the diagonal)
- The consistency between the two classifications can then be estimated with the hit rate
- Example :
  - Carbon source data, comparison of k-means and hierarchical clustering

hierarchical clustering

k-means clustering

	k1	k2	k3	k4	k5	k6	k7	Sum
h1	0	0	2	18	14	1	0	35
h2	0	0	0	4	0	0	0	4
h3	0	0	0	0	10	0	0	10
h4	40	0	10	0	0	9	0	59
h5	2	12	0	0	0	5	0	19
h6	0	0	0	0	0	0	4	4
h7	0	2	0	0	0	0	0	2
Sum	42	14	12	22	24	15	4	133

hierarchical clustering

k-means clustering

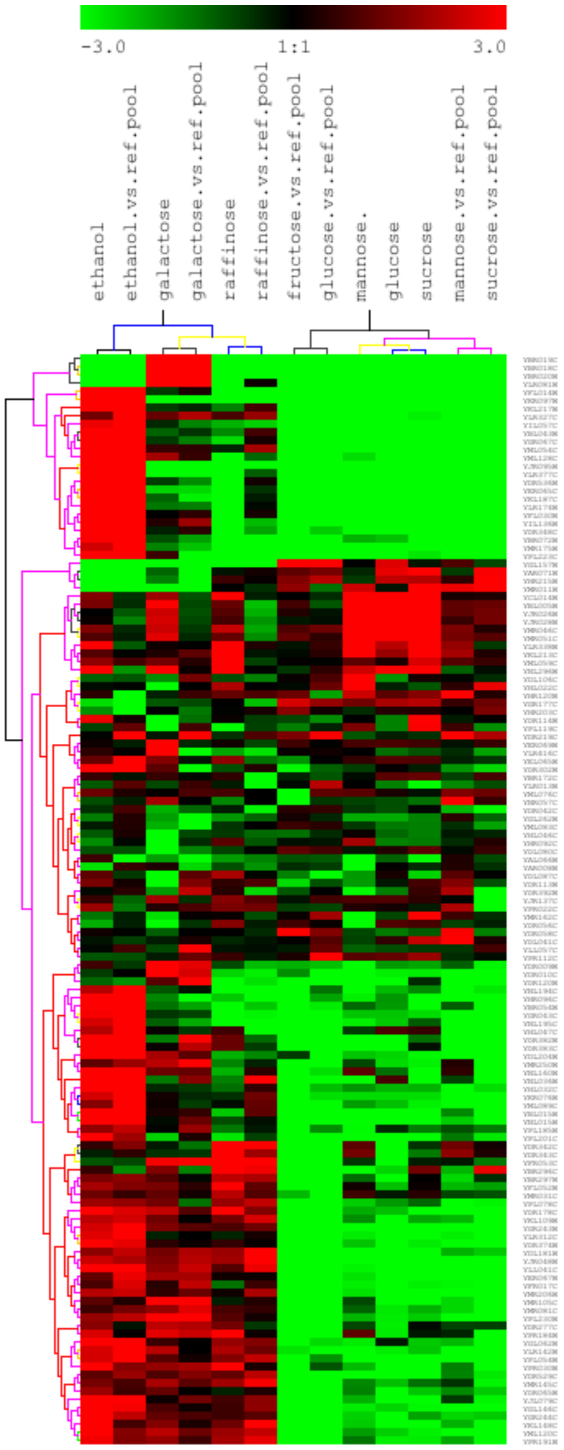
	k4	k3	k5	k1	k2	k7	k6	Sum
h1	18	2	14	0	0	0	1	35
h2	4	0	0	0	0	0	0	4
h3	0	0	10	0	0	0	0	10
h4	0	10	0	40	0	0	9	59
h5	0	0	0	2	12	0	5	19
h6	0	0	0	0	0	4	0	4
h7	0	0	0	0	2	0	0	2
Sum	22	12	24	42	14	4	15	133

Correspondence between clusters

hierarchical	h1	h2	h3	h4	h5	h6	h7
k-means	k4	k3	k5	k1	k2	k7	k6
Matches	84		Hit rate			63.2%	
Mismatches	49		Error rate			36.8%	

Evaluation of robustness - Bootstrap

- The bootstrap consists in repeating r times (for example r=100) the clustering, using each time
  - Either a different subset of variables
  - Or a different subset of objects
- The subset of variables is selected randomly, with resampling (i.e. the same variable can be present several times, whilst other variables are absent.
- On the images the tree is colored according to the reproducibility of the branches during a 100-iterations bootstrap.


1:1
ethanol
ethanol.vs.ref.pool
galactose
galactose.vs.ref.pool
raffinose
raffinose.vs.ref.pool
fructose.vs.ref.pool
glucose.vs.ref.pool
mannose.
glucose
sucrose
mannose.vs.ref.pool
sucrose.vs.ref.pool