Statistics Applied to Bioinformatics

Correlation analysis

Jacques van Helden

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

Mean dot product

- The dot product of two vectors is the sum of the pairwise products of the successive terms.
- The mean dot product is the average of the pairwise products of the successive terms.
- Positive contributions to the dot product:
  - When both terms are positive
  - When both terms are negative
- Negative contributions:
  - When one term is positive, and the other one positive

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

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

Converting the dot product into a dissimilarity metrics

- The dot product is a similarity metrics.
- It can take positive or negative values.
- Is is not bounded.
- The dot product can be converted into a dissimilarity metrics (dpd_{ab}) by substracting it from a constant.
  - For some applications (clustering), the dissimilarity has to be positive. The constant has thus to be adapted to the data, which is a bit tricky.

$$Dmdp_{ab} = k - mdp_{ab}$$

Covariance

- The covariance is the mean dot product of the centred variables (value minus mean).
- The covariance indicates the tendency of two variables to vary in a coordinated way.

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

Pearson's coefficient of correlation

- Pearson's correlation coefficient corresponds to a standardized covariance
  - each term of the product is divided by the standard deviation
- Where
  - a is the index of an object (e.g. gene)
  - b is the index of another object (e.g. gene)
  - i is an index of a dimension in the space of variables (e.g. a sample)
  - m_i is the mean value of the i^{th} dimension
- Note the correspondence with z-scores: computing the coefficient of correlation implicitly includes a standardization of each variable.
- By definition, the correlation is comprised between -1 and 1.
- Positive values indicate correlation, negative values anti-correlation.

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

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

$$= \frac{1}{p} \sum_{i=1}^{p} z_{ai} z_{bi} = \frac{1}{p} \mathbf{z_a} \mathbf{z_b}$$

Correlation distance

- Pearson's correlation coefficient can be converted to a distance metric by a simple operation.
  - This distance has real values comprised between 0 and 2
    - 2 indicates a perfect correlation
    - 1 indicates that there is no linear correlation between a and b
    - 0 indicates a perfect anti- correlation

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

Generalized coefficient of correlation

- Pearson correlation can be generalized by using a various types of references r_a and r_b
  - If the mean values m_a and m_b are used as references, this gives Pearson's correlation.
  - If the references are set to 0, this gives the uncentred coefficient of correlation (see next slide).
  - Other values can be used if this is justified by some particular knowledge about the data.

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

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

Uncentred correlation

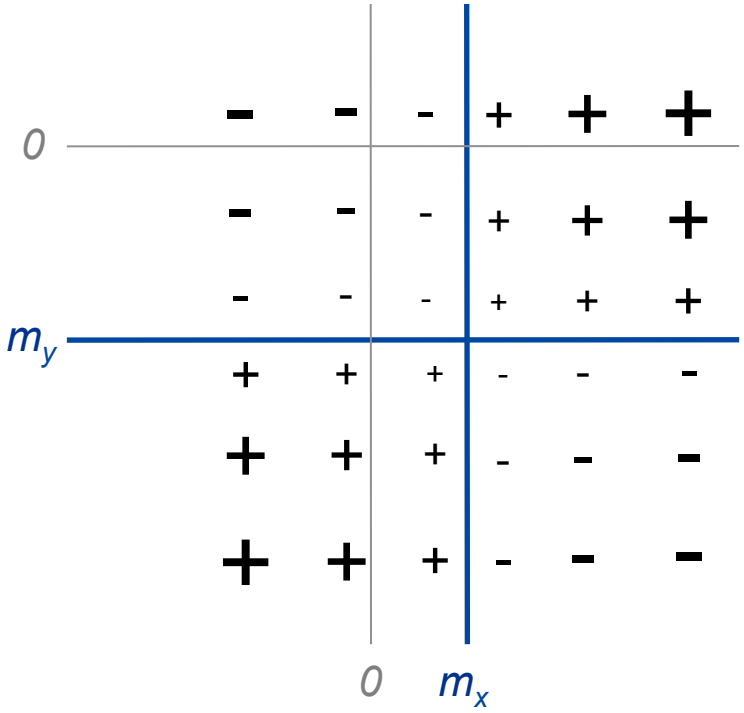
- A particular case of the generalized correlation is to take the value 0 as reference.
- This is called the uncentered correlation.
- This choice can be relevant if the object is a gene, and the value 0 represents non-regulation.

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

$$= \frac{\sum_{i=1}^{p} x_{ai}x_{bi}}{\sqrt{\sum_{j=1}^{p} x_{aj}^{\,2}}\sqrt{\sum_{j=1}^{p} x_{bj}^{\,2}}}$$

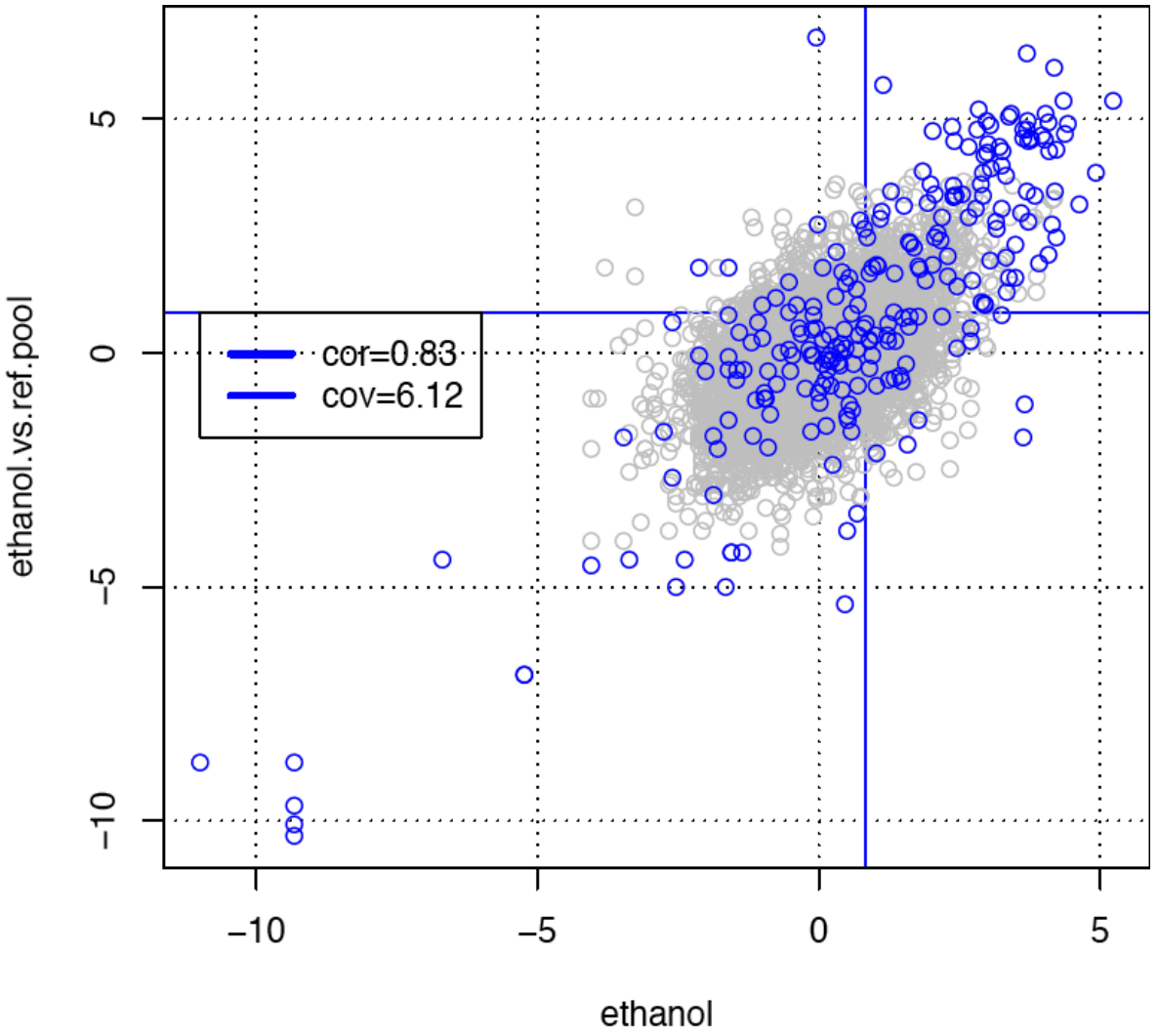
Positive and negative contributions to the coefficient of correlation

- The contribution of points will be positive or negative depending on their positions ***relative to the means*** of the respective dimensions.
- In two dimensions
  - The upper-right and lower-left quadrants (relative to the means) give a positive contribution.
  - The lower-left and upper-right quadrants (relative to the means) give a positive contribution.



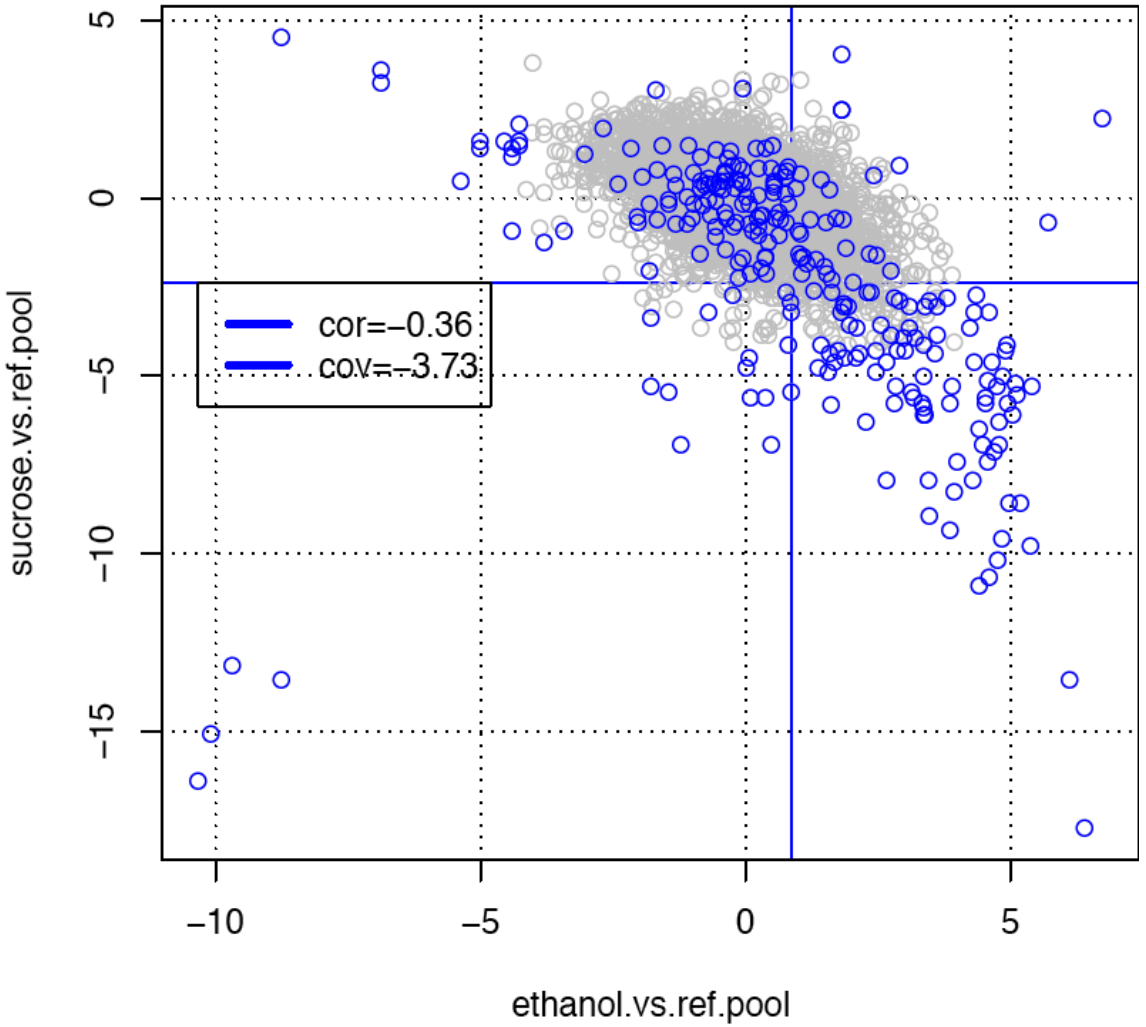
Correlation between the response of yeast transriptome to two carbon sources

- We compared two replicates of an experiment from Gasch, 2000 where ethanol is provided as carbon source.
  - In grey: all the genes
  - In blue: 269 genes showing a significant up- or down-regulation in response to at least one carbon source (13 chips).
- Most points (and in particular the most distant points) are in the upper-right and lower-left quadrants.
- There is a strong positive correlation (cor=0.83).

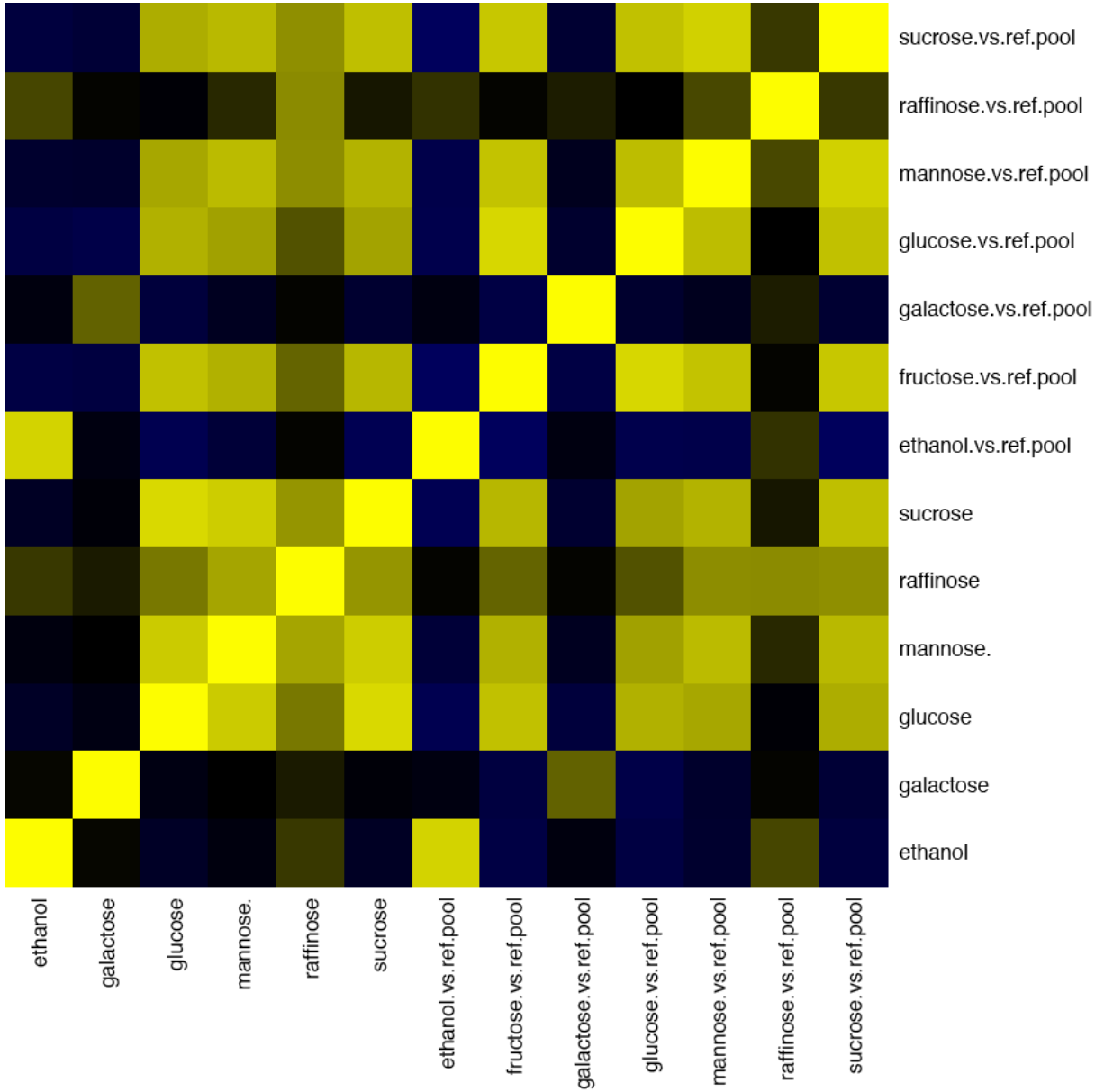


Correlation between the responses of two carbon sources

- We compared two experiments from Gasch, 2000 where either ethanol or sucrose is provided as carbon source.
  - In grey: all the genes
  - In blue: 269 genes showing a significant up- or down-regulation in response to at least one carbon source (13 chips).
- Most selected genes show an opposite behaviour : up-regulated in one condition, down-regulated in the other one.
  - Those genes (upper-left and lower-right quadrants) give negative contributions to the correlation.
- Four genes however are strongly down-regulated in both conditions.
  - Those genes (lower-left quadrant) give positive contributions to the correlation.
- The correlation is negative (cor=-0.36), but not as strong as in the previous slide.

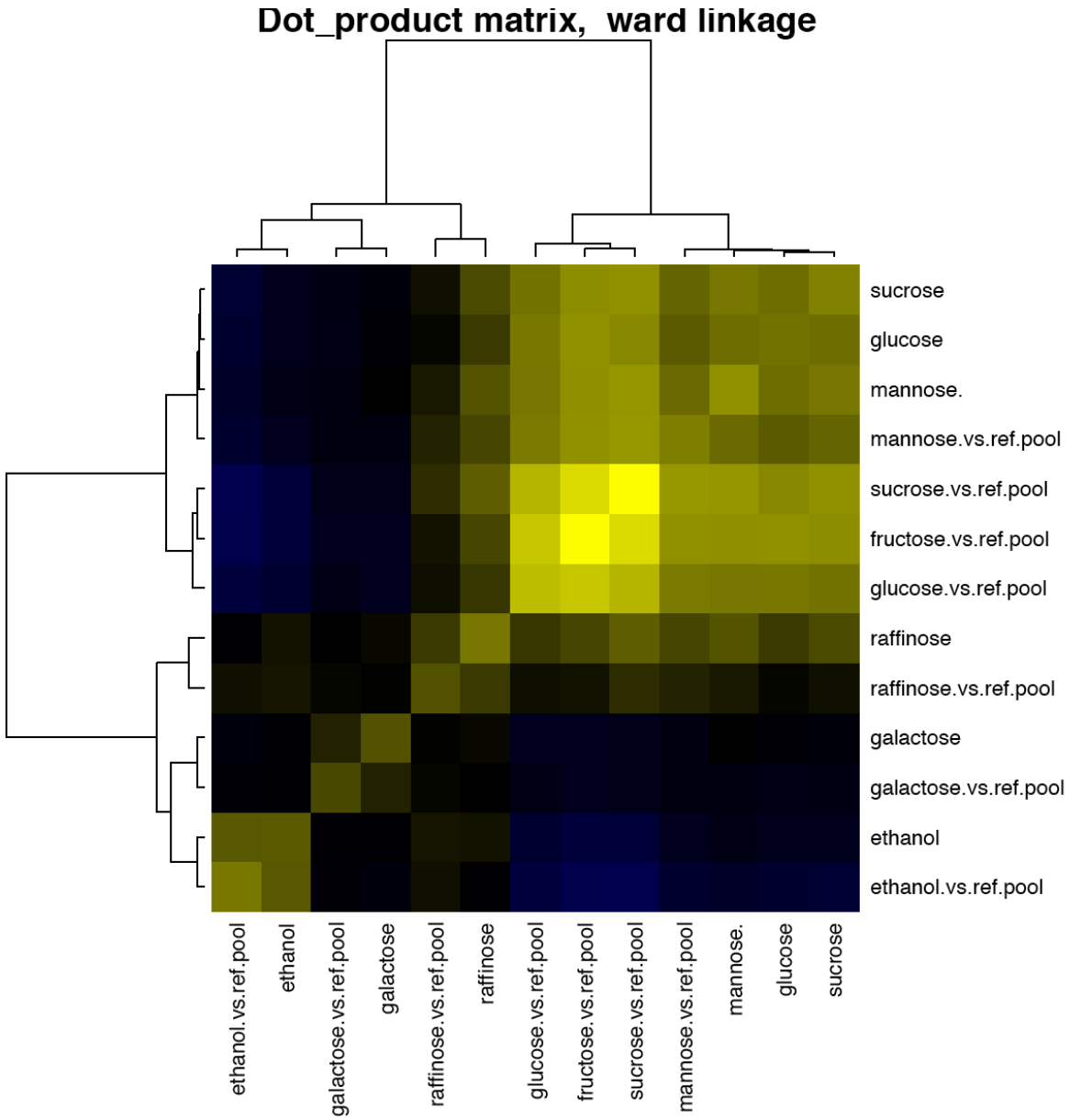


Correlation matrix



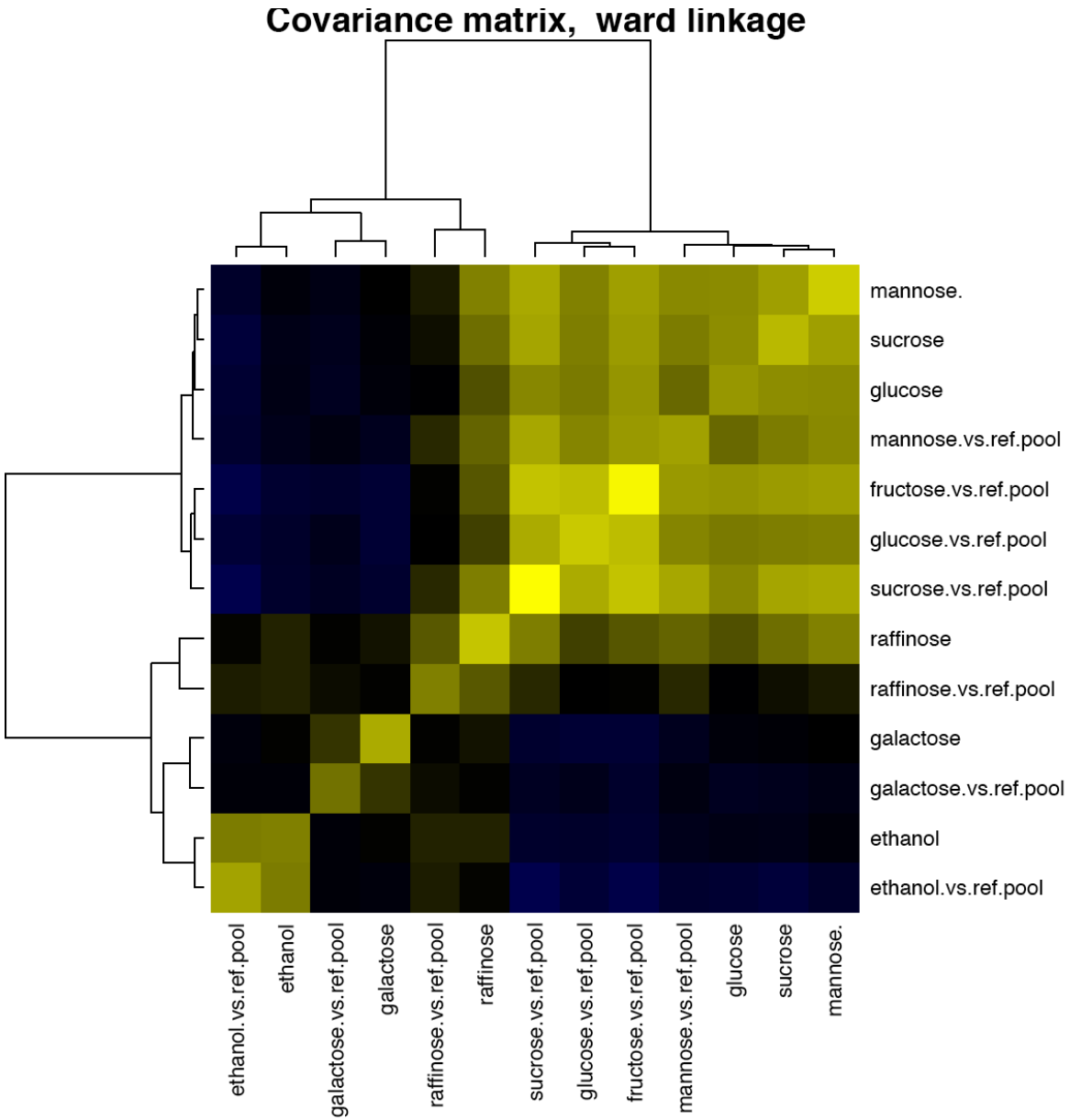
Dot product matrix - carbon sources (Gasch 2000)

- Data set: 269 genes showing a significant up- or down-regulation in response to carbon sources (Gasch, 2000)
- The matrix represents the **dot product** between each pair of conditions.
- Conditions are grouped together (clustered) according to their similarities.



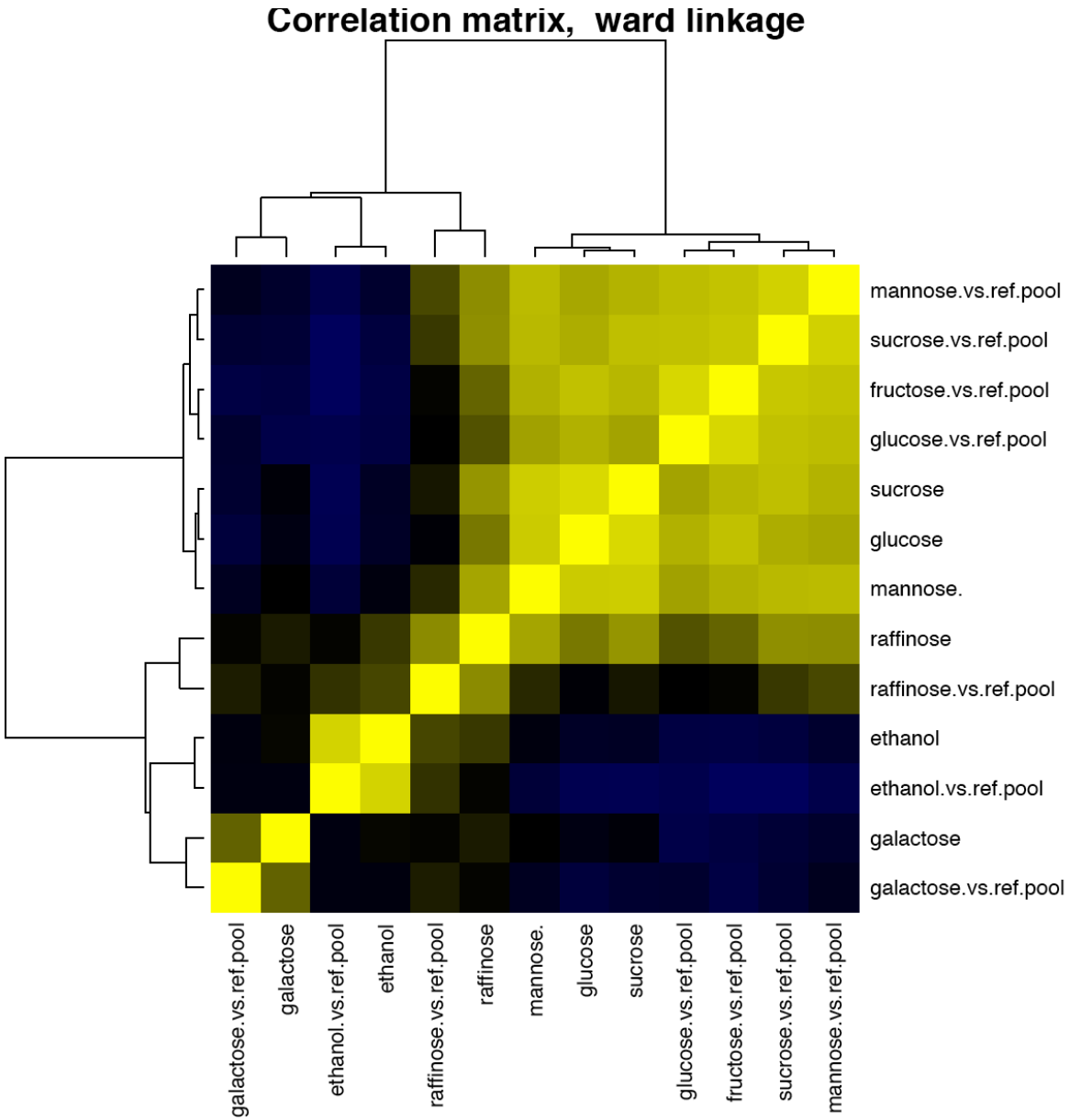
Covariance matrix - carbon sources (Gasch 2000)

- Data set: 269 genes showing a significant up- or down-regulation in response to carbon sources (Gasch, 2000)
- The matrix represents the **covariance** between each pair of conditions.
- Conditions are grouped together (clustered) according to their similarities.
- Note: the diagonal (covariance between a condition and itself) is the variance of each condition.



Correlation matrix - carbon sources (Gasch 2000)

- Data set: 269 genes showing a significant up- or down-regulation in response to carbon sources (Gasch, 2000)
- The matrix represents the **correlation** between each pair of conditions.
- Conditions are grouped together (clustered) according to their similarities.
- Note: the values on the diagonal (correlation between a condition and itself) are always 1.



Euclidian distance - carbon sources (Gasch 2000)

- Data set: 269 genes showing a significant up- or down-regulation in response to carbon sources (Gasch, 2000)
- The matrix represents the **Euclidian distance** between each pair of conditions.
- Conditions are grouped together (clustered) according to their similarities.
- Note:
  - The values on the diagonal (distance between a condition and itself) are always 0.
  - The Euclidian distance is always positive, we loose the distinction between correlation and anti-correlation.