Exploratory Data Analysis

Hubert Rehrauer

Data Generation Workflow

Experiment

↓

Omics method

↓

Raw "machine" data

↓

Raw quantitative data →

- Normalization
- Transformation
- Visualization
- Quality Control
- Outlier Detection

→

- Quantitative sample differences

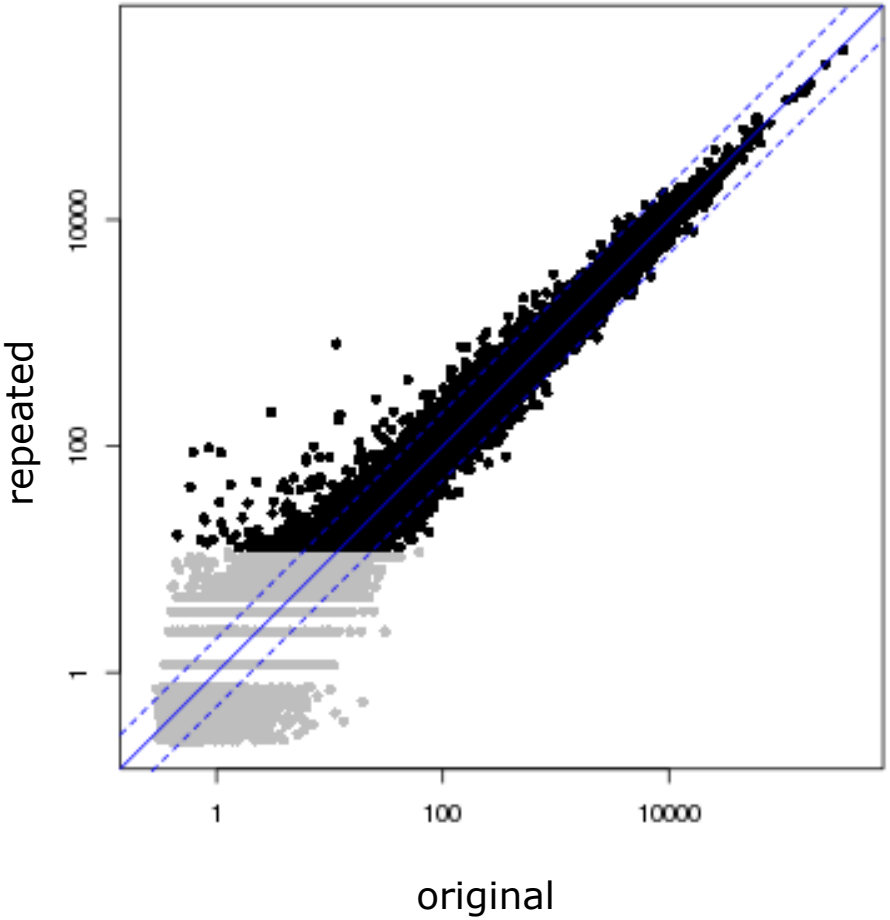
Quantitative Omics Data

- Gene expression data:
  - Quantification of relative mRNA abundance in cells/tissues
- Protein expression data
  - Quantification of relative protein abundance in cells/tissue
- Methylation status
- …

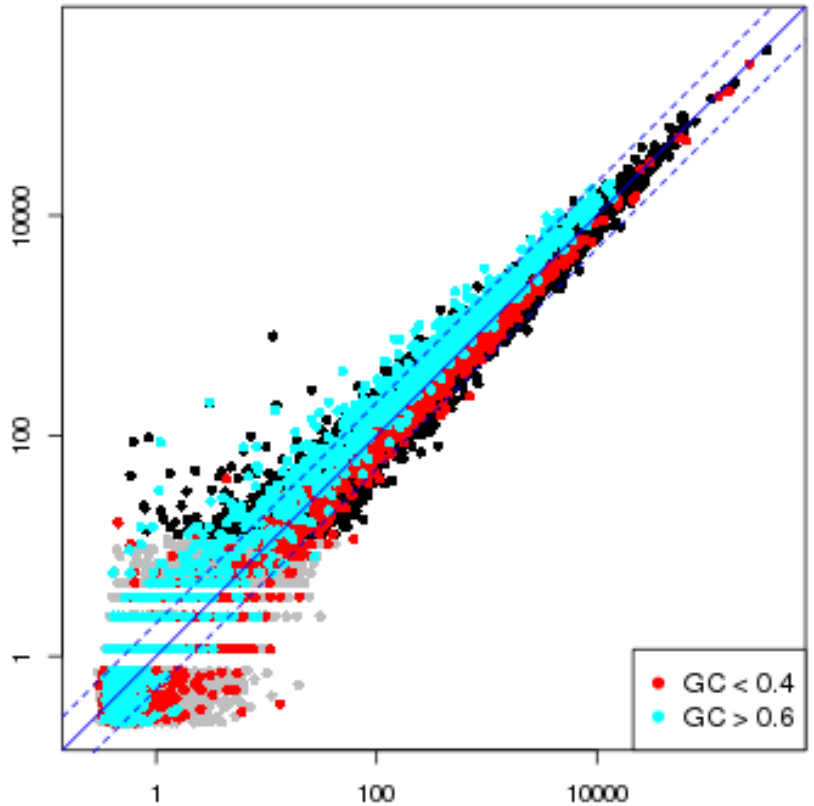
- Characteristics
  - obtained after analog signal transduction and amplification
  - not calibrated; no physical units
  - thousands or millions of measurements
  - measurements are at molecular scale

Example of Quantitative Data

Comparison of a repeated experiment



Systematic effect of GC content on quantitative values



Technical Issues

- Dynamic Range
  - ~10^4 to 10^6: largest value can be a million times higher than lowest values
  - non-linearity of measurement device
- Zero Measurements
  - additive background signal
  - zero value can have to explanations
    - failed-to-detect
    - value is truly zero
- Errors
  - Non-Gaussian noise

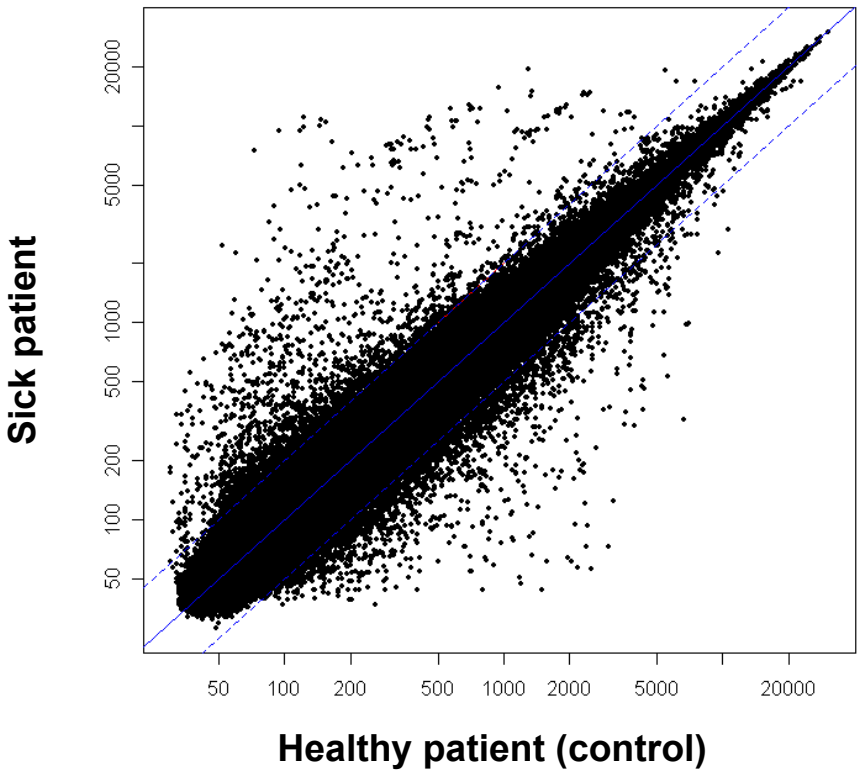
Representation as Data Matrix

Samples →

Features →

Example: Good comparability

Scatter plot showing all probes



Background signal comparable for both samples

Majority of the genes on the first diagonal

Signal response comparable (slope of the backbone of the cloud on the diagonal)

Some genes up- or down-regulated

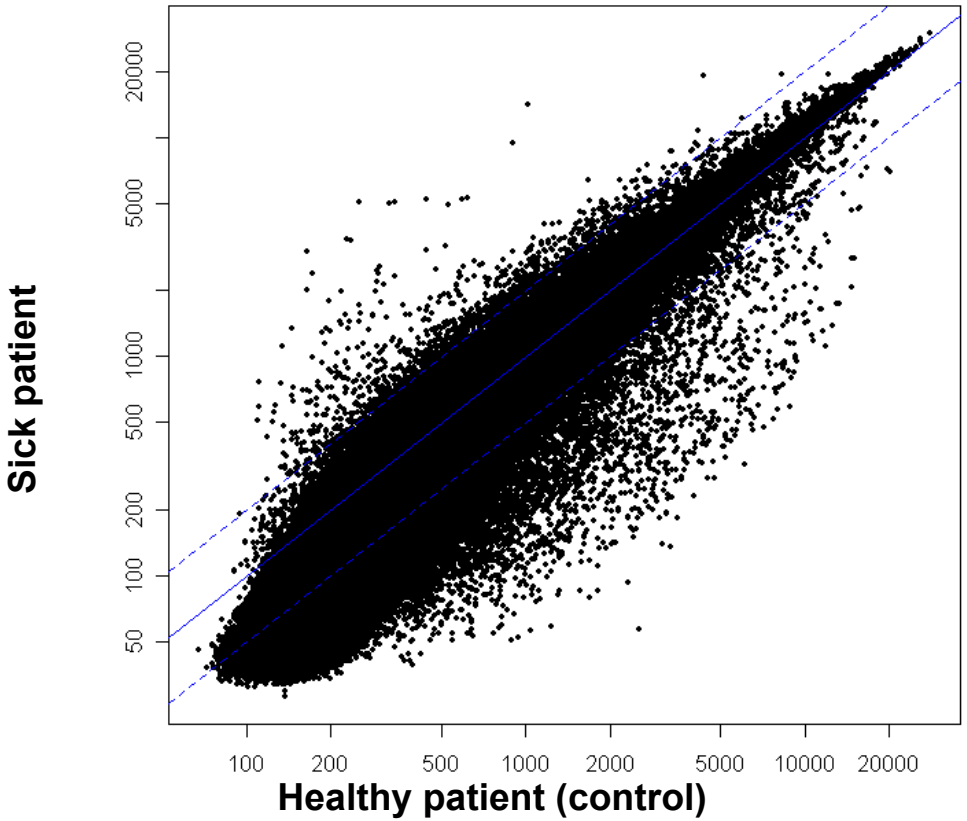
Dashed blue line indicates 2-fold changes

Data range is typically: 0 to 65000

Display is always logarithmic

Example: Bad comparability

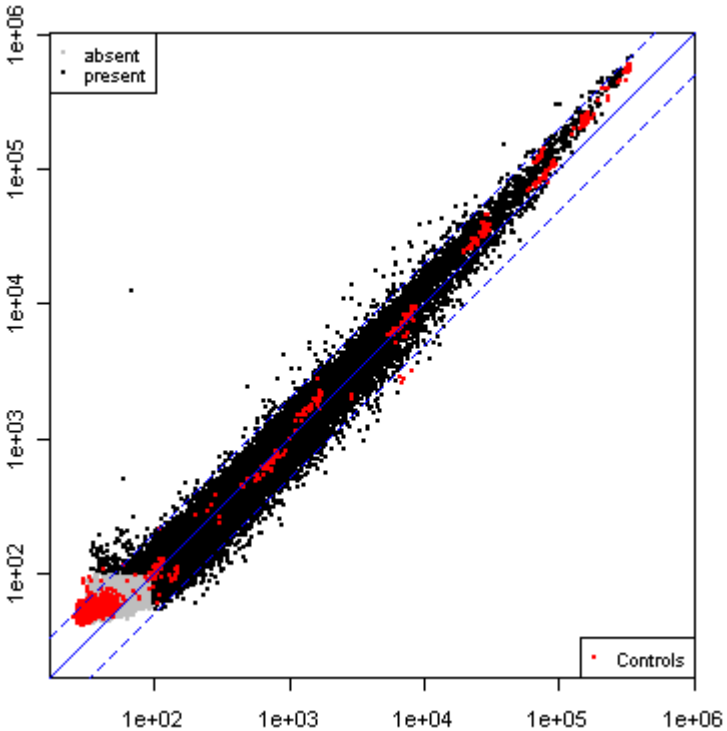
Scatter plot showing all probes



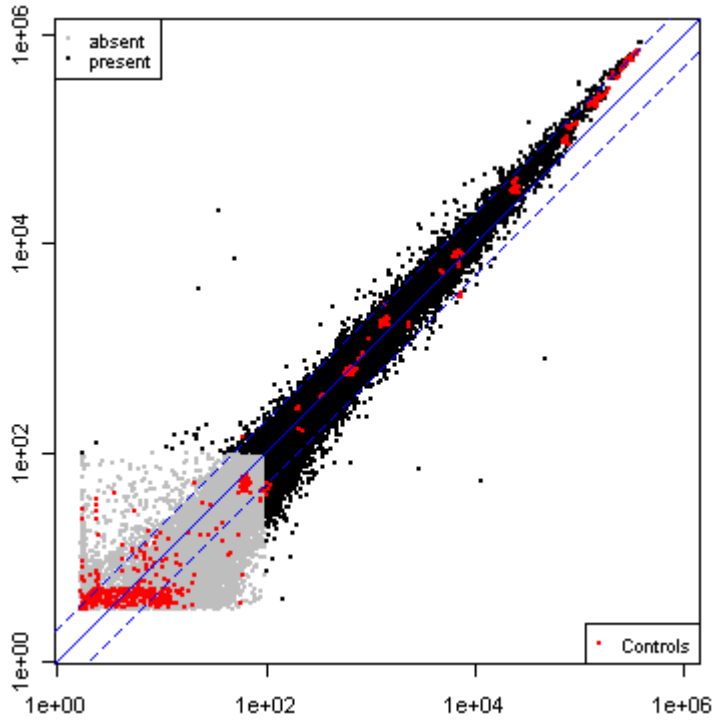
Background signal differs

Background subtraction

without background subtraction



with background subtraction



Background subtraction leads to inflated fold-changes for low signal probes
Naïve background subtraction leads to negative values which are meaningless
Background subtraction does not impact high signal values

Normalization

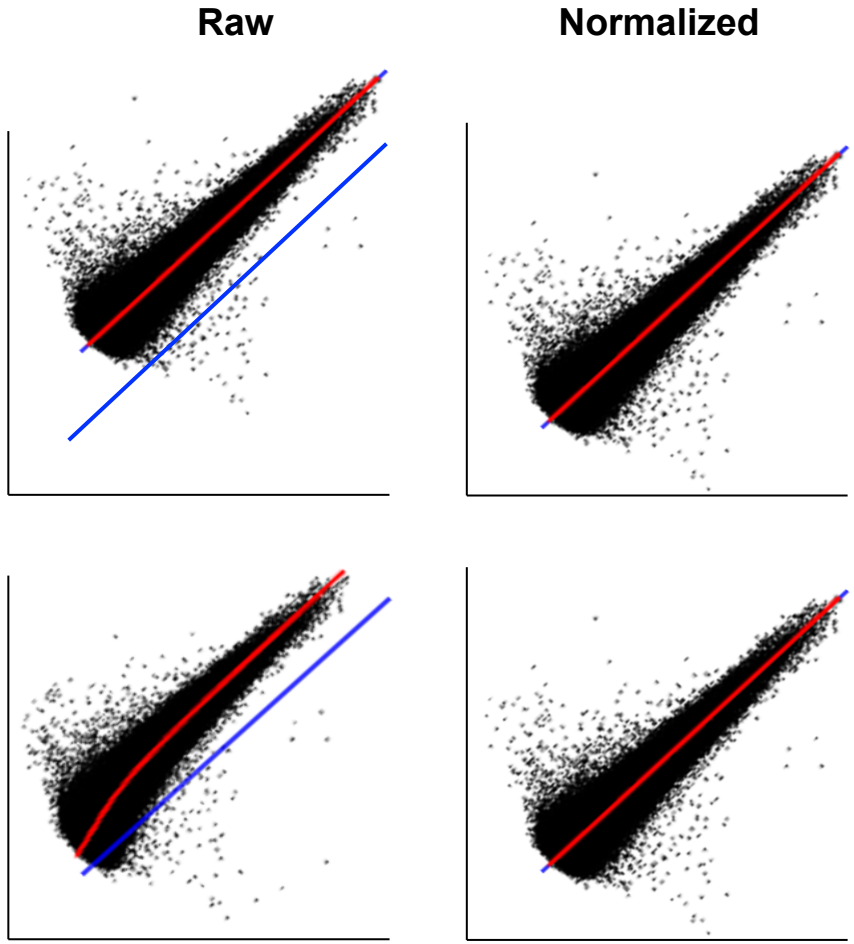
- Typically there is no calibration standard
- Normalization is always done relative to a biological control
- Assumptions:
  - majority of features is unchanged
  - minority of features has changes that go in both directions

- Assumptions may not be fulfilled under certain circumstances, e.g. comparison of free-living with symbiotically-living bacteria, …

- Note:
  - Compare with normalization when measuring a single feature
  - e.g. quantitative PCR uses one reference feature for normalizations

Normalization

- Global normalization:
  - Compute for each chip the mean/ median of log intensities
  - Scale the values so that the chip-wide mean/median is identical
- Signal dependent scaling
  - Quantile Normalization



Blue: first diagonal

Quantile Normalization

- Extends the idea of median normalization: Do not only make the 50% quantile equal, but equalize all quantiles
- Quantile normalization leads to the result that the signal histograms of all arrays are identical

- Quantile normalization is model free; no assumptions on the cause of the biases that are removed

Quantile Normalization

Sort columns of original matrix

$$\begin{bmatrix} 1 & 5 & 3 & 5 \\ 2 & 1 & 6 & 7 \\ 3 & 2 & 2 & 6 \\ 4 & 6 & 1 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix}$$

Take averages across rows

$$\begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix}$$

Set average as value for All elements in the row

$$\begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix}$$

Unsort columns of matrix to original order

$$\begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 4.5 & 4.5 & 2 \\ 3 & 2 & 6 & 4.5 \\ 4.5 & 3 & 3 & 3 \\ 6 & 6 & 2 & 6 \end{bmatrix}$$

Bolstad et al, 2003

Data Exploration

- How similar are the expression profiles of the samples?
- Do the similarities match the experimental design and the anticipated effect sizes?
- Are samples within an experimental condition more similar than across conditions?
- Are there outliers?

- Distance measures for two expression profiles **X** und **Y**:
  - 1 – correlation(**X, Y**)
  - Euclidian distance
  - …

- Most frequently the correlation is used

Correlation measure

Correlation of two profiles with p genes:

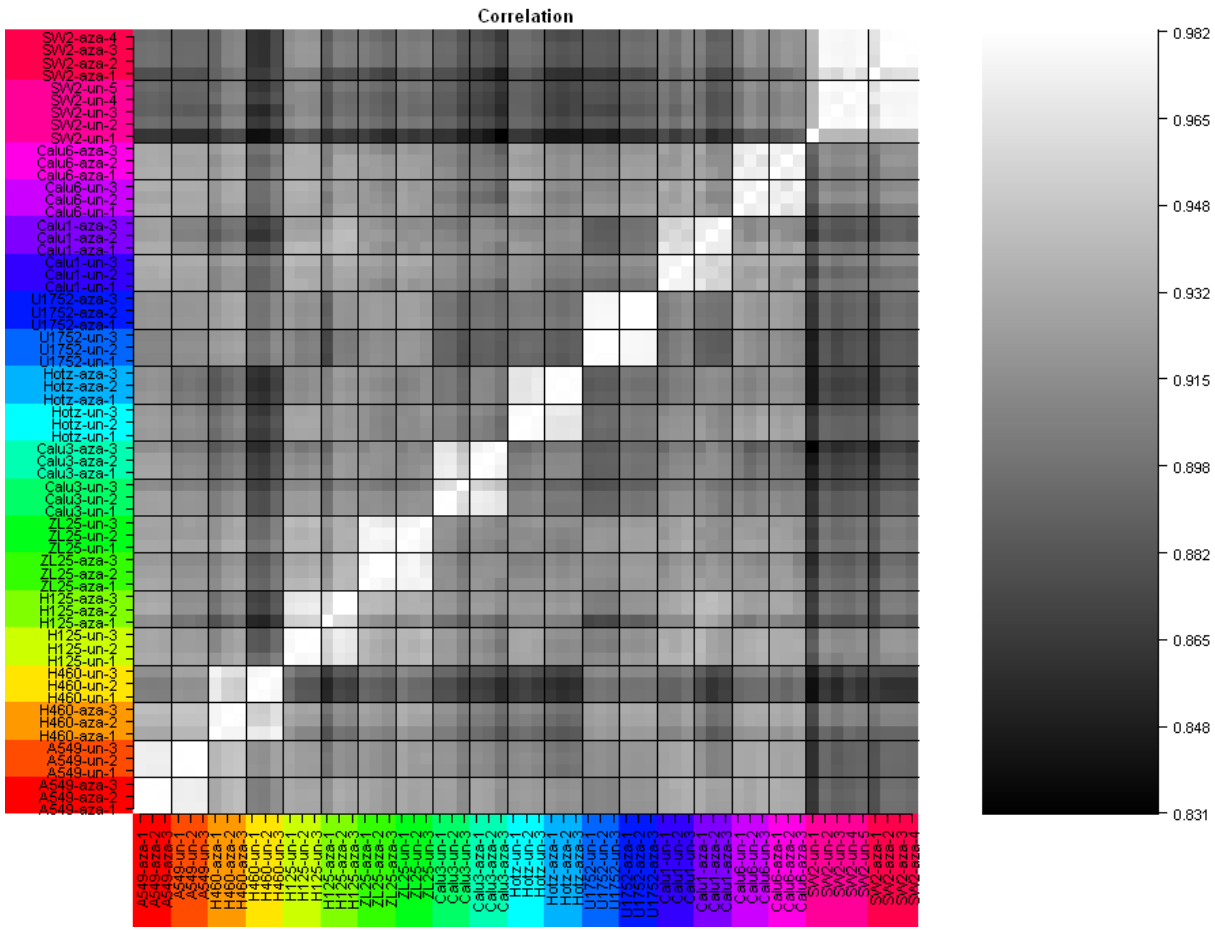
$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

averages: $\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$ and $\bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i$.

Correlation matrix for samples

The matrix shows the correlation for all sample pairs.

This gives a quick overview on the presence of outliers.



Hierarchical Clustering

Goal: Grouping of samples according to similarity

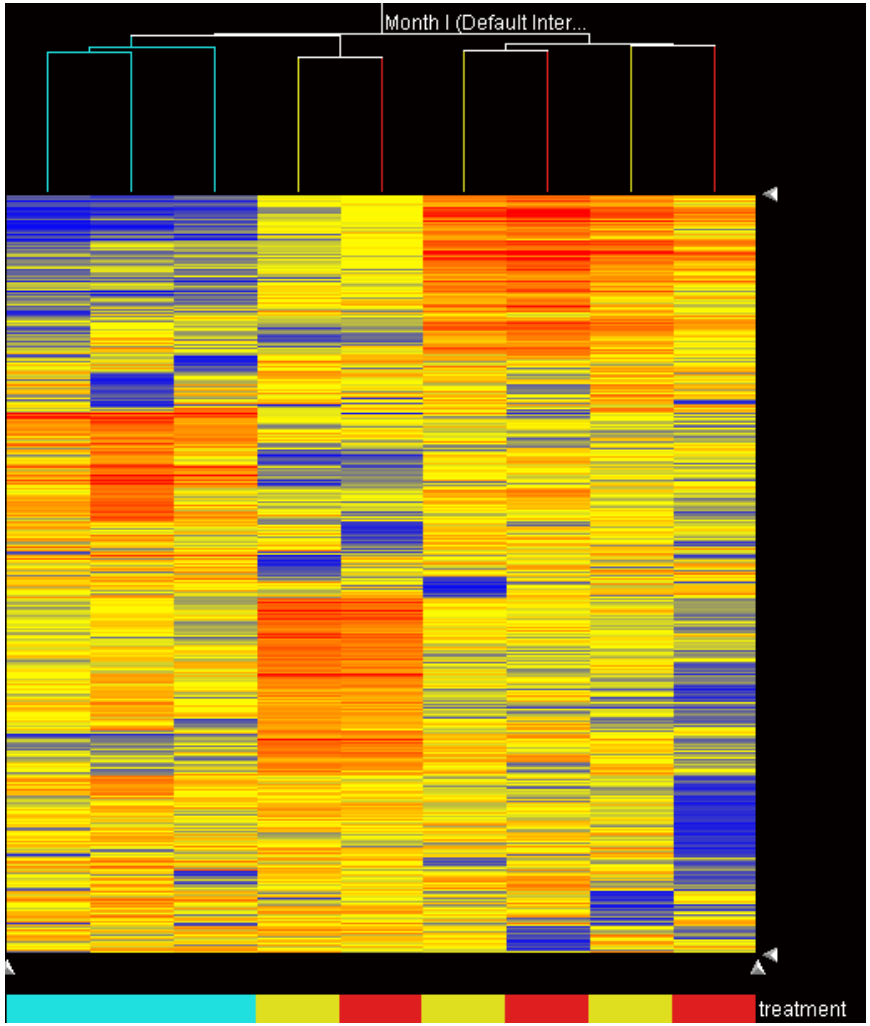
Procedure:

Initialization

- Every sample is a cluster

Iteration:

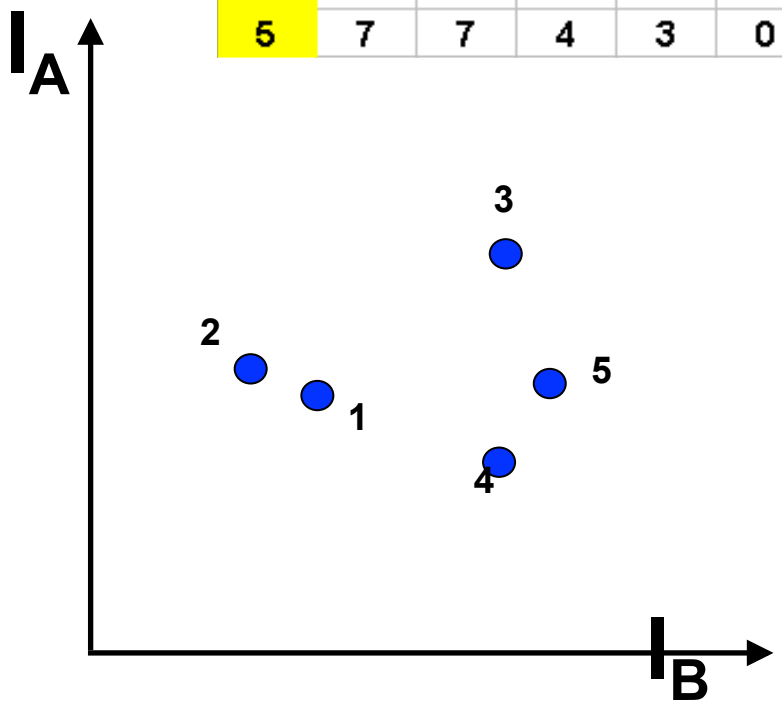
- Recursive joining of the most similar clusters



Example

Cluster distances

d(ij)	1	2	3	4	5
1	0	2	6	9	7
2	2	0	5	7	7
3	6	5	0	5	4
4	9	7	5	0	3
5	7	7	4	3	0

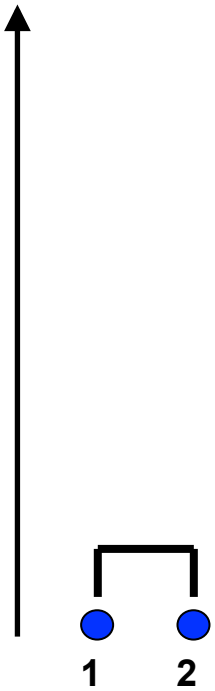


Dendrogram

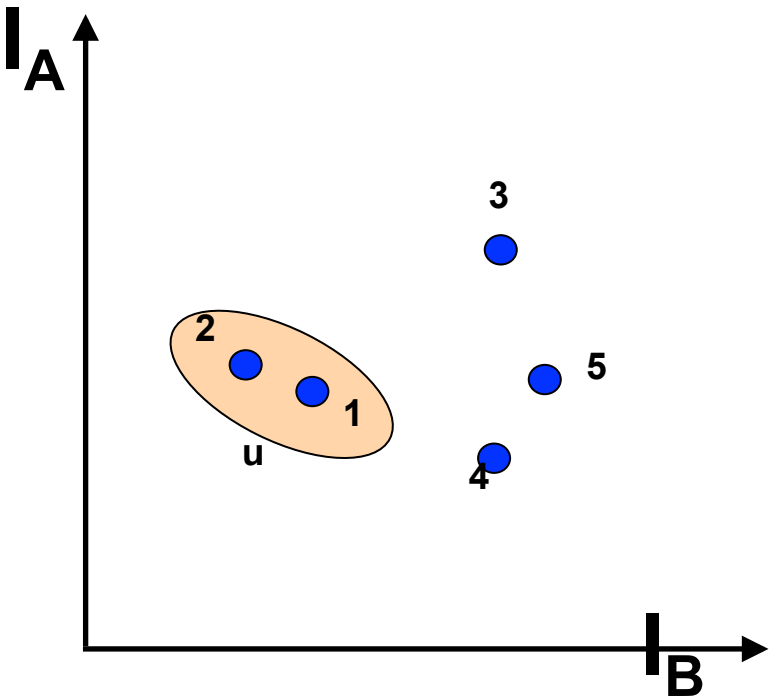
Example

Cluster distances

d(ij)	u	3	4	5
u		5.5	8	7
3	5.5	0	5	4
4	8	5	0	3
5	7	4	3	0



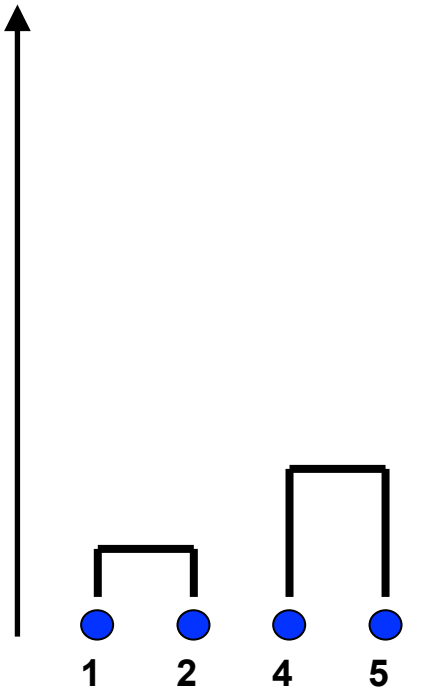
Dendrogram



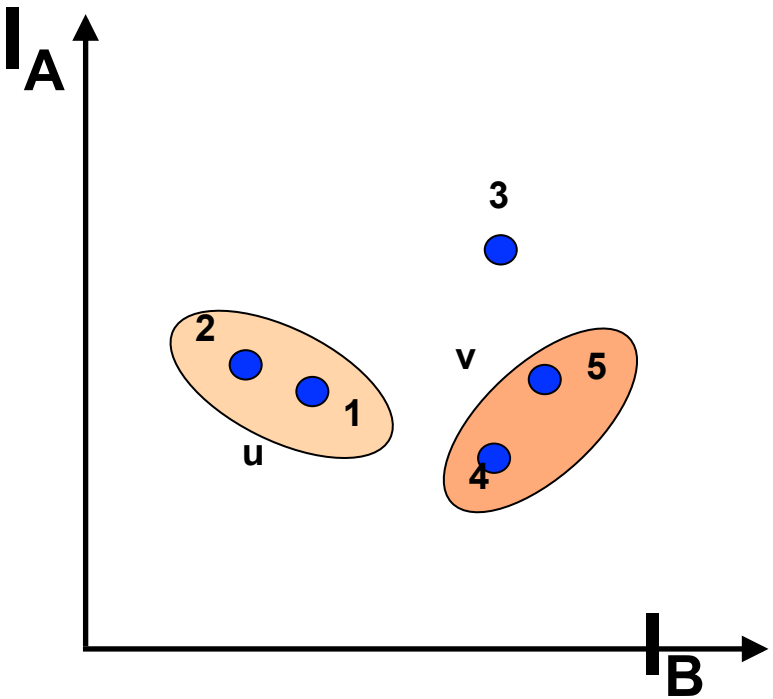
Example

Cluster distances

d(ij)	u	3	v
u	0	5.5	7.6
3	5.5	0	4.5
v	7.5	4,5	0



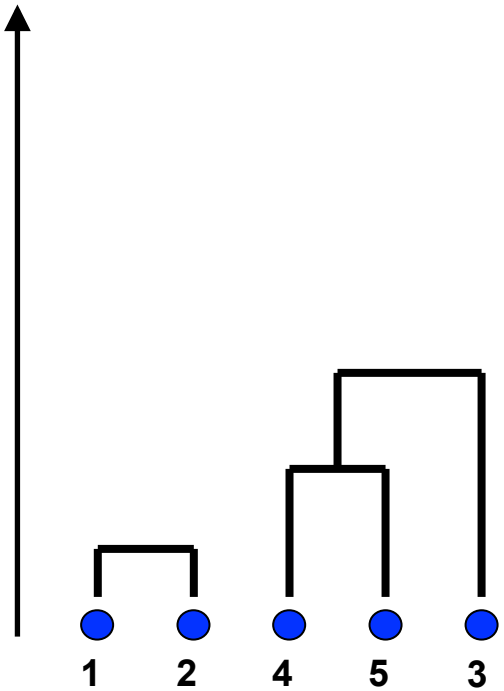
Dendrogram



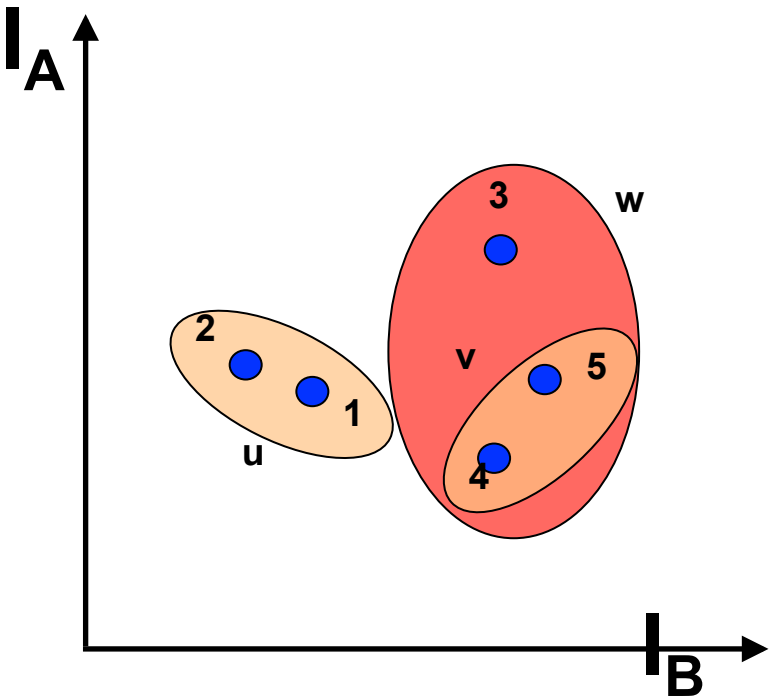
Example

Cluster distances

d(ij)	u	w
u	0	6.5
w	6.5	0

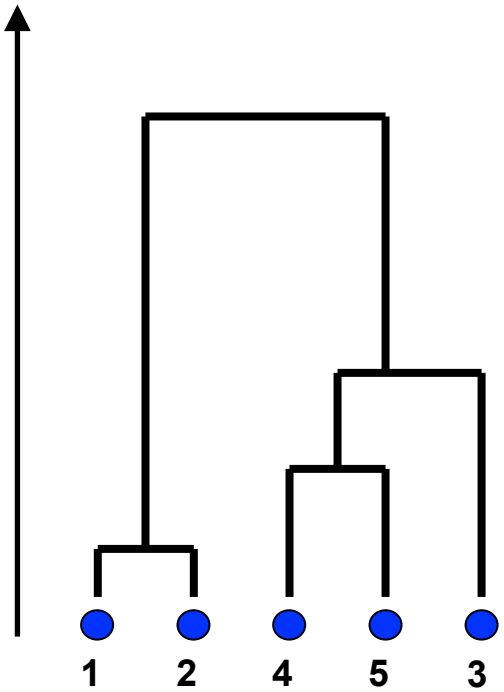


Dendrogram

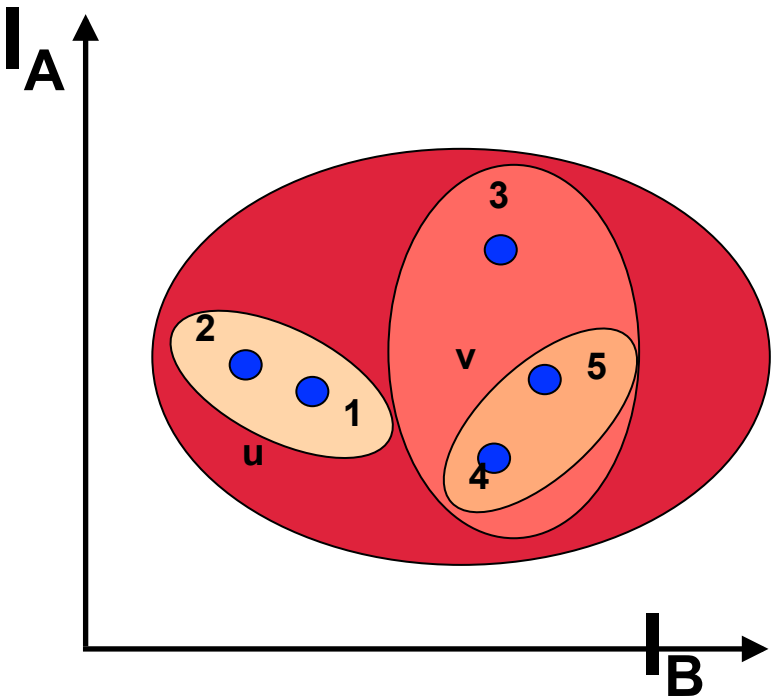


Example

Cluster distances



Dendrogram



Hierarchical Clustering: Algorithm

Algorithm

1. Compute matrix of pair-wise distances
2. Find pair with minimal distance and merge them
3. Update distance matrix
4. Continue with step 3 until a single cluster is left

Parameters:

- Distance measure for individual samples
  - For gene expression this is typically the correlation
- Distance measure for clusters of samples (linkage rule)

Hierarchical Clustering

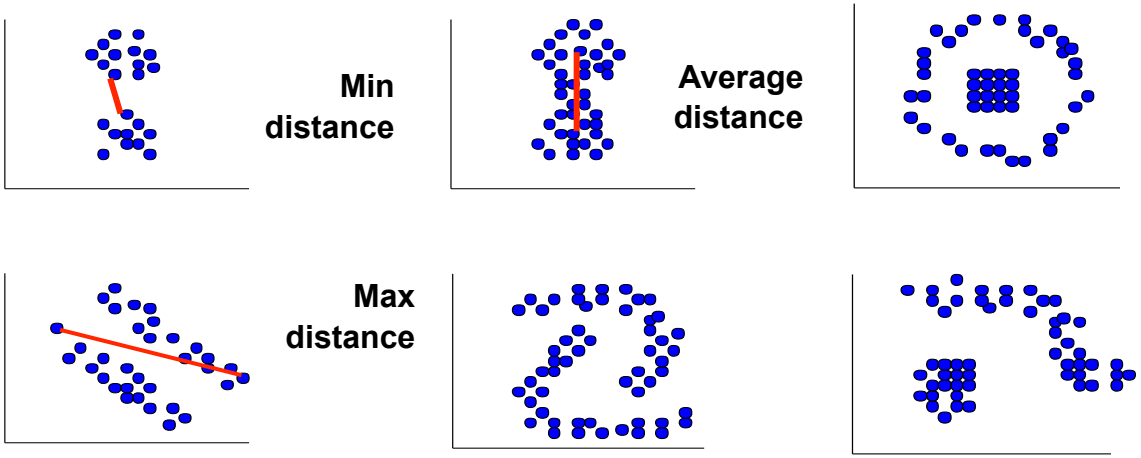
Linkage Rules:
How to compute the distance of two clusters (groups of samples)

- Single linkage: minimal distance of two members
- Complete linkage: maximal distance of two members
- Average linkage: average distance of all members
- Ward‘s linkage: minimal increase in intra-cluster variance
- …

Hint:
- The above cluster distances can be derived directly from the distance matrix of the samples
- Cluster algorithm only needs the distance matrix as input not the measurements of the individual samples

Hierarchical Clustering

Examples for cluster distances



K-means Clustering

1. Initialization: Randomly assign each sample to a cluster
2. Compute the cluster centers as the average of the assigned samples
3. Assign each sample to the closest cluster center
4. Repeat steps 3 and 4 until convergence or maximum number of steps is reached

Characteristics:

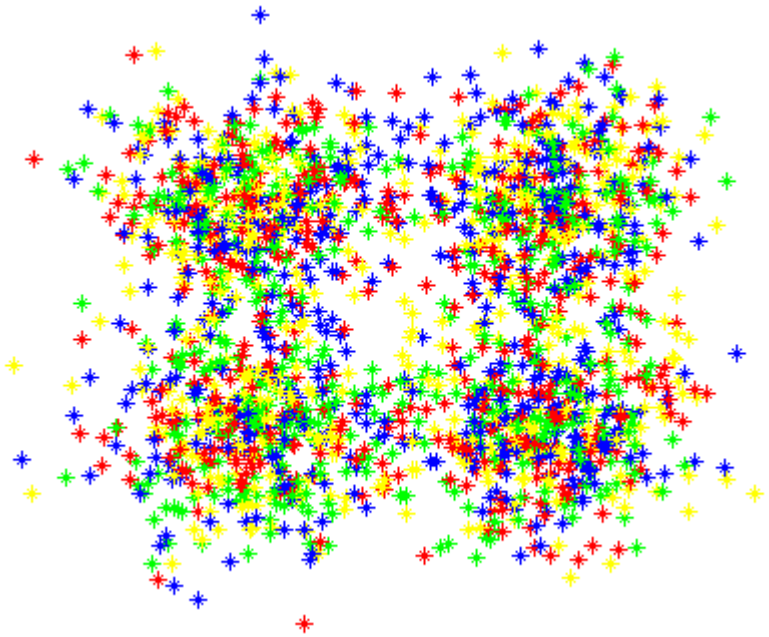
- Finds clusters that minimize intra-cluster variance

Parameters:

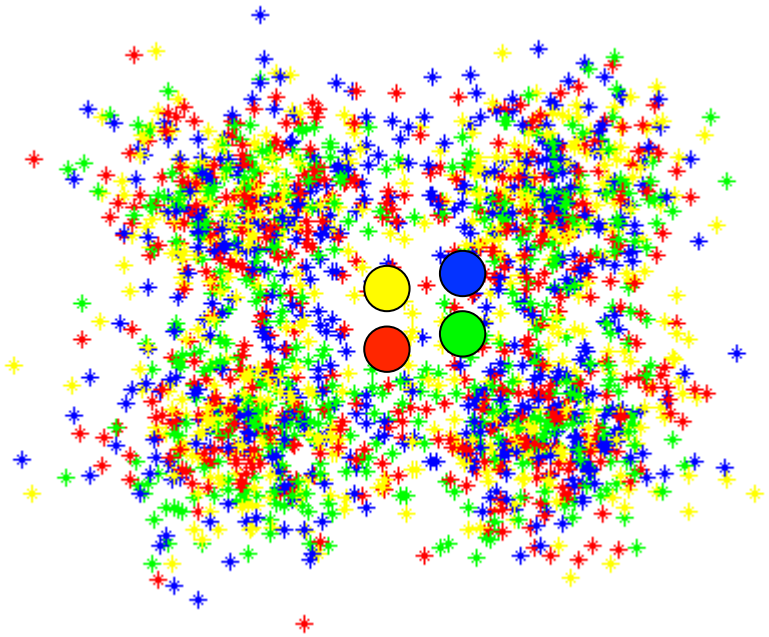
- number of clusters
- distance measure

K-means Example

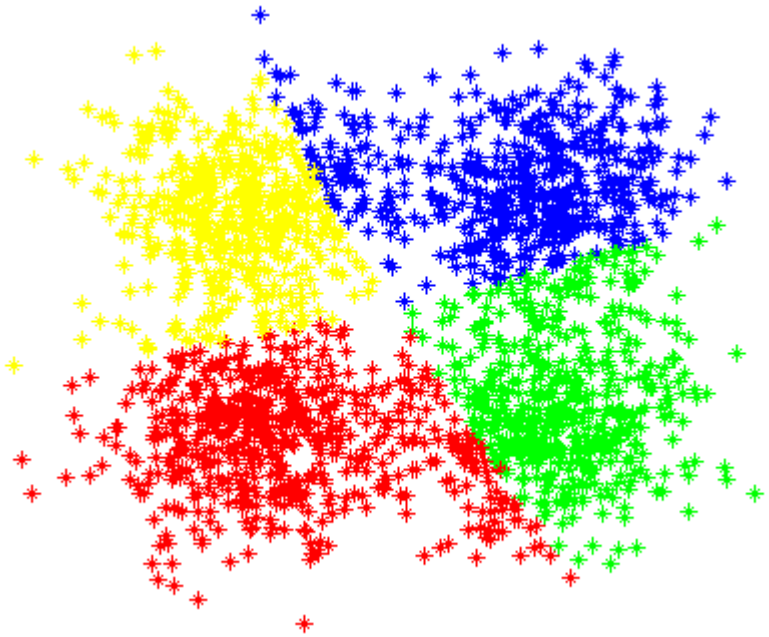
K-means: Initialization



K-means: Computing cluster centers



K-means: Reassignment of points



Comparison of Clustering Methods

Computing time:
Hierarchical clustering
- $O(n^2 \log(n))$

K-means clustering
- t: number of iterations
- k: number of clusters
- O(k t n)

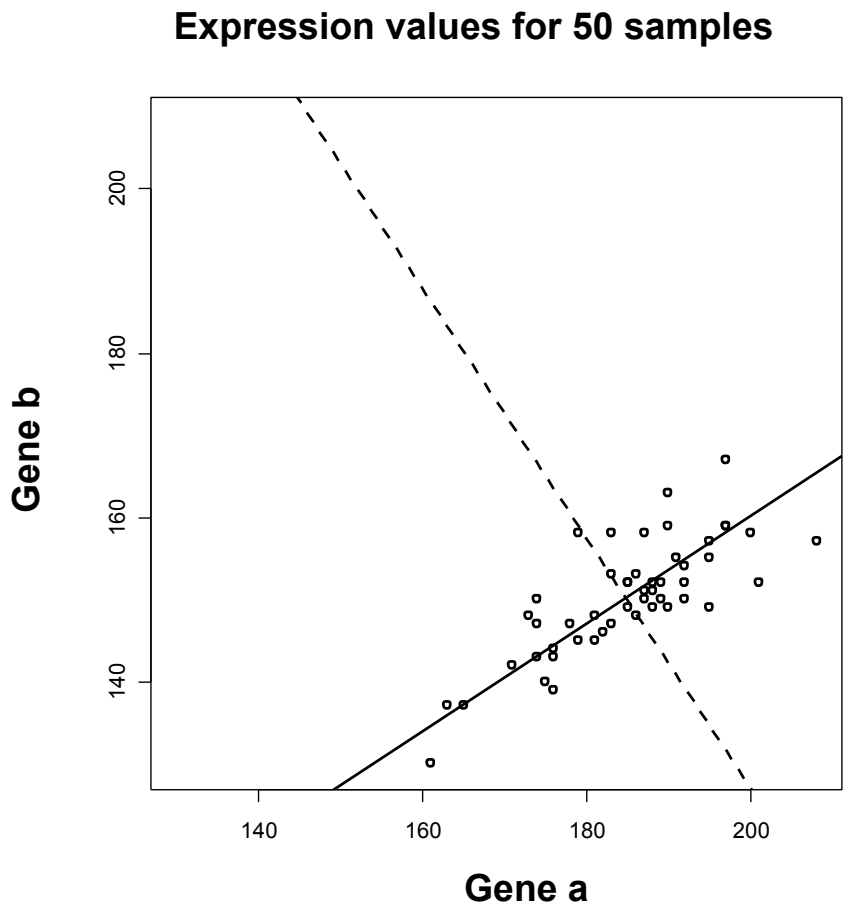
Memory requirements:
Hierarchical clustering
- O(n2)

K-means clustering
- O(kn)

Note:
When clustering large numbers of genes (>1e4, hierarchical clustering is not practical)

Principal Component Analysis

- An expression profile characterizes the state of a sample with ~25 000 genes (variables)
- Can we get a representation that uses less variables? Reduction of dimensionality?
- Yes, genes that are highly correlated can be summarized without major loss of information content
- Goal is to represent the samples in a low-dimensional space where the distance relationships of the samples are similar to the relationships in the full space



Principal Component Analysis (PCA)

Gene expression matrix: p genes, n samples

Goal: Coordinate transformation so that the relevant information is contained in the first k rows (genes)

$$Y \in R^{p \times n}$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdot & \cdot & \cdot & y_{1n} \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ y_{i1} & y_{i2} & y_{i3} & \cdot & \cdot & \cdot & y_{in} \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ y_{p1} & y_{p2} & y_{p3} & \cdot & \cdot & \cdot & y_{pn} \end{bmatrix}$$

Principal Component Analysis

Procedure:

- Center the matrix: Subtract average
- Compute covariance matrix of the centered matrix (gives an n x n Matrix)
- Compute Eigenvalues and Eigenvectors of the covariance matrix
- Sort Eigenvectors according to the magnitude of the associated Eigenvalues
- Transform to the Eigenspace
- Only show the first k variables in the Eigenspace

Principal Component Analysis

Covariance matrix:

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

Properties:
symmetric
positive definite
→ Eigenvalue decomposition possible

Principal Component Analysis

Eigenvalue problem:

- λ is an Eigenvalue of the matrix Σ, if there is a vector v such that:

$$\Sigma \vec{v} = \lambda \vec{v} \qquad \vec{v} \neq \vec{0}$$

Computation of the Eigenvalues from the determinant

$$\det(\Sigma - \lambda I) = 0$$

Principal Component Analysis

Quality of the approximation

Proportion of the explained variance:

- Let the Eigenvalues:

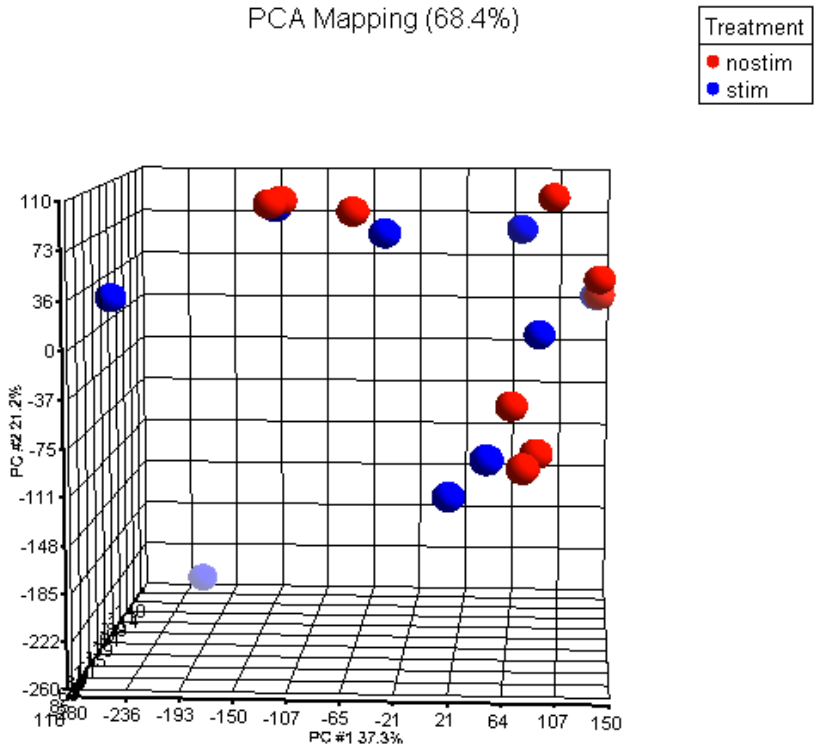
$$\lambda_i \quad i = 1, \ldots, n$$

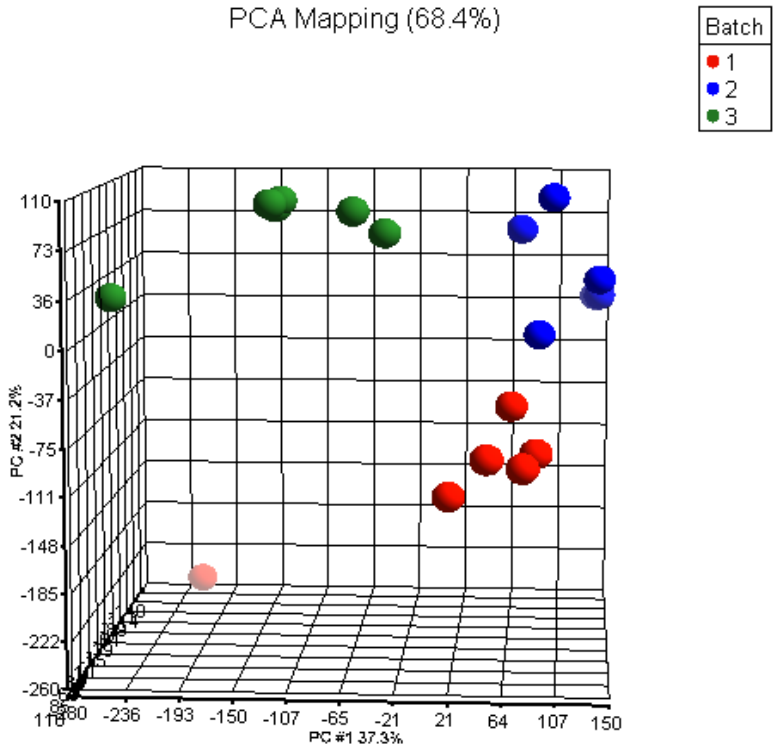
The proportion of the variance that is explained by component i is:

$$\frac{\lambda_i}{\sum_{i'} \lambda_i}$$

Example: Stimulation experiment

- Plot showing the 18 samples in the PCA coordinates
- Coloring by treatment of the samples
  - stimulated
  - not stimulated
- The samples do not separate
- There are two outliers on the left



Example: Stimulation experiment

- Coloring by batch shows that the major effect is the batch effect
- Global expression profile is majorily determined by the batch
- Stimulation leads only to a minor modulation of the expression profile