Exploratory data analysis for microarray data

Anja von Heydebreck

Max–Planck–Institute for Molecular Genetics,
Dept. Computational Molecular Biology, Berlin, Germany

heydebre@molgen.mpg.de

Outline

- Goals
- Cluster analysis
  - Distance measures
  - Clustering methods
- Projection methods
- Class discovery

Exploratory data analysis/unsupervised learning

- "Look at the data"; identify structures in the data and visualize them.

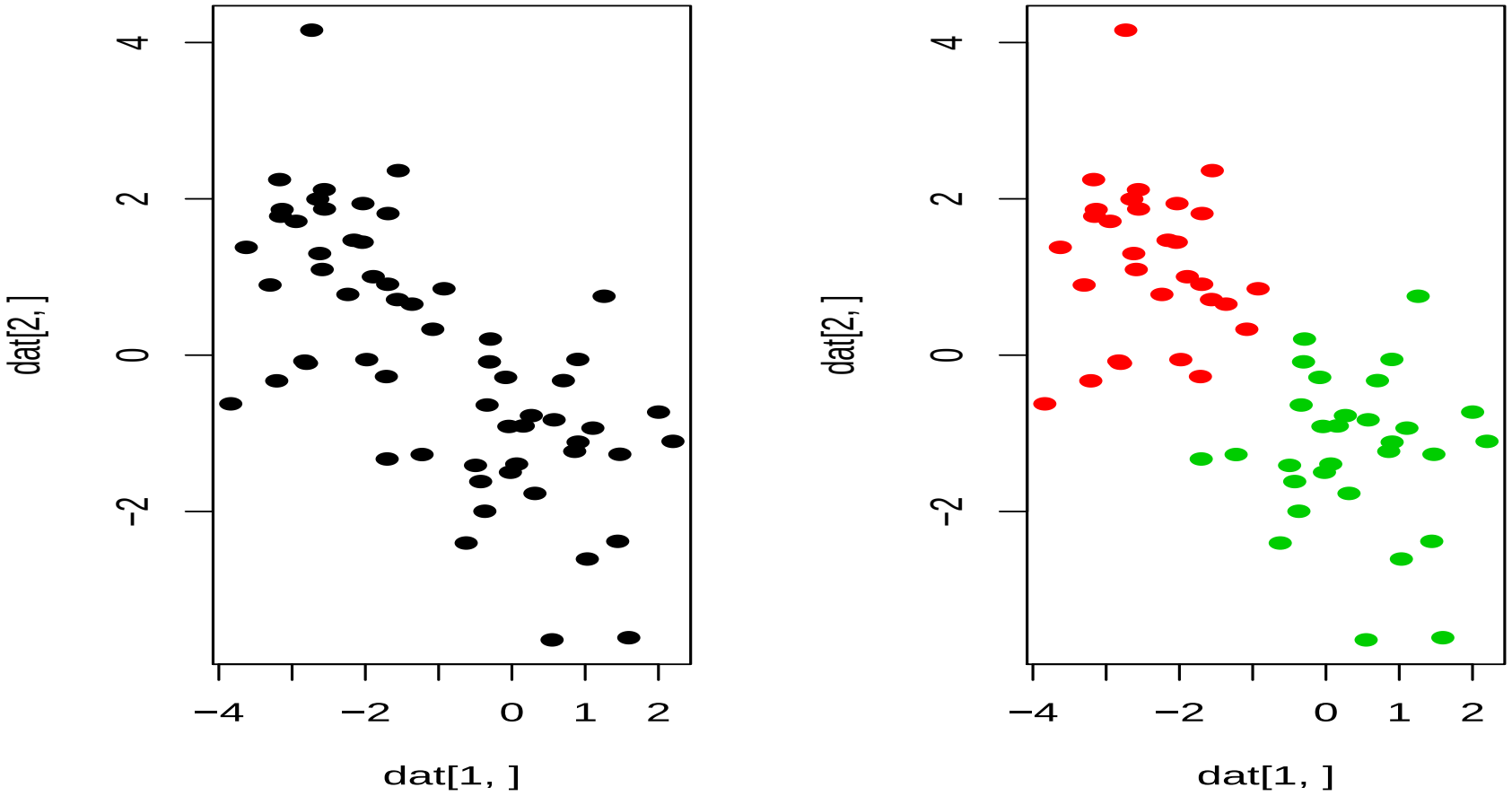
- Can we see biological/experimental parameters; are there outliers?

- Find groups of genes and/or samples sharing similarity.

- Unsupervised learning: The analysis makes no use of gene/sample annotations.

Clustering

Aim: Group objects according to their similarity.



Clustering gene expression data

- Clustering can be applied to rows (genes) and/or columns (samples/arrays) of an expression data matrix.

- Clustering may allow for reordering of the rows/columns of an expression data matrix which is appropriate for visualization (heatmap).

Clustering genes

Aims:

- identify groups of co-regulated genes
- identify typical spatial or temporal expression patterns (e.g. yeast cell cycle data)
- arrange a set of genes in a linear order which is at least not totally meaningless

Clustering samples

Aims:

- detect experimental artifacts/bad hybridizations (quality control)

- check whether samples are grouped according to known categories (meaning that these are clearly visible in terms of gene expression)

- identify new classes of biological samples (e.g. tumor subtypes)

Clustering: Distance measures

- Aim: Group objects according to their similarity.

- Clustering requires a definition of distance between the objects, quantifying a notion of (dis)similarity. After this has been specified, a clustering algorithm may be applied.

- The result of a cluster analysis may strongly depend on the chosen distance measure.

Metrics and distances

A metric d is a function satisfying:

1. non-negativity: $d(a, b) \geq 0$;
2. symmetry: $d(a, b) = d(b, a)$;
3. $d(a, a) = 0$.
4. definiteness: $d(a, b) = 0$ if and only if $a = b$;
5. triangle inequality: $d(a, b) + d(b, c) \geq d(a, c)$.

A function only satisfying 1.-3. is called a distance.

Distance measures: Examples

Vectors $\boldsymbol{x} = (x_1, \ldots, x_n), \boldsymbol{y} = (y_1, \ldots, y_n)$

- Euclidean distance: $d_M(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Manhattan distance: $d_E(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} |x_i - y_i|$
- One minus Pearson correlation:

$$d_C(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{1/2}(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{1/2}}$$

Distance measures/standardization

❍ The correlation distance is invariant wrt shifting and scaling of its arguments:

$$d_C(\boldsymbol{x}, \boldsymbol{y}) = d_C(\boldsymbol{x}, a\boldsymbol{y} + b), a > 0.$$

❍ One may apply standardization to observations or variables:

$$\boldsymbol{x} \mapsto \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}}{\sigma(\boldsymbol{x})},$$

where $\sigma(\boldsymbol{x})$ is the standard deviation of $\boldsymbol{x}$.

❍ The correlation distance and the Euclidean distance between standardized vectors are closely related:

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{2n d_C(\boldsymbol{x}, \boldsymbol{y})}.$$

Distances between clusters

Extend a distance measure d to a measure of distance between clusters.

❍ Single linkage The distance between two clusters is the minimal distance between two objects, one from each cluster.

❍ Average linkage The distance between two clusters is the average of the pairwise distance between members of the two clusters.

❍ Complete linkage The distance between two clusters is the maximum of the distances between two objects, one from each cluster.

❍ Centroid linkage The distance between two clusters is the distance between their *centroids*.

Hierarchical clustering

- Build a cluster tree/dendrogram, starting from the individual objects as clusters.

- In each step, merge the two clusters with the minimum distance between them - using one of the above linkage principles.

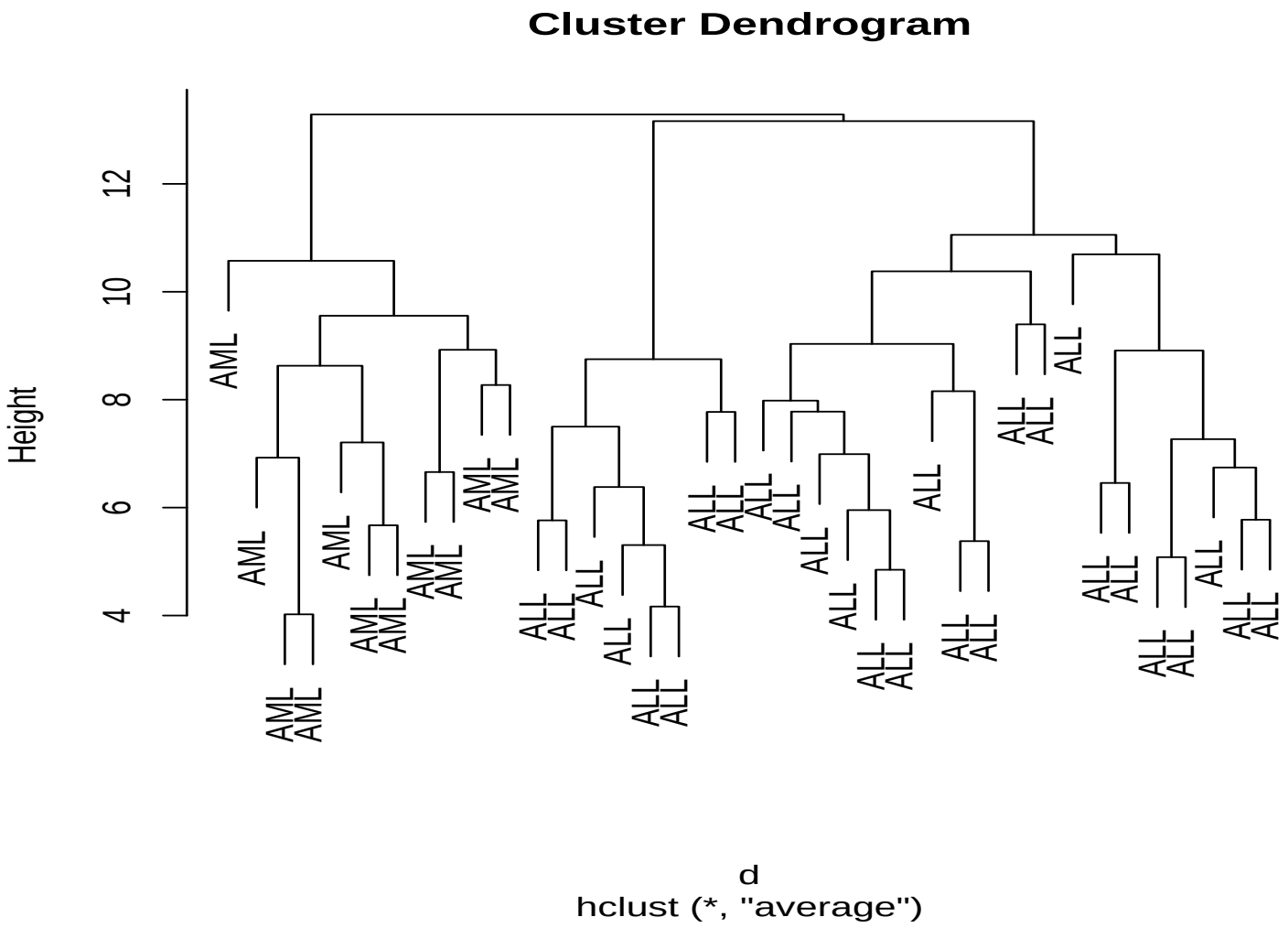
- Continue until everything is in one cluster.

- If you want a partition of the set of objects, cut the tree at a certain height.

- R function `hclust` in package `mva`.

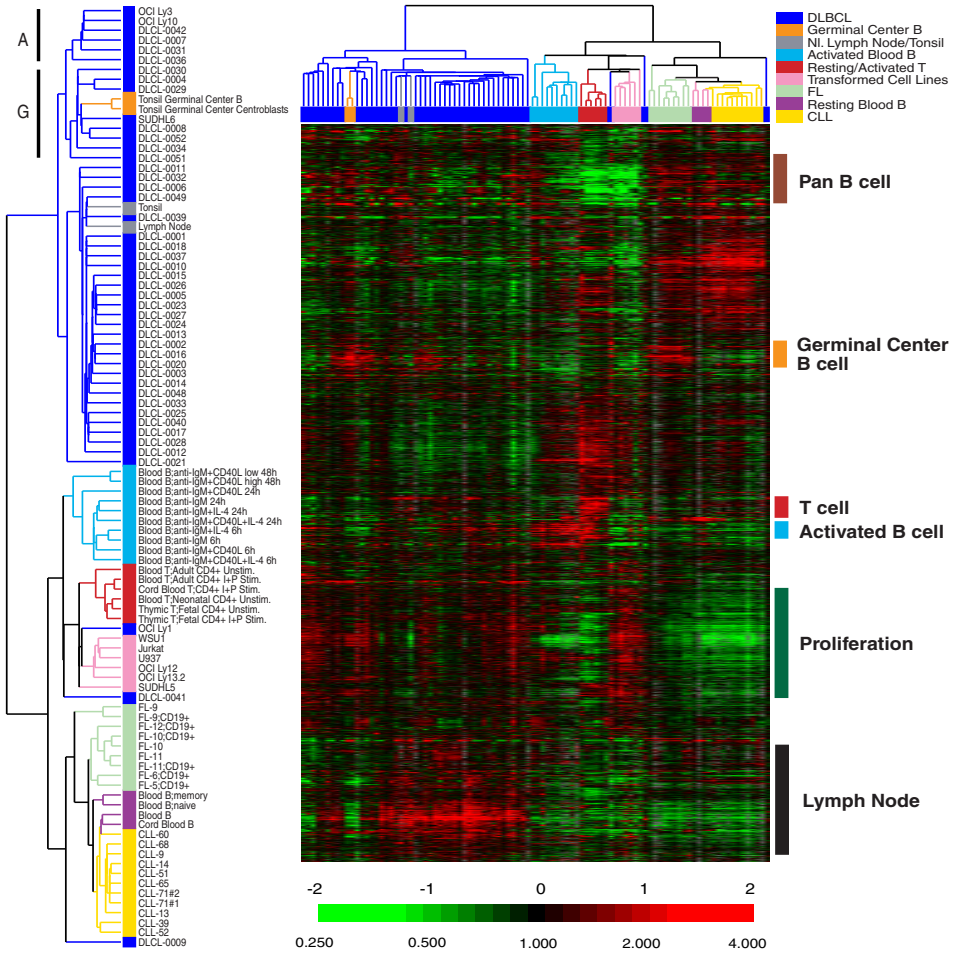
Hierarchical clustering, example

Golub data, 150 genes with highest variance



Example: Clustering of rows and columns

Alizadeh et al.(2000): Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature.



k-means clustering

❍ User specifies the number k of desired clusters. Input: Objects given as vectors in n-dimensional space (Euclidean distance is used).

❍ For an initial choice of k cluster centers, each object is assigned to the closest of the centers.

❍ The centroids of the obtained clusters are taken as new cluster centers.

❍ This procedure is iterated until convergence.

How many clusters?

❍ Many methods require the user to specify the number of clusters. Generally it is not clear which number is appropriate for the data at hand.

❍ Several authors have proposed criteria for determining the number of clusters, see Dudoit and Fridlyand 2002.

❍ Sometimes there may not be a clear answer to this question - there may be a hierarchy of clusters.

Which scale, which distance measure to use for clustering?

❍ Data should be normalized and transformed to an appropriate scale before clustering ($\log$ or the generalized $\log$ resulting from variance stabilization (R package `vsn`)).

❍ Clustering genes: Standardization of gene vectors or the use of the correlation distance is useful when looking for patterns of relative changes - independent of their magnitude.

❍ Clustering samples: Standardizing genes gives relatively smaller weight for genes with high variance across the samples - not generally clear whether this is desirable.

❍ Gene filtering (based on intensity/variability) may be reasonable - also for computational reasons.

Some remarks on clustering

- A clustering algorithm will always yield clusters, whether the data are organized in clusters or not.

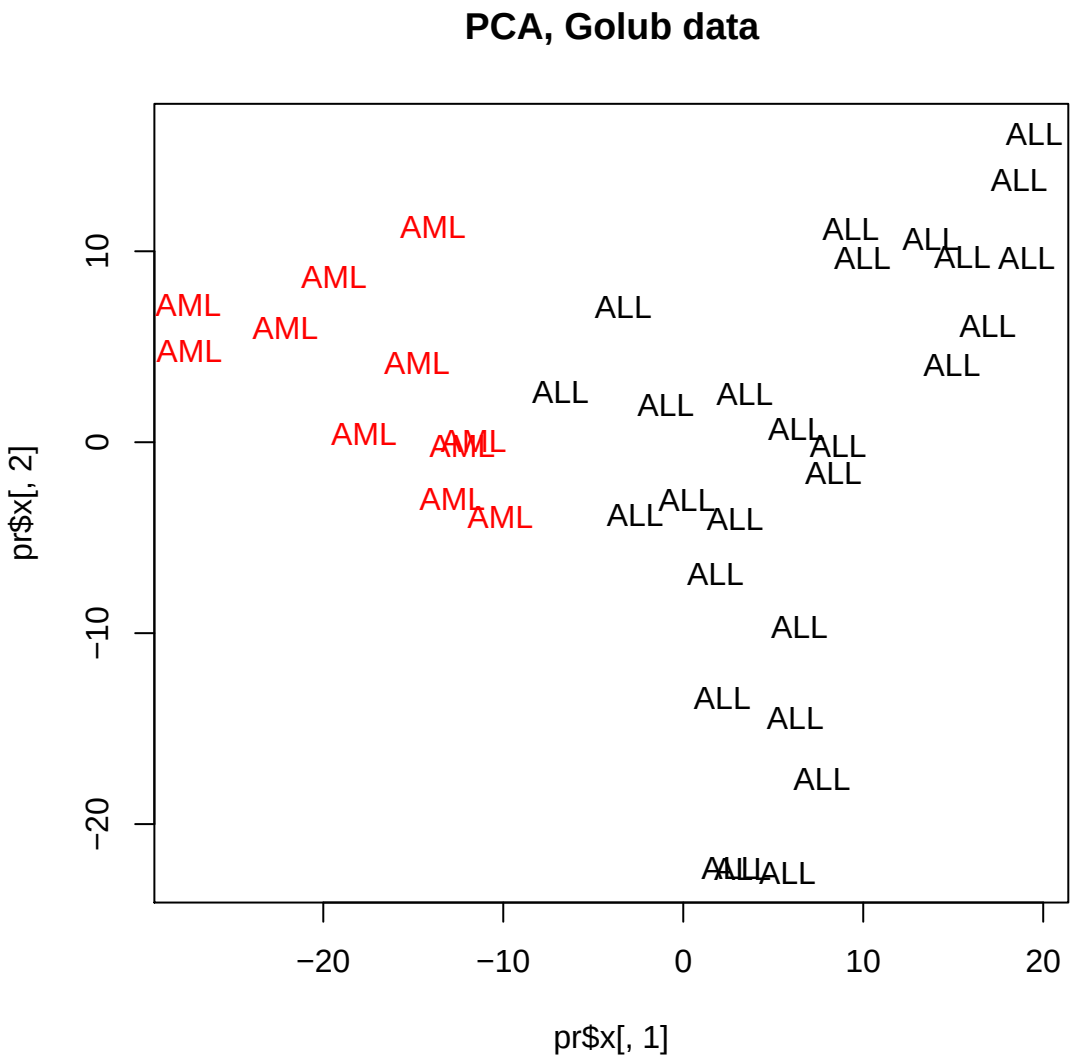
- The bootstrap may be used to assess the variability of a clustering (Kerr/Churchill 2001, Pollard/van der Laan 2002).

- If a class distinction is not visible in cluster analysis, it may still be accessible for supervised methods (e.g. classification).

Projection methods

- Map the rows and/or columns of the data matrix to a plane such that similar rows/columns are located close to each other.

- Different methods (principal component analysis, multidimensional scaling, correspondence analysis) use different notions of similarity.



Principal component analysis

❍ Imagine k observations (e.g. tissue samples) as points in n-dimensional space (here: n is the number of genes).

❍ Aim: Dimension reduction while retaining as much of the variation in the data as possible.

❍ Principal component analysis identifies the direction in this space with maximal variance (of the observations projected onto it).

❍ This gives the first principal component (PC). The $i+1$st PC is the direction with maximal variance among those orthogonal to the first i PCs.

❍ The data projected onto the first PCs may then be visualized in scatterplots.

Principal component analysis

❍ PCA can be explained in terms of the eigenvalue decomposition of the covariance/correlation matrix Σ:

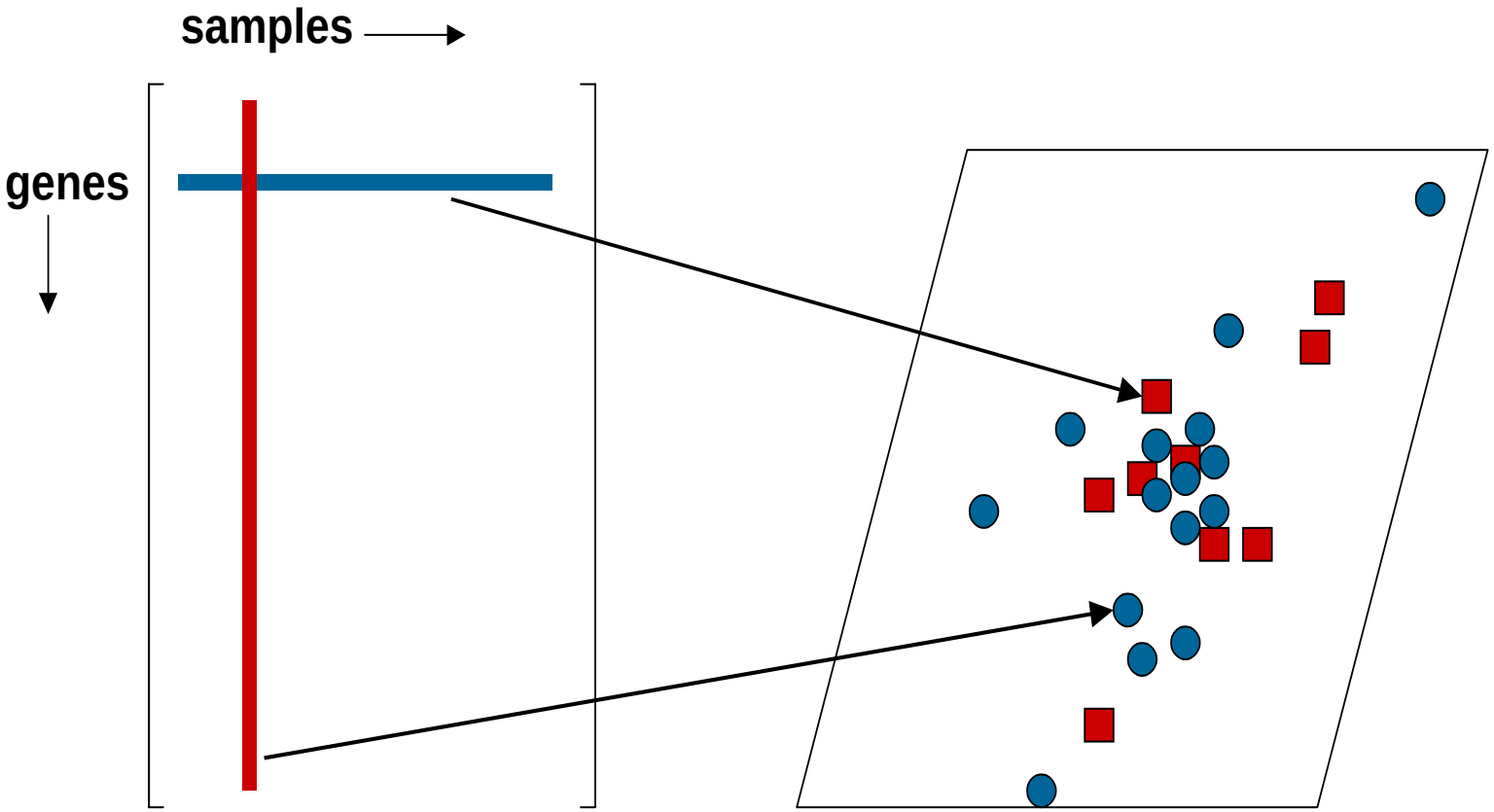
$$\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^{\boldsymbol{t}},$$

where the columns of $\boldsymbol{S}$ are the eigenvectors of Σ (the principal components), and $\boldsymbol{\Lambda}$ is the diagonal matrix with the eigenvalues (the variances of the principal components).

❍ Use of the correlation matrix instead of the covariance matrix amounts to standardizing variables (genes).

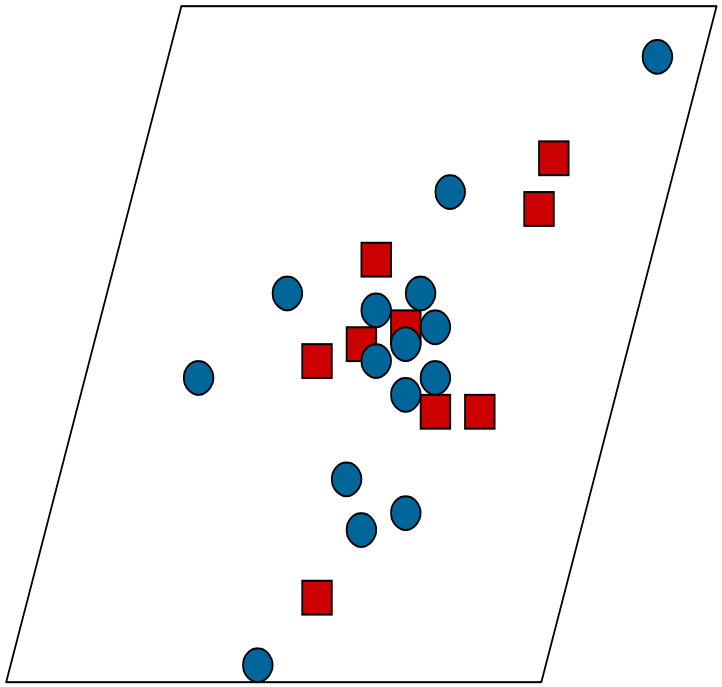
❍ R function `prcomp` in package `mva`.

Correspondence analysis: Projection onto plane



Correspondence analysis: Properties of projection

- Similar row/column profiles (small χ^2-distance) are projected close to each other.
- A gene with positive/negative association with a sample will lie in the same/opposite direction from the centroid.



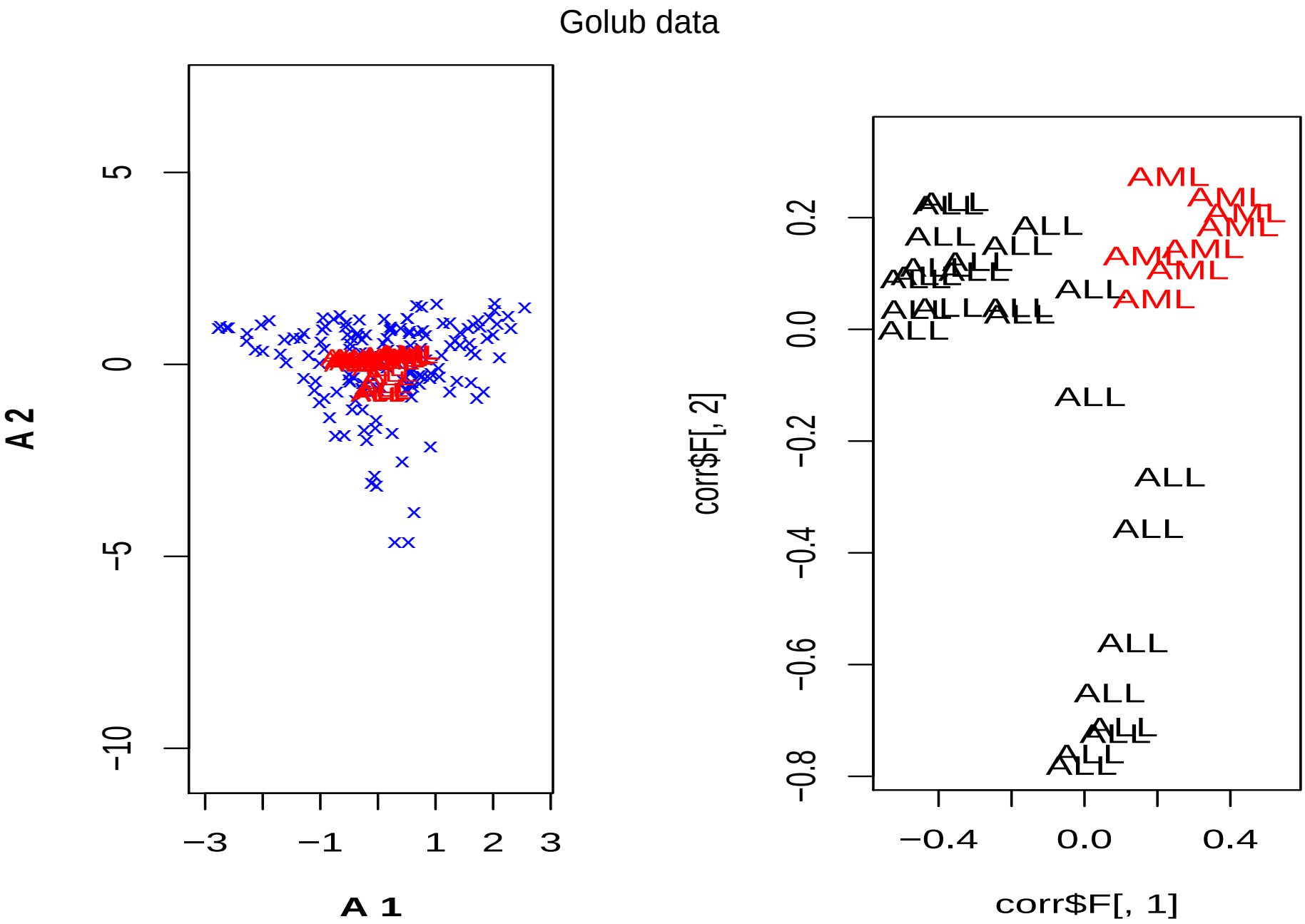
Projection methods: Correspondence analysis

❍ Correspondence analysis is usually applied to tables of frequencies (contingency tables) in order to show associations between particular rows and columns – in the sense of deviations from homogeneity, as measured by the χ^2-statistic.

❍ Data matrix is supposed to contain only positive numbers - may apply global shifting.

❍ R packages `CoCoAn, multiv`.

Correspondence analysis - Example

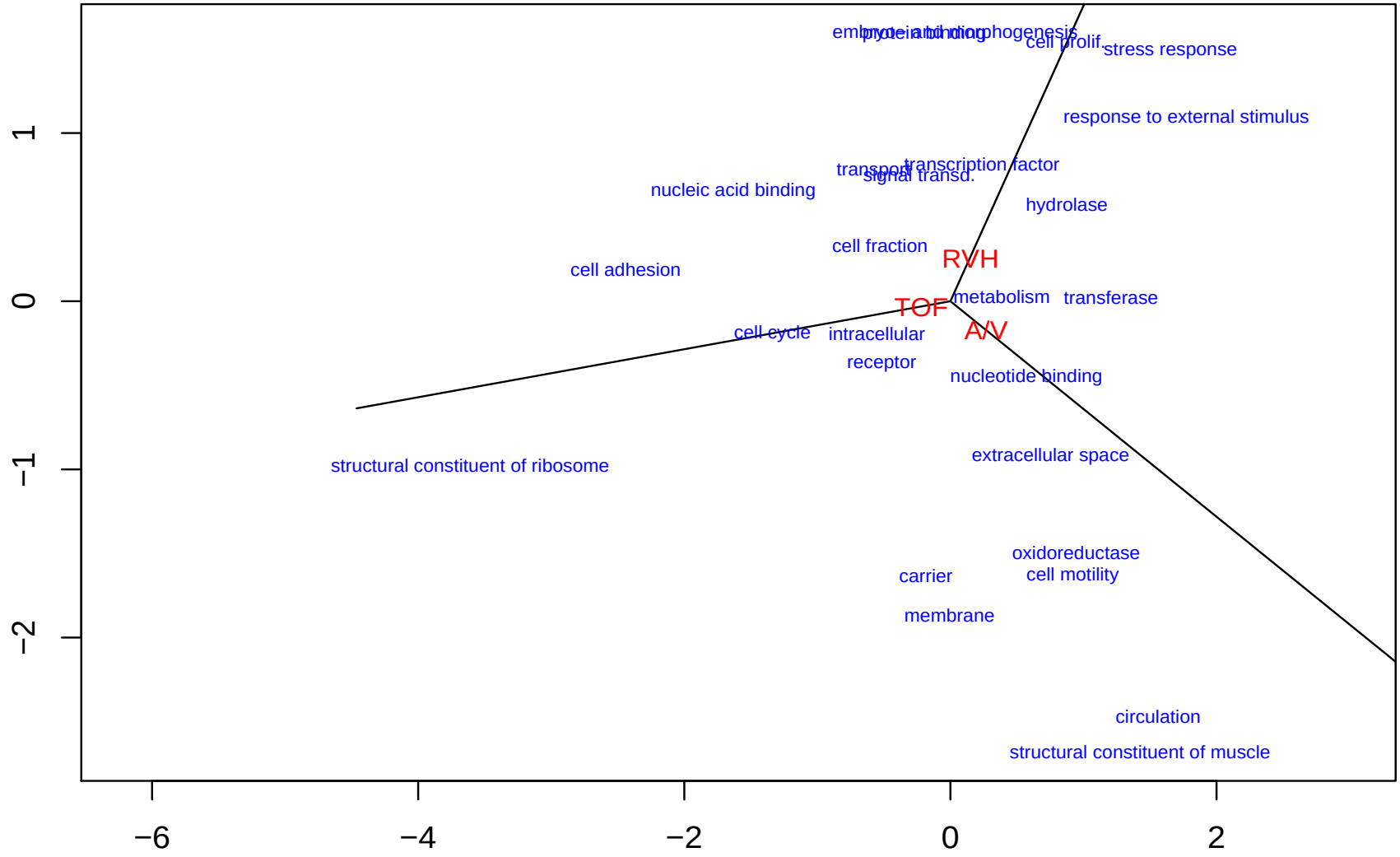


Contingency table of differentially expressed genes

	right ventricular hypertrophy	tetralogy of Fallot	atrium/ ventricle
stress response	11	8	9
constituent of muscle	7	29	20
constituent of ribosome	9	20	8
cell proliferation	7	7	5
signal transduction	14	25	11
metabolism	38	66	44
cell motility	5	12	12
...	...	...	...

Correspondence analysis

Association between Gene Ontology categories and tissue/disease phenotypes



ISIS - a class discovery method

❍ Aim: detect subtle class distinctions among a set of tissue samples/gene expression profiles (application: search for disease subtypes)

❍ Idea: Such class distinctions may be characterized by differential expression of just a small set of genes, not by global similarity of the gene expression profiles.

❍ The method quantifies this notion and conducts a search for interesting class distinctions in this sense.

❍ R package `ISIS` available at
`http://www.molgen.mpg.de/~heydebre`

References

❍ Duda, Hart and Stork (2000). Pattern Classification. 2nd Edition. Wiley.

❍ Dudoit and Fridlyand (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, Vol. 3(7), research 0036.1-0036.21.

❍ Eisen et al. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, Vol 95, 14863–14868.

❍ Fellenberg et al. (2001): Correspondence analysis applied to microarray data. *PNAS*, Vol. 98, p. 10781–10786.

❍ v. Heydebreck et al. (2001). Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, Suppl. 1, S107–114.

❍ Kerr and Churchill (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS*, Vol. 98, p. 8961-8965.

❍ Pollard and van der Laan (2002). Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences*, Vol. 176, 99-121.