*Statistics Applied to Bioinformatics*

# *Multivariate analysis*
# *Introduction*

**Jacques van Helden**
ORCID 0000-0002-8799-8584

Institut Français de Bioinformatique (**IFB**)
French node of the European **ELIXIR** bioinformatics infrastructure

Aix-Marseille Université (AMU)
Lab. Theory and Approaches of Genomic Complexity (**TAGC**)

# Multivariate data

- Each row represents one object (also called unit)
- Each column represents one variable

|  | variable 1 | variable 2 | … | variable p |
|---|---|---|---|---|
| individual 1 | $x_{11}$ | $x_{21}$ | … | $x_{p1}$ |
| individual 2 | $x_{12}$ | $x_{22}$ | … | $x_{p2}$ |
| individual 3 | $x_{13}$ | $x_{23}$ | … | $x_{p3}$ |
| individual 4 | $x_{14}$ | $x_{24}$ | … | $x_{p4}$ |
| individual 5 | $x_{15}$ | $x_{25}$ | … | $x_{p5}$ |
| individual 6 | $x_{16}$ | $x_{26}$ | … | $x_{p6}$ |
| individual 7 | $x_{17}$ | $x_{27}$ | … | $x_{p7}$ |
| individual 8 | $x_{18}$ | $x_{28}$ | … | $x_{p8}$ |
| … | … | … | … | … |
| individual n | $x_{1n}$ | $x_{2n}$ | … | $x_{pn}$ |

# *Multivariate data with an outcome variable*

- The outcome variable (also called criterion variable) can be
  - qualitative (nominal) : classes (e.g. cancer type)
  - quantitative (e.g. survival expectation for a cancer patient)

| | **Predictor variables** | | | | **Outcome variable** |
|---|---|---|---|---|---|
| | **variable 1** | **variable 2** | **...** | **variable p** | **variable p+1** |
| **individual 1** | $x_{11}$ | $x_{21}$ | … | $x_{p1}$ | $y_{1}$ |
| **individual 2** | $x_{12}$ | $x_{22}$ | … | $x_{p2}$ | $y_{2}$ |
| **individual 3** | $x_{13}$ | $x_{23}$ | … | $x_{p3}$ | $y_{3}$ |
| **individual 4** | $x_{14}$ | $x_{24}$ | … | $x_{p4}$ | $y_{4}$ |
| **individual 5** | $x_{15}$ | $x_{25}$ | … | $x_{p5}$ | $y_{5}$ |
| **individual 6** | $x_{16}$ | $x_{26}$ | … | $x_{p6}$ | $y_{6}$ |
| **individual 7** | $x_{17}$ | $x_{27}$ | … | $x_{p7}$ | $y_{7}$ |
| **individual 8** | $x_{18}$ | $x_{28}$ | … | $x_{p8}$ | $y_{8}$ |
| **...** | … | … | … | … | … |
| **individual n** | $x_{1n}$ | $x_{2n}$ | … | $x_{pn}$ | $y_{n}$ |

# Predictive approaches - Training set

- The training set is used to build a predictive function
- This function is used to predict the value of the outcome variable for new objects

**Training set**

| | Predictor variables | | | | Outcome variable |
|---|---|---|---|---|---|
| | **variable 1** | **variable 2** | **...** | **variable p** | **variable p+1** |
| **individual 1** | $x_{11}$ | $x_{21}$ | ... | $x_{p1}$ | $y_1$ |
| **individual 2** | $x_{12}$ | $x_{22}$ | ... | $x_{p2}$ | $y_2$ |
| **individual 3** | $x_{13}$ | $x_{23}$ | ... | $x_{p3}$ | $y_3$ |
| **...** | ... | ... | ... | ... | ... |
| **individual N_train** | $x_{1n}$ | $x_{2n}$ | ... | $x_{pn}$ | $y_n$ |

**Set to predict**

| | Predictor variables | | | | Outcome variable |
|---|---|---|---|---|---|
| | **variable 1** | **variable 2** | **...** | **variable p** | **variable p+1** |
| **individual 1** | $x_{11}$ | $x_{21}$ | ... | $x_{p1}$ | ? |
| **individual 2** | $x_{12}$ | $x_{22}$ | ... | $x_{p2}$ | ? |
| **individual 3** | $x_{13}$ | $x_{23}$ | ... | $x_{p3}$ | ? |
| **...** | ... | ... | ... | ... | ... |
| **individual N_pred** | $x_{1n}$ | $x_{2n}$ | ... | $x_{pn}$ | ? |

Training set

| | Predictor variables | | | | Outcome variable |
|---|---|---|---|---|---|
| | variable 1 | variable 2 | ... | variable p | variable p+1 |
| individual 1 | $x_{11}$ | $x_{12}$ | … | $x_{1p}$ | $y_1$ |
| individual 2 | $x_{21}$ | $x_{22}$ | … | $x_{2p}$ | $y_2$ |
| individual 3 | $x_{31}$ | $x_{32}$ | … | $x_{3p}$ | $y_3$ |
| ... | … | … | … | … | … |
| individual ntrain | $x_{n1}$ | $x_{n2}$ | … | $x_{np}$ | $y_n$ |

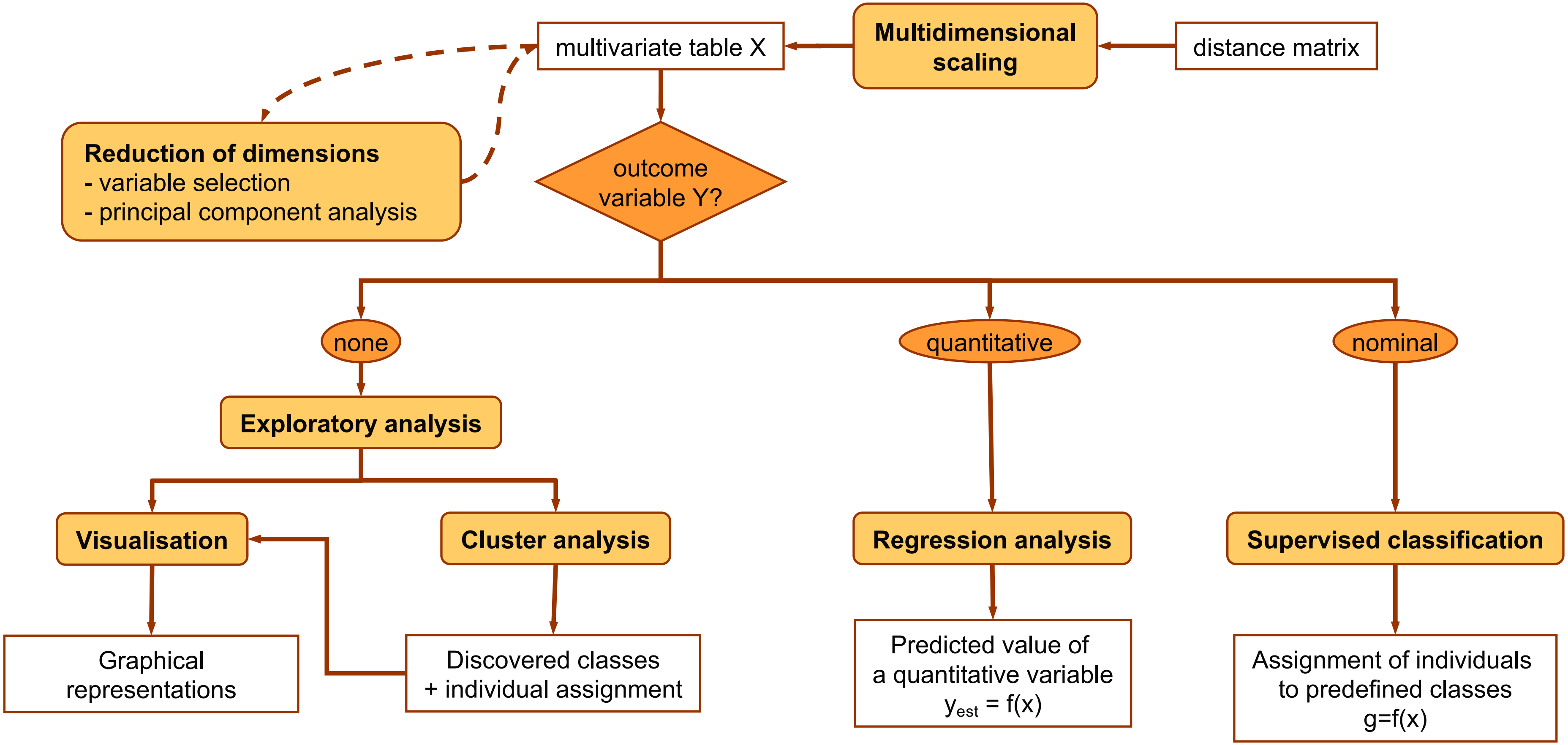
Testing set

| | Predictor variables | | | | Outcome variable | |
|---|---|---|---|---|---|---|
| | variable 1 | variable 2 | ... | variable p | variable p+1 (known value) | variable p+1 (predicted) |
| individual 1 | $x_{11}$ | $x_{12}$ | … | $x_{1p}$ | $y_1$ | $y'_1$ |
| individual 2 | $x_{21}$ | $x_{22}$ | … | $x_{2p}$ | $y_2$ | $y'_2$ |
| individual 3 | $x_{31}$ | $x_{32}$ | … | $x_{3p}$ | $y_3$ | $y'_3$ |
| ... | … | … | … | … | … | … |
| individual ntest | $x_{n1}$ | $x_{n2}$ | … | $x_{np}$ | $y_{ntest}$ | $y'_{ntest}$ |

et to predict

| | Predictor variables | | | | Outcome variable |
|---|---|---|---|---|---|
| | variable 1 | variable 2 | ... | variable p | variable p+1 |
| individual 1 | $x_{11}$ | $x_{12}$ | … | $x_{1p}$ | ? |
| individual 2 | $x_{21}$ | $x_{22}$ | … | $x_{2p}$ | ? |
| individual 3 | $x_{31}$ | $x_{32}$ | … | $x_{3p}$ | ? |
| ... | … | … | … | … | … |

# Flowchart of the approaches in multivariate analysis


multivariate table X
Multidimensional scaling
distance matrix
Reduction of dimensions
- variable selection
- principal component analysis
outcome variable Y?
none
quantitative
nominal
Exploratory analysis
Visualisation
Cluster analysis
Regression analysis
Supervised classification
Graphical representations
Discovered classes + individual assignment
Predicted value of a quantitative variable $y_{est} = f(x)$
Assignment of individuals to predefined classes g=f(x)

# *Quizz*

Check your understanding of the concepts presented in the previous slides by applying them to your own data.

1. Describe in one sentence a typical case of multidimensional data that is handled in your domain.
2. Explain how you would organise this dataset into a multivariate structure
   - ❑ What would correspond to the individuals?
   - ❑ What would correspond to the variables?
   - ❑ How many individuals (n) would you have?
   - ❑ How many variables (p) would you have?
   - ❑ Do you dispose of one or several outcome variable(s)?
   - ❑ If so, are these quantitative, qualitative or both?
3. Based on the conceptual framework defined above, which kind of approaches would be you envisage to extract which kind of relevant information from this data? Note that several approaches can be combined to address different questions.

# *Historical (vintage) examples*

# Historical example of clustering heat map

- Spellman et al. (1998).
- Systematic detection of genes regulated in a periodic way during the 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)

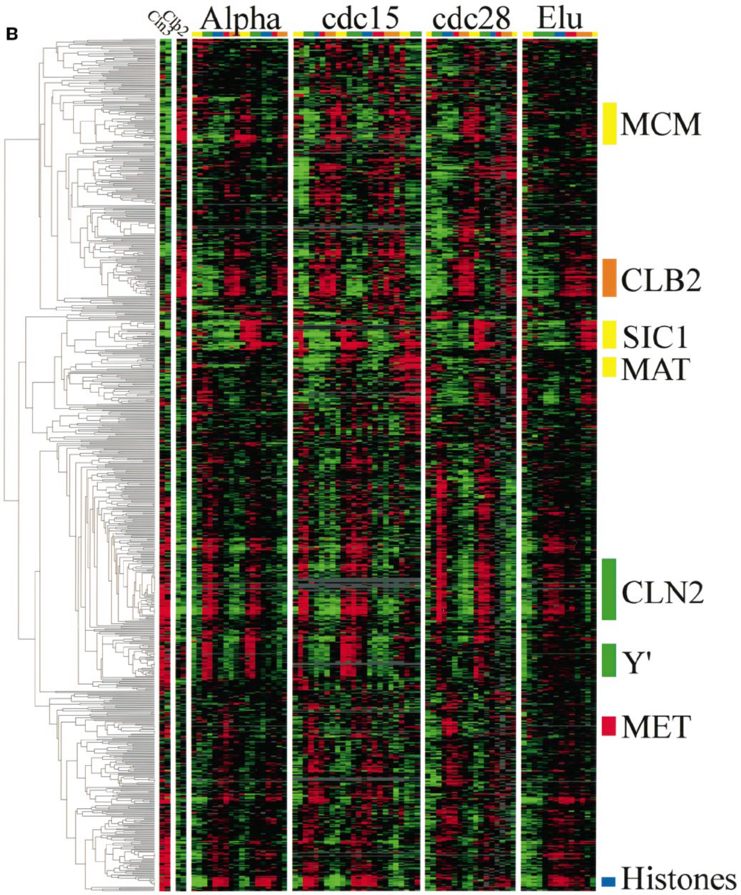
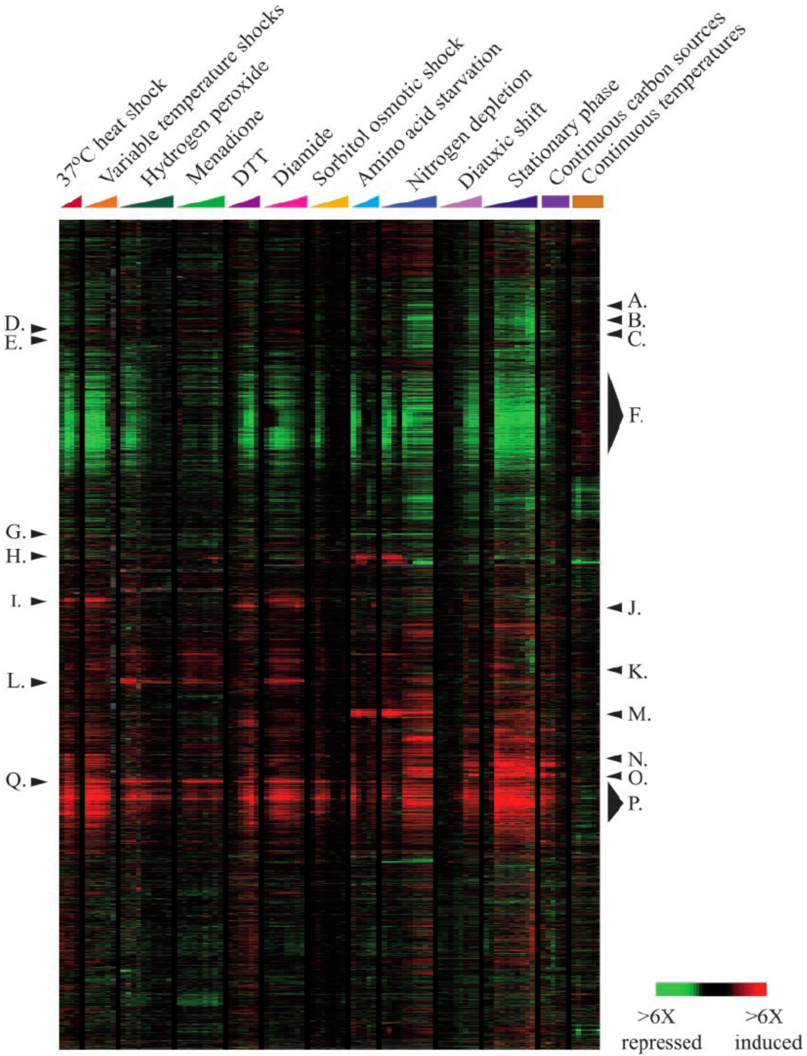


Figure 1. (cont).

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 9, 3273-97.Time profiles of yeast cells followed during cell cycle.

# Stress response in yeast

- Gasch et al. (2000) tested the transcriptional response of yeast genome to
  - Various stress conditions (heat shock, osmotic shock, …)
  - Drugs
  - Alternative carbon sources
  - …
- The heatmap shows clusters of genes having similar profiles of responses to the different types of stress.



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

# *Cancer types (Golub, 1999)*

- Compared the profiles of expression of ~7000 human genes in patients suffering from two different cancer types: ALL or AML, respectively.
- Selected the 50 genes most correlated with the cancer type.
- Goal: use these genes as molecular signatures for the diagnostic of new patients.

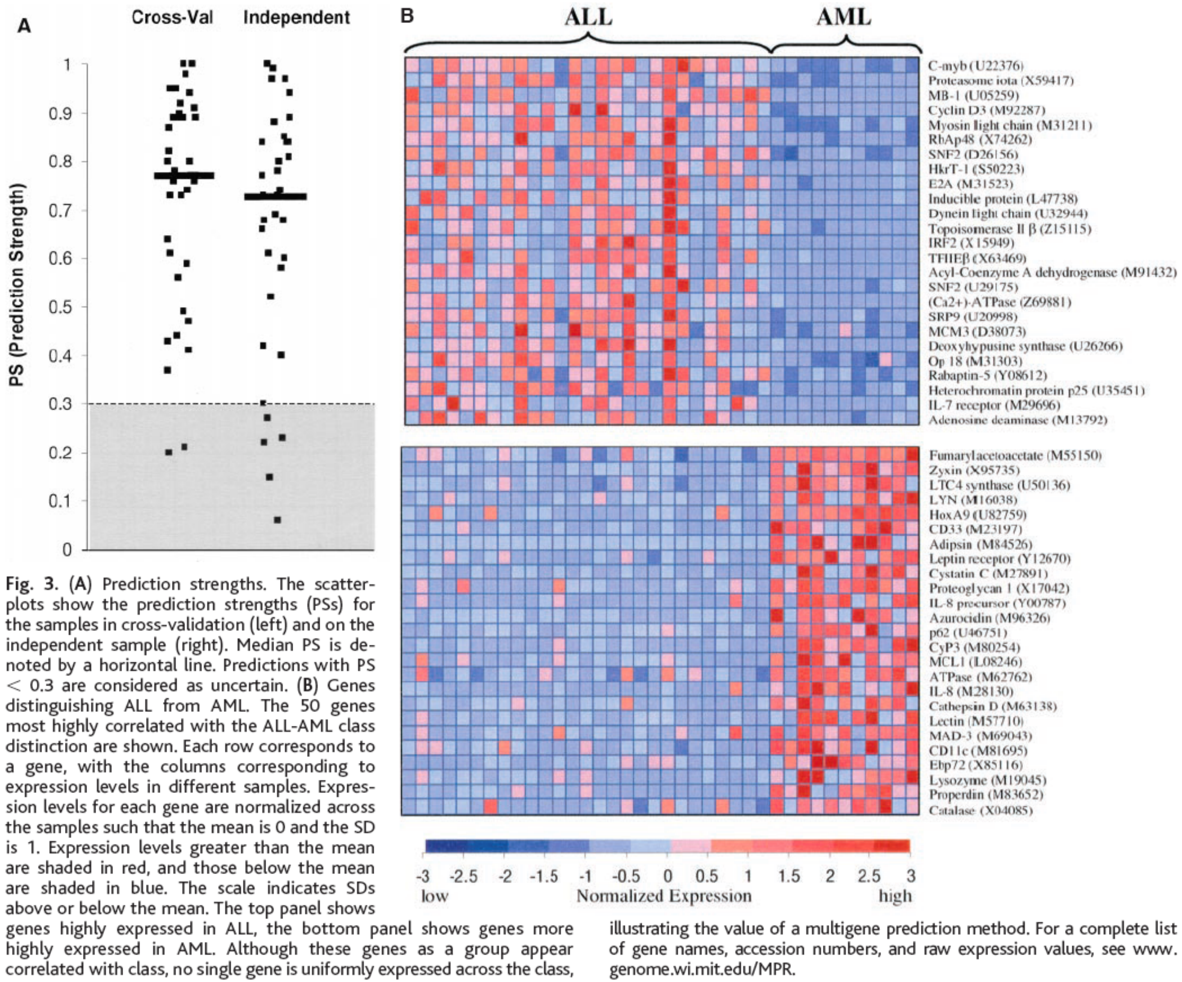


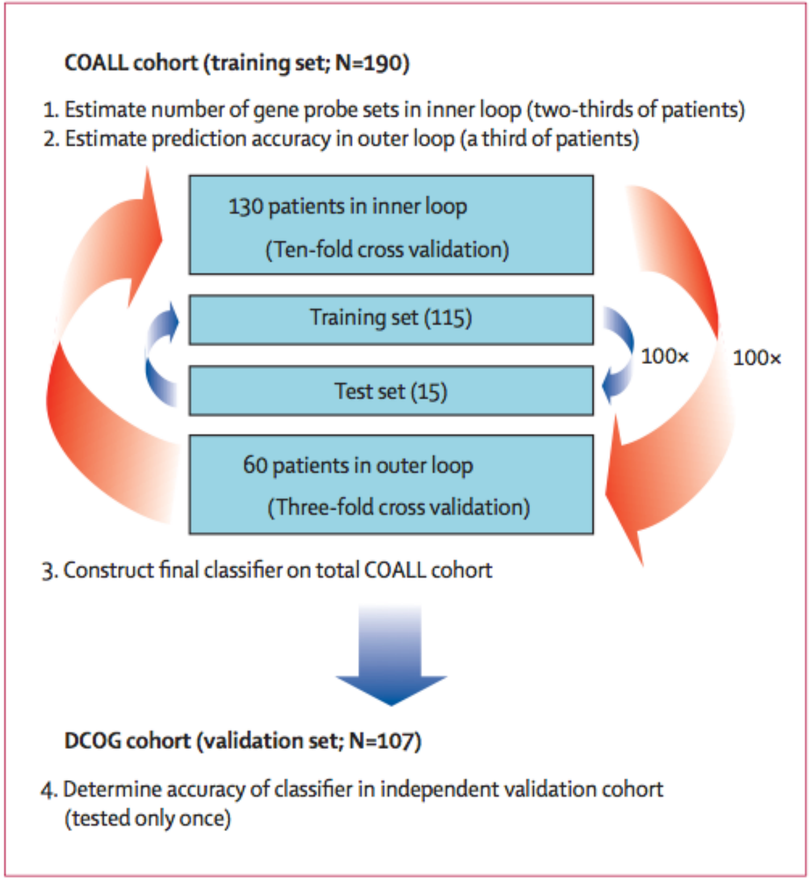
**Fig. 3.** (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS $< 0.3$ are considered as uncertain. (**B**) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.

- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-7.

# Den Boer et al., 2009 : procedure

- Den Boer et al (2009) use Affymetrix microarrays to characterize the transcriptome of 190 Acute Lymphoblastic Leukemia of different types.
- They use these profiles to select “transcriptome signatures” that will serve for diagnostics purposes: assigning new samples to one of the cancer types.
- They apply an elaborate procedure relying on an inner and an outer loop of cross-validation.

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



COALL cohort (training set; N=190)

1. Estimate number of gene probe sets in inner loop (two-thirds of patients)
2. Estimate prediction accuracy in outer loop (a third of patients)

130 patients in inner loop
(Ten-fold cross validation)
Training set (115)
Test set (15)
60 patients in outer loop
(Three-fold cross validation)
100×
100×

3. Construct final classifier on total COALL cohort

DCOG cohort (validation set; N=107)

4. Determine accuracy of classifier in independent validation cohort (tested only once)

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

- 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.

# Den Boer 2009 - The transcriptomic signature

- The training procedure selects 100 gens whose combined expression levels can be used to assign samples to cancer subtypes.
- The heatmaps show that the selected genes are differentially expressed
  - between subtypes of the training set (left);
  - between subtypes of the testing set (right).
- The heatmap is bi-clustered, in order to identify  simultaneously the groups of patients (rows), and groups of genes (columns) based on the similarity between expression profiles.

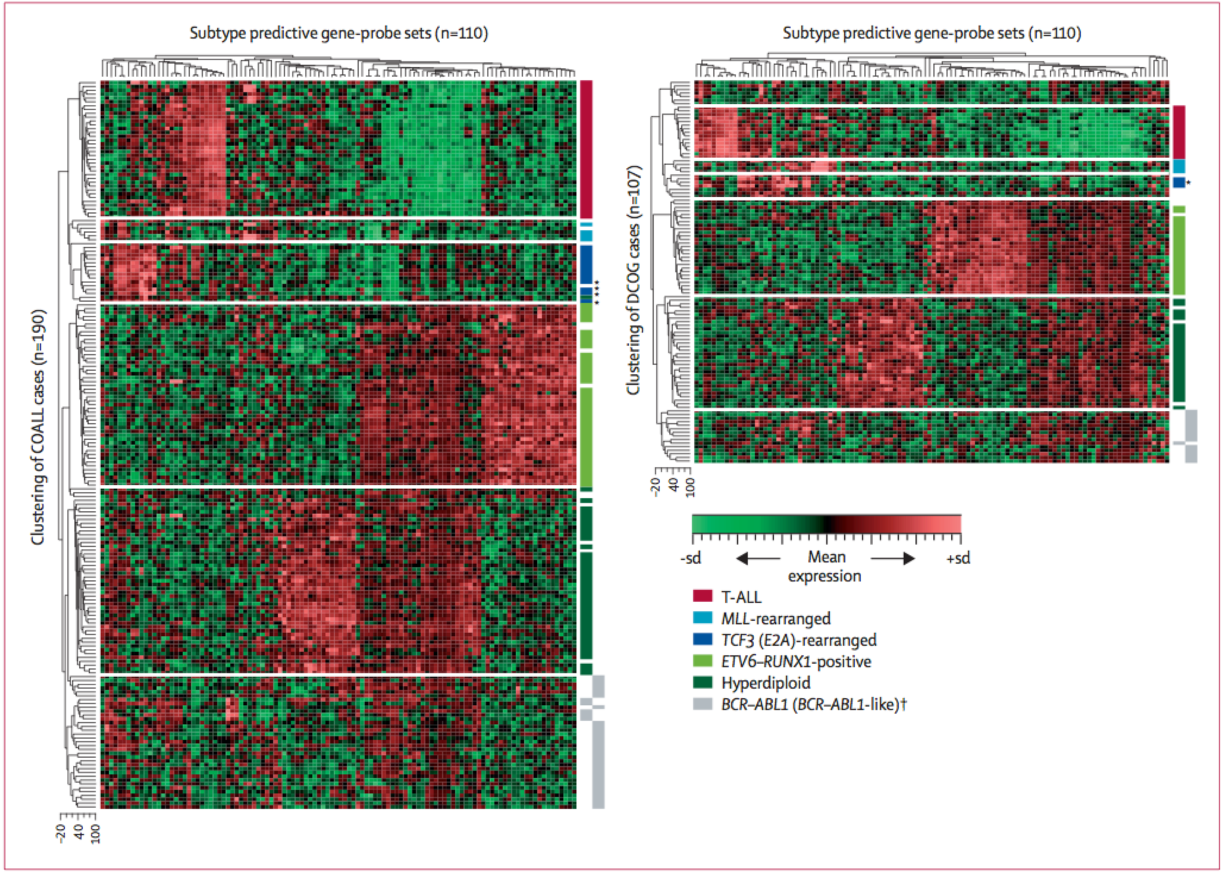
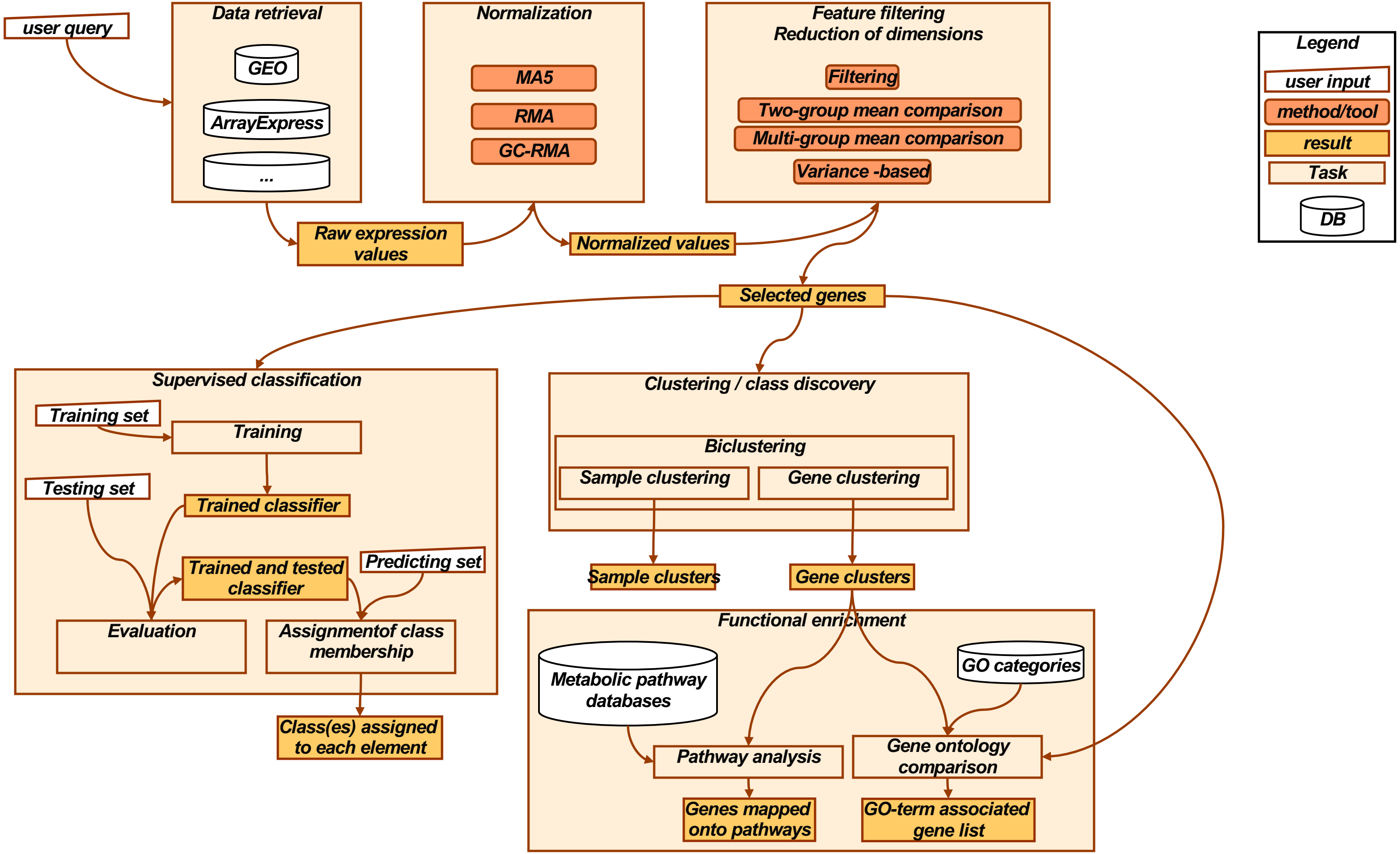


***Figure 2:* Clustering of ALL subtypes by gene-expression profiles**

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

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

# Conceptual workflow – transcriptome analysis



user query
Data retrieval
GEO
ArrayExpress
...
Raw expression values
Normalization
MA5
RMA
GC-RMA
Normalized values
Feature filtering
Reduction of dimensions
Filtering
Two-group mean comparison
Multi-group mean comparison
Variance -based
Selected genes
Legend
user input
method/tool
result
Task
DB
Supervised classification
Training set
Training
Testing set
Trained classifier
Trained and tested classifier
Predicting set
Evaluation
Assignmentof class membership
Class(es) assigned to each element
Clustering / class discovery
Biclustering
Sample clustering
Gene clustering
Sample clusters
Gene clusters
Functional enrichment
Metabolic pathway databases
GO categories
Pathway analysis
Gene ontology comparison
Genes mapped onto pathways
GO-term associated gene list