Study cases - multivariate analysis

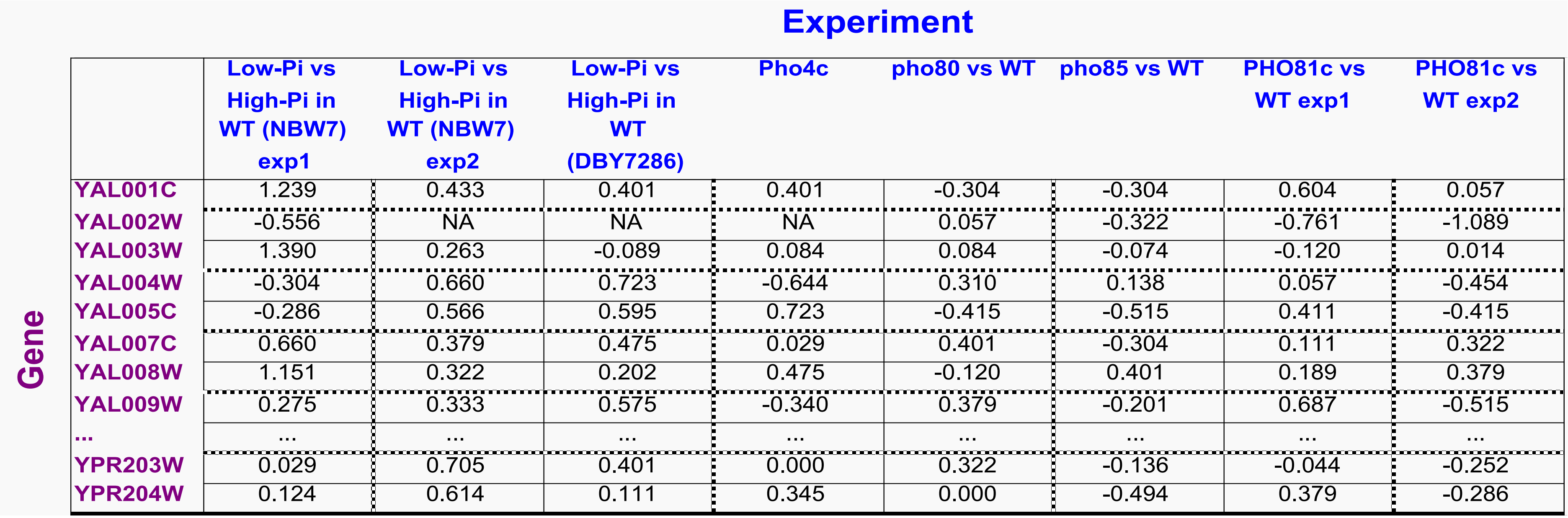
Multivariate data

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

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

Example of multivariate table - gene expression data

- Data:
  - Ogawa, N., DeRisi, J. & Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. Mol Biol Cell 11, 4309-21.
- Transcriptional response to phosphate in yeast.
  - Each column represents one DNA chip (8 conditions were tested)
  - Each row represents a gene
  - The cells of the matrix represent the level of regulation of the gene on a given DNA chip



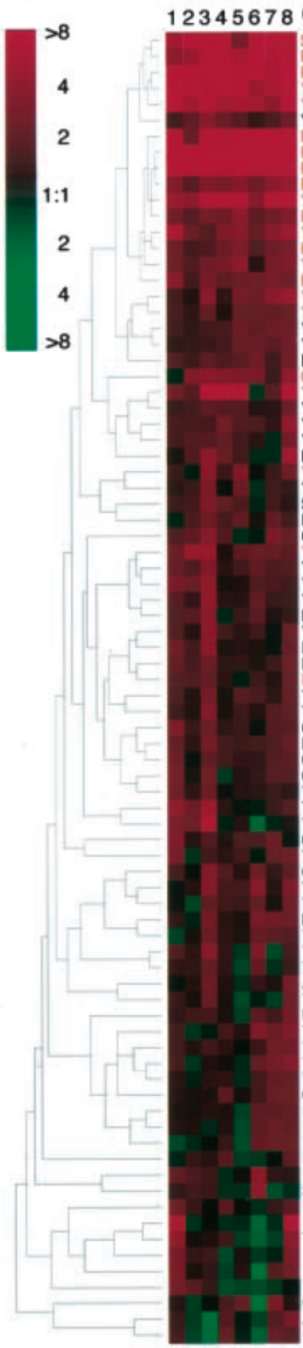
Experiment

Gene

	Low-Pi vs High-Pi in WT (NBW7) exp1	Low-Pi vs High-Pi in WT (NBW7) exp2	Low-Pi vs High-Pi in WT (DBY7286)	Pho4c	pho80 vs WT	pho85 vs WT	PHO81c vs WT exp1	PHO81c vs WT exp2
YAL001C	1.239	0.433	0.401	0.401	-0.304	-0.304	0.604	0.057
YAL002W	-0.556	NA	NA	NA	0.057	-0.322	-0.761	-1.089
YAL003W	1.390	0.263	-0.089	0.084	0.084	-0.074	-0.120	0.014
YAL004W	-0.304	0.660	0.723	-0.644	0.310	0.138	0.057	-0.454
YAL005C	-0.286	0.566	0.595	0.723	-0.415	-0.515	0.411	-0.415
YAL007C	0.660	0.379	0.475	0.029	0.401	-0.304	0.111	0.322
YAL008W	1.151	0.322	0.202	0.475	-0.120	0.401	0.189	0.379
YAL009W	0.275	0.333	0.575	-0.340	0.379	-0.201	0.687	-0.515
...	...	...	...	...	...	...	...	...
YPR203W	0.029	0.705	0.401	0.000	0.322	-0.136	-0.044	-0.252
YPR204W	0.124	0.614	0.111	0.345	0.000	-0.494	0.379	-0.286

Phosphate metabolism

- Ogawa, N., DeRisi, J. & Brown, P. O. (2000).
- Analysis of genes repsonding to changes in phosphate concentration, or deregulated in mutants of genes involved phosphate regulation.
- 8 conditions were tested
- 81 genes are selected as responding to phosphate stress.

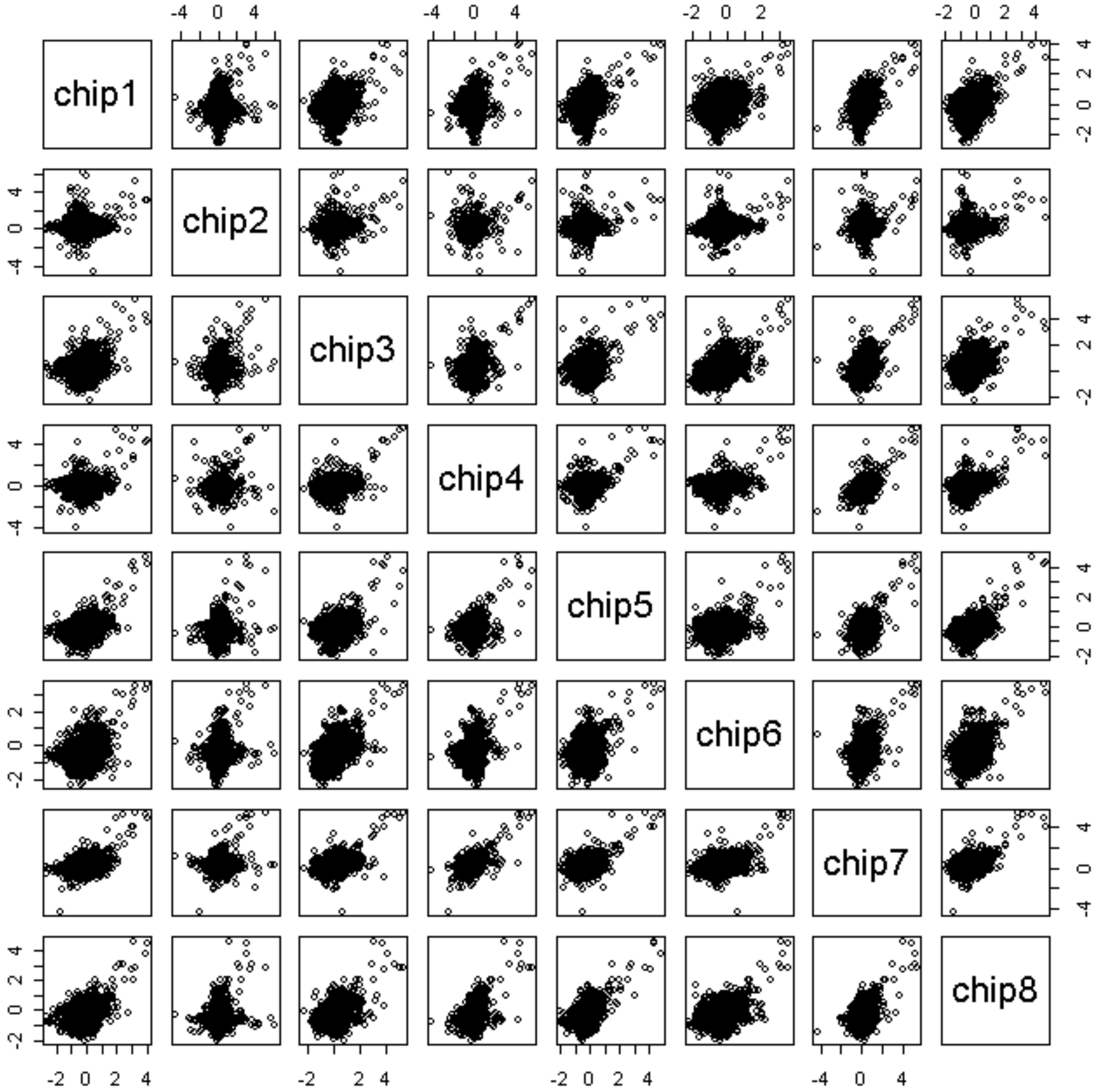


Gene	Description
SPL2	PHO81 ANKYRIN HOMOLOG
PHO89	NA+/PI COTRANSPORTER
PHO84	HIGH-AFFINITY PI/H+ SYMPORTER
YPL019C	HOMOLOGY TO YFL004W
YER072W	SIMILARITY TO YFL004W C-TERMINAL HALF
YMR291W	PUTATIVE PROTEIN KINASE
PHO5	REPRESSIBLE ACID PHOSPHATASE
PHO12	SECRETED ACID PHOSPHATASE
PHO11	SECRETED ACID PHOSPHATASE,56 KDA ISOZYME
PHO8	REPRESSIBLE VACUOLAR ALKALINE PHOSPHATASE
YJL012C	SIMILARITY TO YFL004W
PHO86	INVOLVED IN PI TRANSPORT
YDR452W	SIMILAR TO HUMAN SPHINGOMYELIN PHOSPHODIESTERASE
HIS1	ATP PHOSPHORIBOSYLTRANSFERASE
YFL004W	HOMOLOGY TO YPL019C
HOR2	DL-GLYCEROL PHOSPHATASE
YER037W	SIMILARITY TO HYPOTHETICAL PROTEIN YGL224C
YER038C	FUNCTION UNKNOWN
YNL217W	SIMILARITY TO E.COLI BIS 5'-NUCLEOSYL-TETRAPHOSPHATASE
YDR501W	SIMILARITY TO YLR183C
NTH2	ALPHA,ALPHA-TREHALASE
CTF19	IMPORTANT FOR CHROMOSOME SEGREGATION
YDR281C	HYPOTHETICAL PROTEIN
YLR410W	HYPOTHETICAL PROTEIN
YAR069C	3 POTENTIAL MEMBRANE SPANNING REGIONS
YJL119C	SIMILAR TO HUMAN RETINOBLASTOMA BINDING PROTEIN2
KRE2	ALPHA-1,2-MANNOSYLTRANSFERASE
YPL110C	PHO81 HOMOLOG
YMR195W	HYPOTHETICAL PROTEIN
SUR1	SUPPRESSOR OF RVS161, RVS167, AND CLS2 MUTATIONS
PYK2	PUTATIVE PYRUVATE KINASE
GPD1	GLYCEROL-3-PHOSPHATE DEHYDROGENASE NAD+ PRECURSOR
YOL084W	SIMILARITY TO S.POMBE SPAC24H6.13 AND YMR266W
TSL1	ALPHA,ALPHA-TREHALOSE-PHOSPHATE SYNTHASE UDP-FORMING SUBUNIT
YHR087W	HYPOTHETICAL PROTEIN
YML128C	HYPOTHETICAL PROTEIN
MGA1	SIMILARITY TO HEAT SHOCK TRANSCRIPTION FACTORS
YLR149C	HYPOTHETICAL PROTEIN
PGM2	PHOSPHOGLUCOMUTASE, MAJOR ENZYME
PTK2	PUTATIVE PROTEIN KINASE
PHO81	CYCLIN-DEPENDENT KINASE INHIBITOR
YDR005C	REQUIRED FOR SORTING OF MOD5P
YHL021C	WEAK SIMILARITY TO P.SP. GAMMA-BUTYROBETAINE HYDROXYLASE
GSY1	UDP GLUCOSE--STARCH GLUCOSYLTRANSFERASE 1
DDR48	STRESS PROTEIN
GLK1	ALDOHEXOSE SPECIFIC GLUCOKINASE
GLC3	1,4-GLUCAN BRANCHING ENZYME GLYCOGEN BRANCHING ENZYME
YDR516C	PUTATIVE GLUCOKINASE
TPS2	TREHALOSE-6-PHOSPHATE PHOSPHATASE
YJL107C	HYPOTHETICAL PROTEIN
CIT2	PEROXISOMAL CITRATE SI -SYNTHASE
YER044C	HYPOTHETICAL PROTEIN
GRE3	HOMOLOGY TO D-XYLOSE 1-DEHYDROGENASE NADP+
YBR051W	QUESTIONABLE ORF
YNL208W	HYPOTHETICAL PROTEIN
LAS1	INVOLVED IN CELL MORPHOGENESIS, CYTOSKELETAL REGULATION AND BUD FORMATION
YOR289W	HYPOTHETICAL PROTEIN
YLR327C	HOMOLOGY TO STF2 PROTEIN
YGR161C	HYPOTHETICAL PROTEIN
AMS1	ALPHA-MANNOSIDASE
YJR061W	SIMILARITY TO YKL200C AND YKL201C
PMC1	CA2+-TRANSPORTING P-TYPE ATPASE
YBR005W	HOMOLOGY TO YDR003W
YOR385W	HOMOLOGY TO HYPOTHETICAL PROTEIN YMR316W
YCR007C	MEMBER OF THE YBR302P/YCR007P/YHL048P/YKL219P FAMILY
YKL159C	HYPOTHETICAL PROTEIN
CMK2	CA/CALMODULIN-DEPENDENT S/T PROTEIN KINASE, TYPE II
PLB1	PHOSPHOLIPASE B LYSOPHOSPHOLIPASE
YMR316W	SIMILARITY TO YOR385W AND YNL165W
YMR007W	HYPOTHETICAL PROTEIN
NUF1	SPINDLE POLE BODY COMPONENT
HSP42	HEAT SHOCK PROTEIN WITH SIMILARITY TO HSP26P
SSA4	HEAT SHOCK PROTEIN
MET6	5-METHYLTETRAHYDROPTEROYLTRIGLUTAMATE-HOMOCYSTEINE METHYLTRANSFERASE
DAN1	SIMILARITY TO PROTEINS OF THE SRP1P/TIP1P FAMILY
TIR1	COLD-SHOCK INDUCED PROTEIN OF THE TIR1P,TIP1P FAMILY
YJR151C	SIMILARITY TO PROTEINS OF THE SRP1P/TIP1P FAMILY
YIL011W	SIMILARITY TO YIL176P,YIR041P AND OTHER MEMBERS OF THE SRP1P/TIP1P FAMILY
YPK2	SER/THR PROTEIN KINASE
CHA1	L-SERINE/L-THREONINE DEAMINASE
ZRT1	HIGH-AFFINITY ZINC TRANSPORT PROTEIN
CTT1	CYTOSOLIC CATALASE T

- Ogawa, N., DeRisi, J. & Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. Mol Biol Cell 11, 4309-21.

Gene expression data - relationships between variables

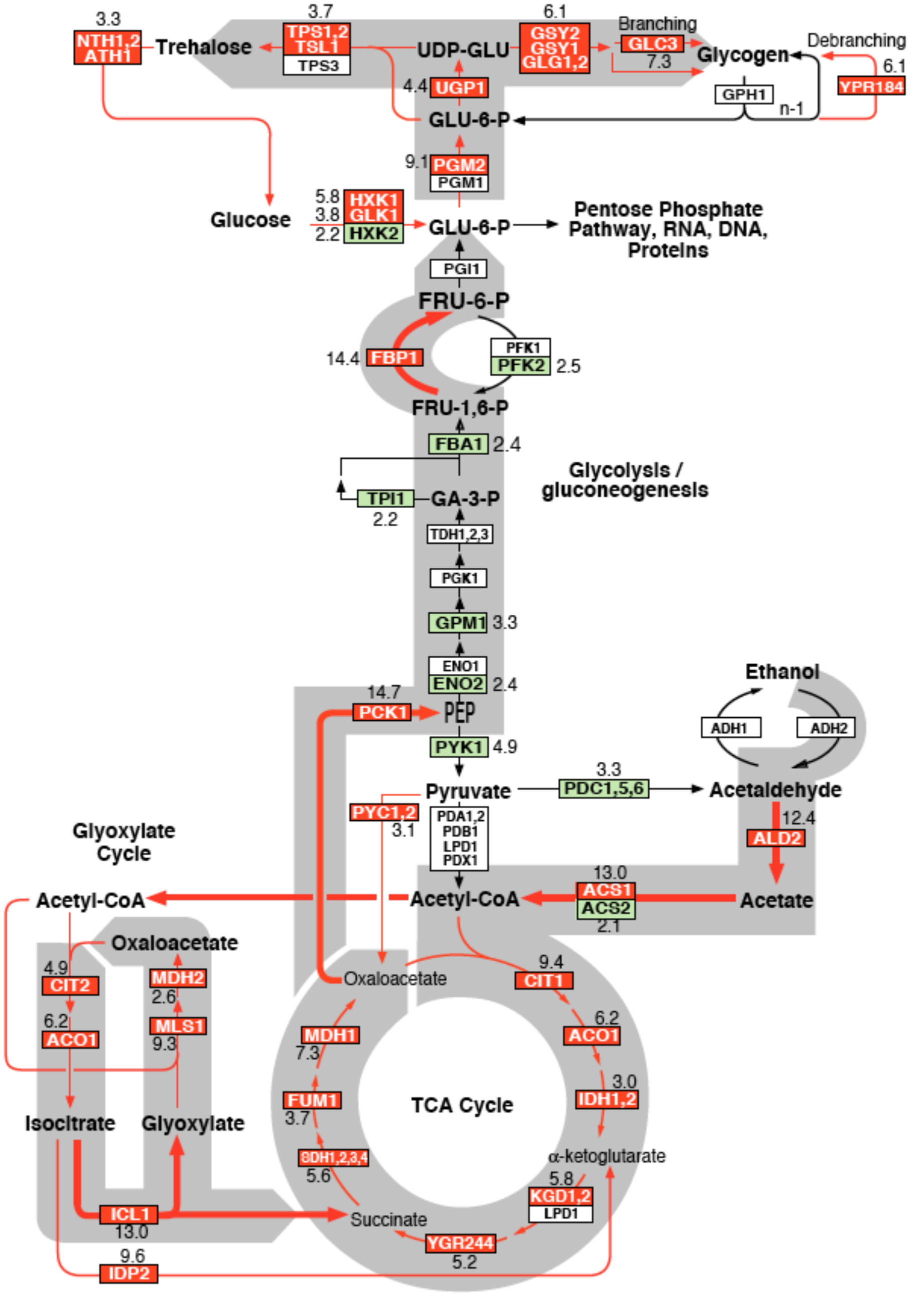
- Dot plots allow to compare the levels of regulation in the 8 experiments from Ogama et al. (2000).



- Ogawa, N., DeRisi, J. & Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. Mol Biol Cell 11, 4309-21.

Diauxic shift

- DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680-6.
- The first publication on full-genome monitoring of gene expression data.
- Time profile 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



Cell cycle data

- 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)

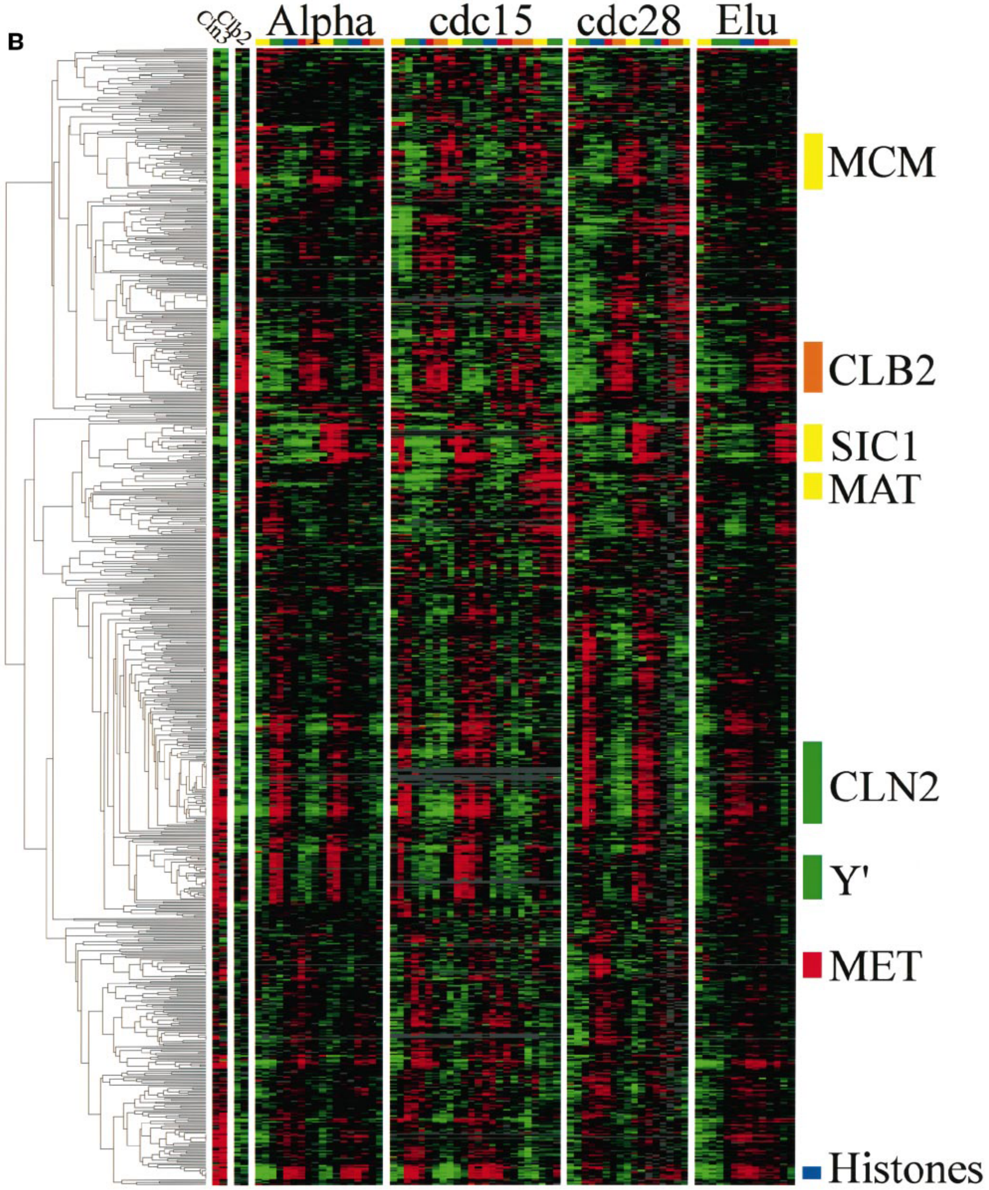
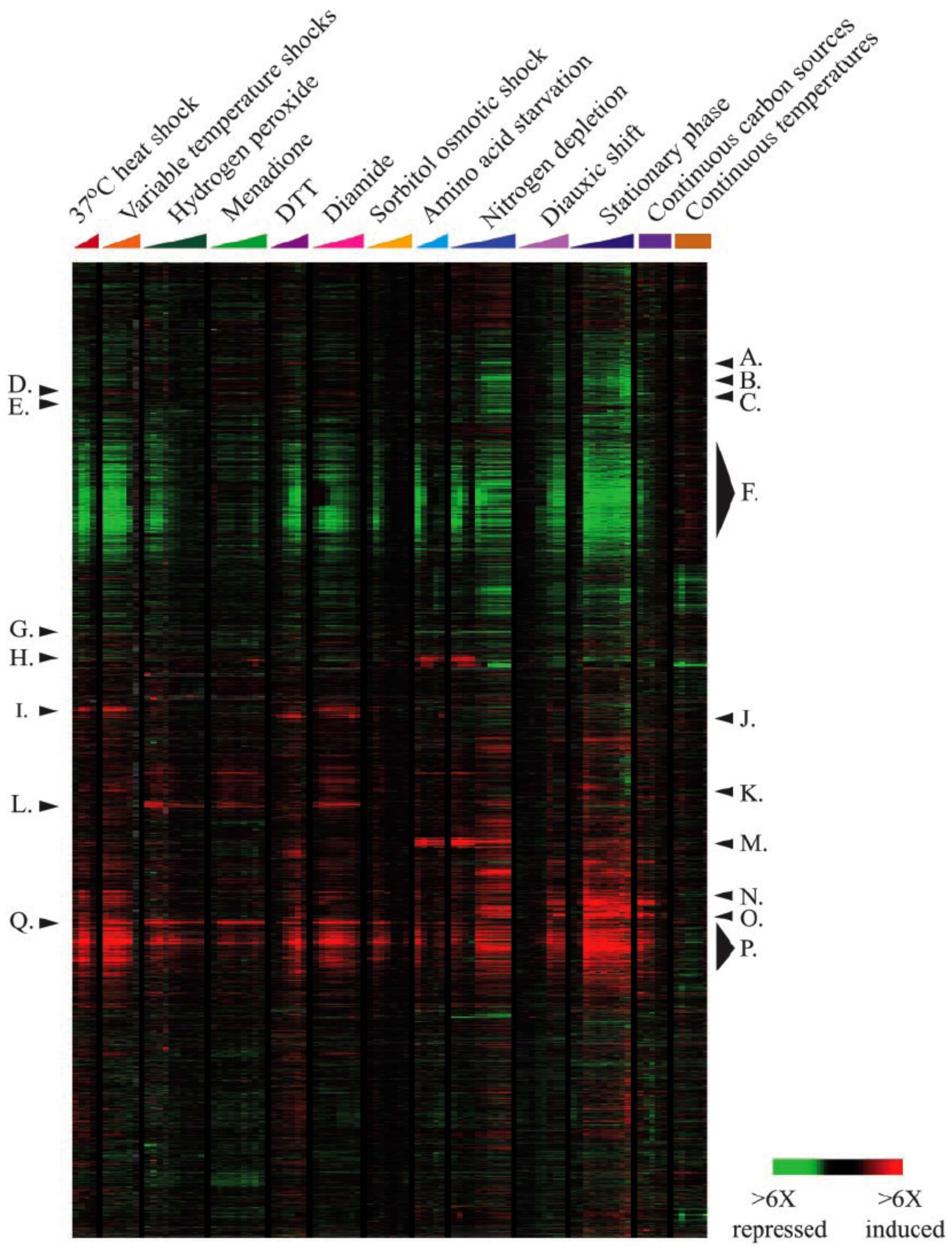


Figure 1. (cont).

- 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
  - …

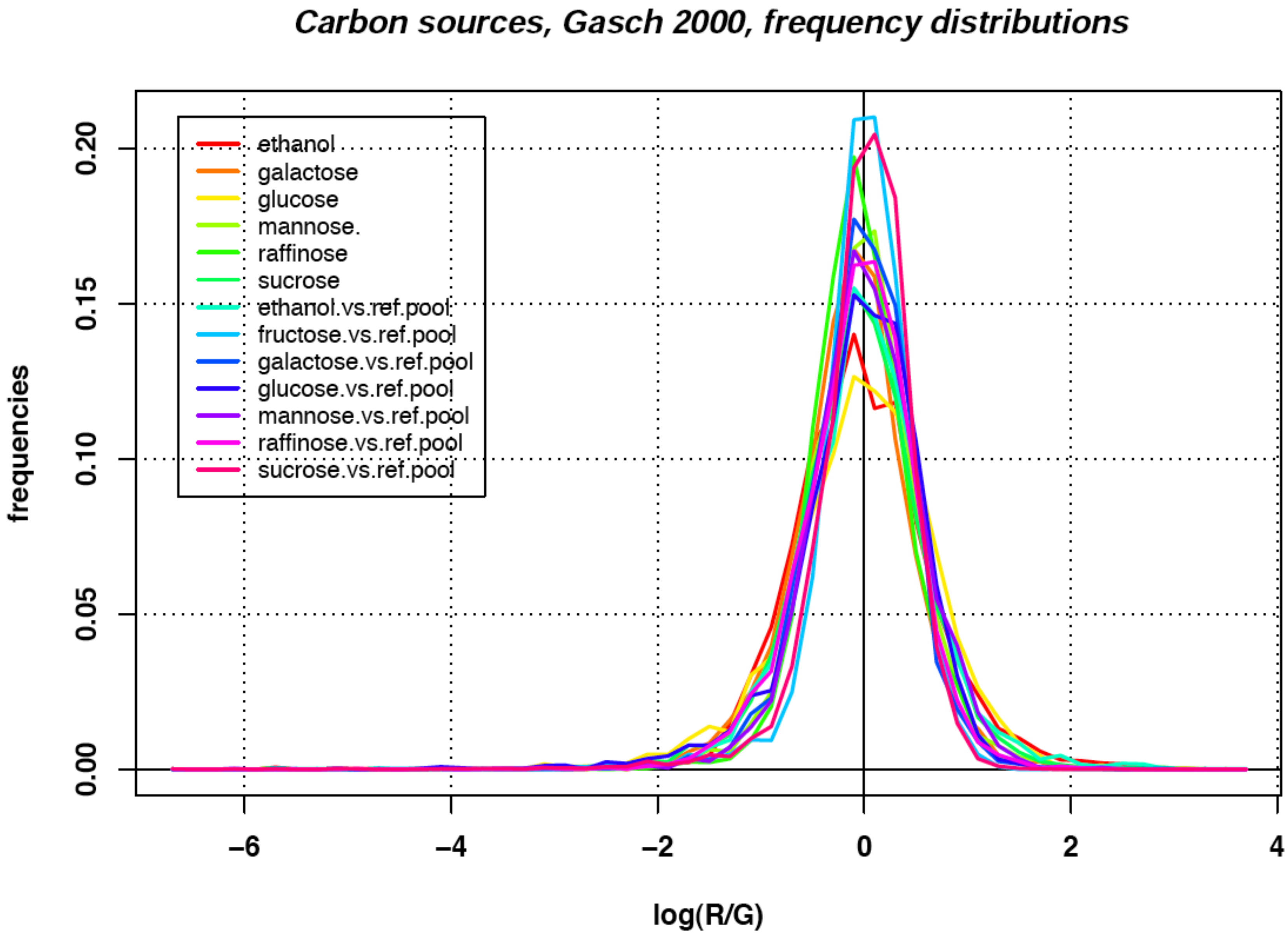


- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eis M. B., Storz, G., Botstein, D. & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241-57.

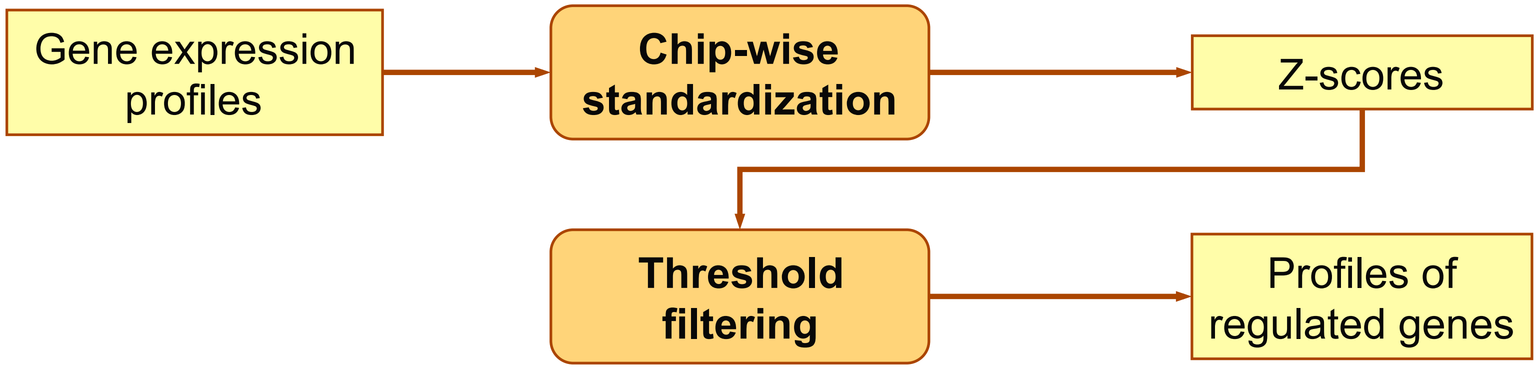
Data standardization and filtering

- 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.

Carbon sources - value distributions



Filtering of carbon source data



Gene expression profiles

Chip-wise standardization

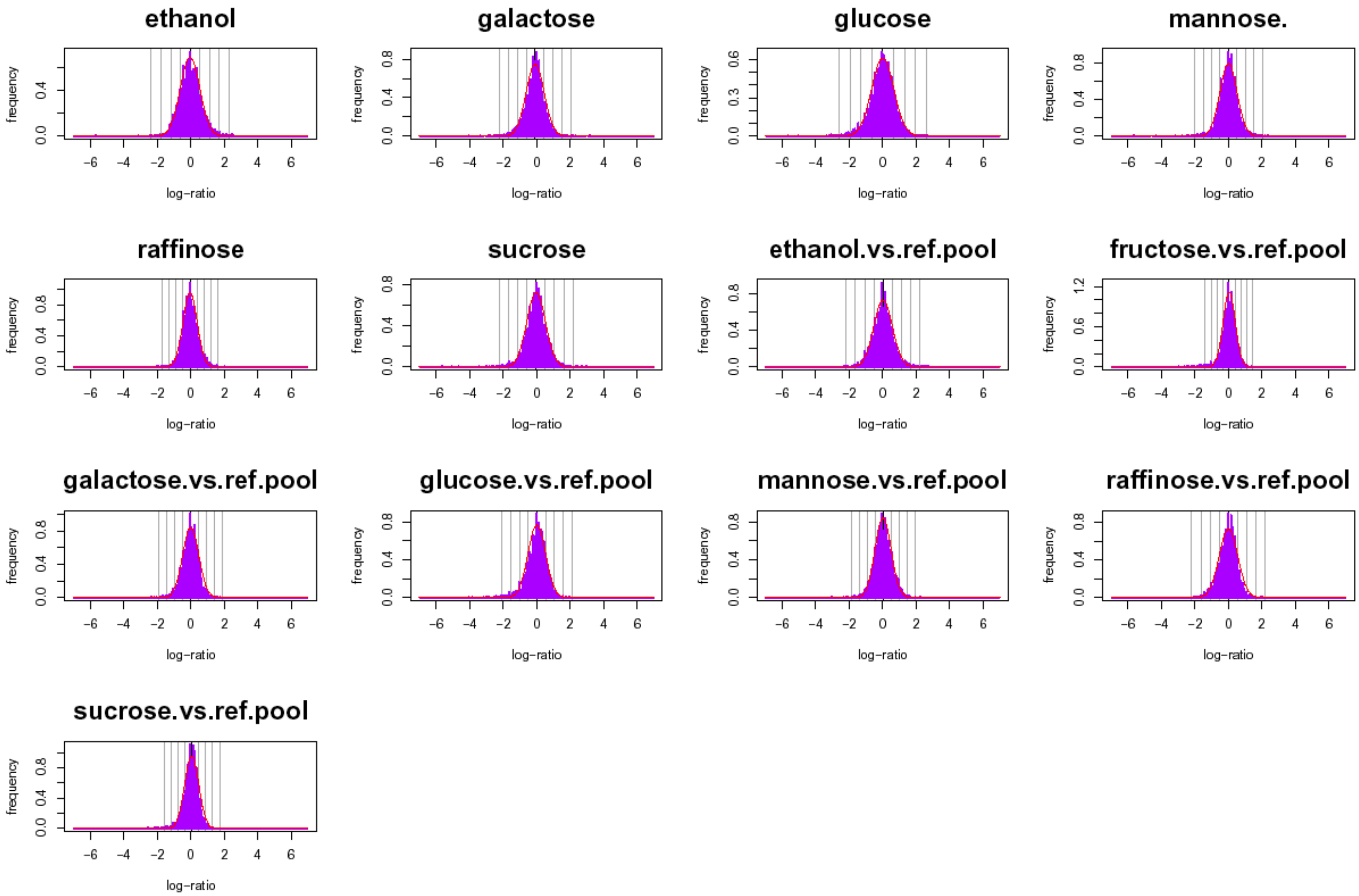
Z-scores

Threshold filtering

Profiles of regulated genes

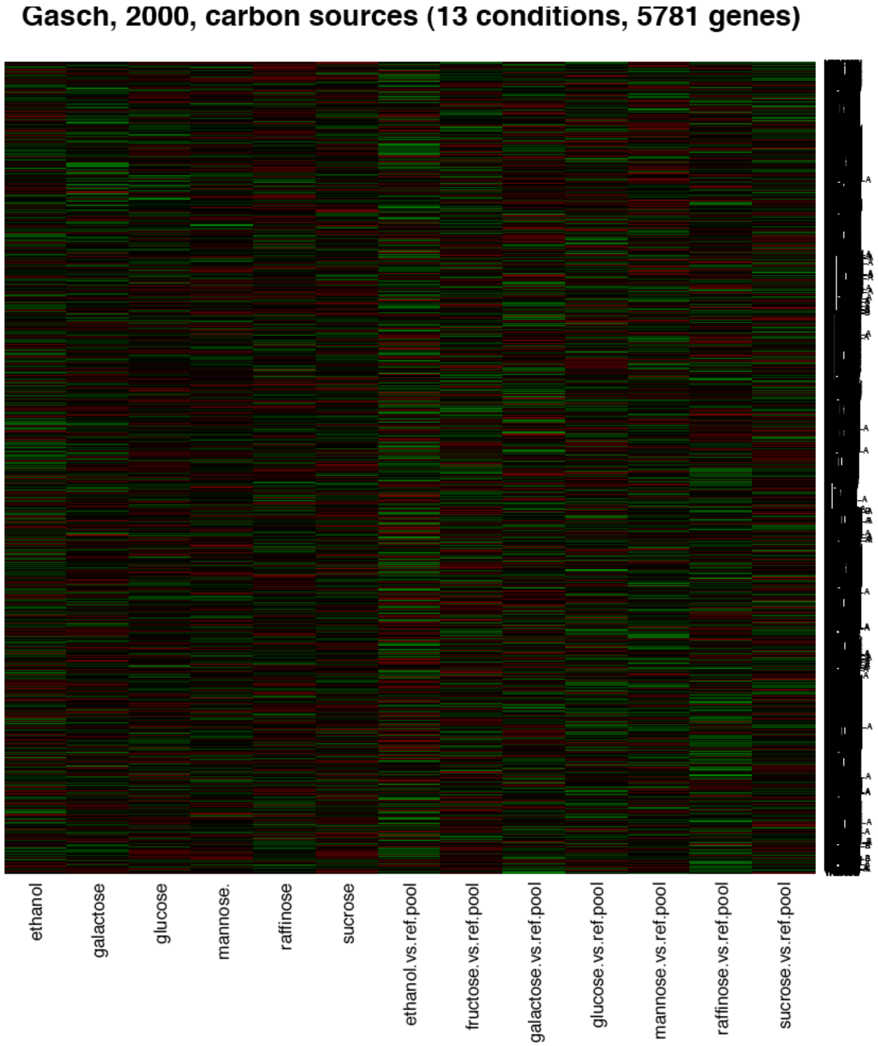
	Carbon sources												
ORF	ethanol	galactose	glucose	mannose.	raffinose	sucrose	ethanol.vs.ref.pool	fructose.vs.ref.pool	galactose.vs.ref.pool	glucose.vs.ref.pool	mannose.vs.ref.pool	raffinose.vs.ref.pool	sucrose.vs.ref.pool
YAL066W	0.71	-1.87	-1.15	-4.90	-1.85	-3.81	-3.34	-0.96	-3.41	-1.04	0.36	-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
…	…	…	…	…	…	…	…	…	…	…	…	…	…

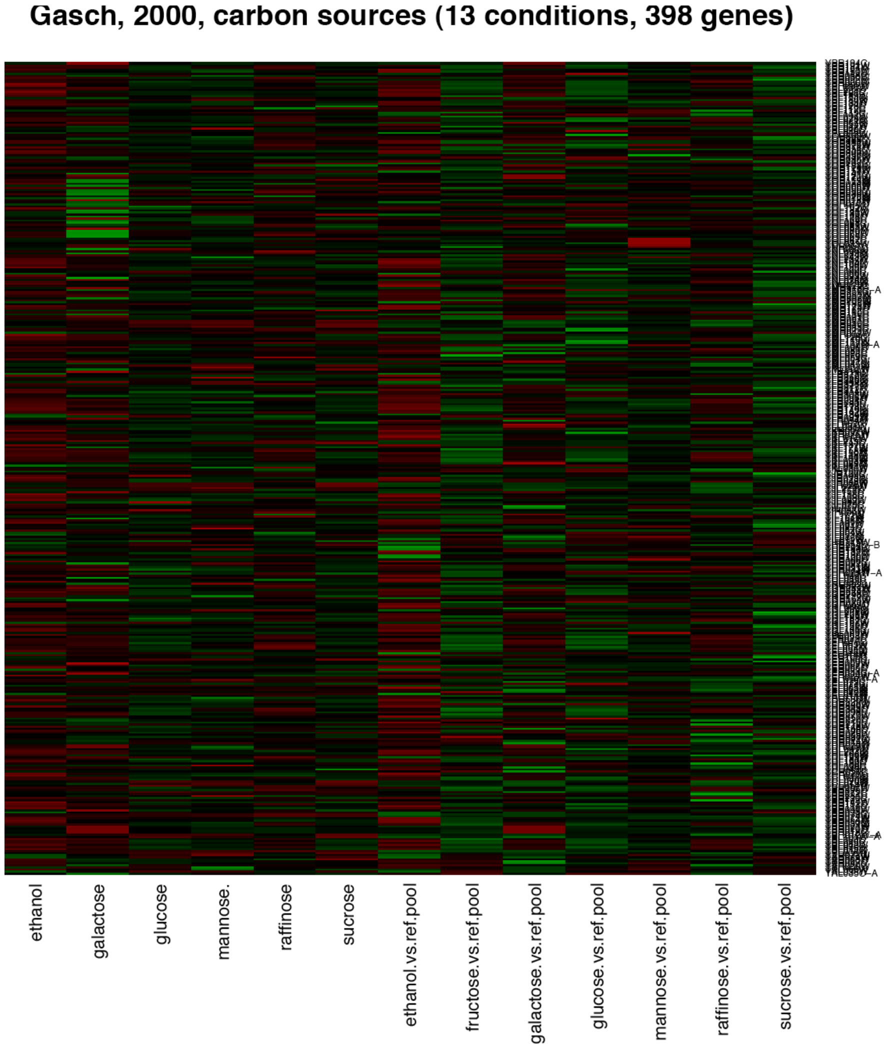
Alternative carbon sources (Gasch, 2000)



Carbon sources: filtering result

- Among the 5781 genes analyzed by Gasch (left heat map), 398 show a significant up- or down-regulation in at least one of the 13 chips on alternative carbon sources (right heat map).





Cancer types (Golub, 1999)

- 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.
- 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.

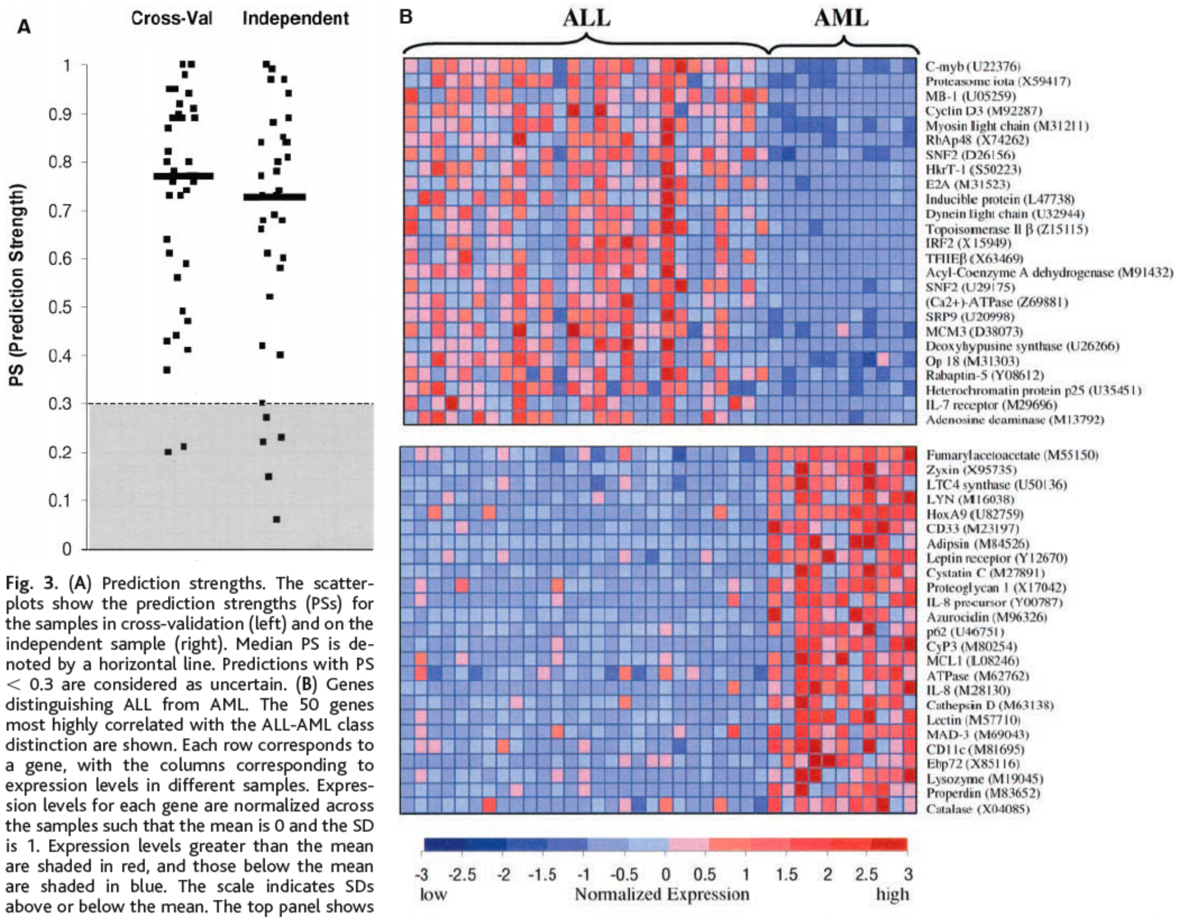
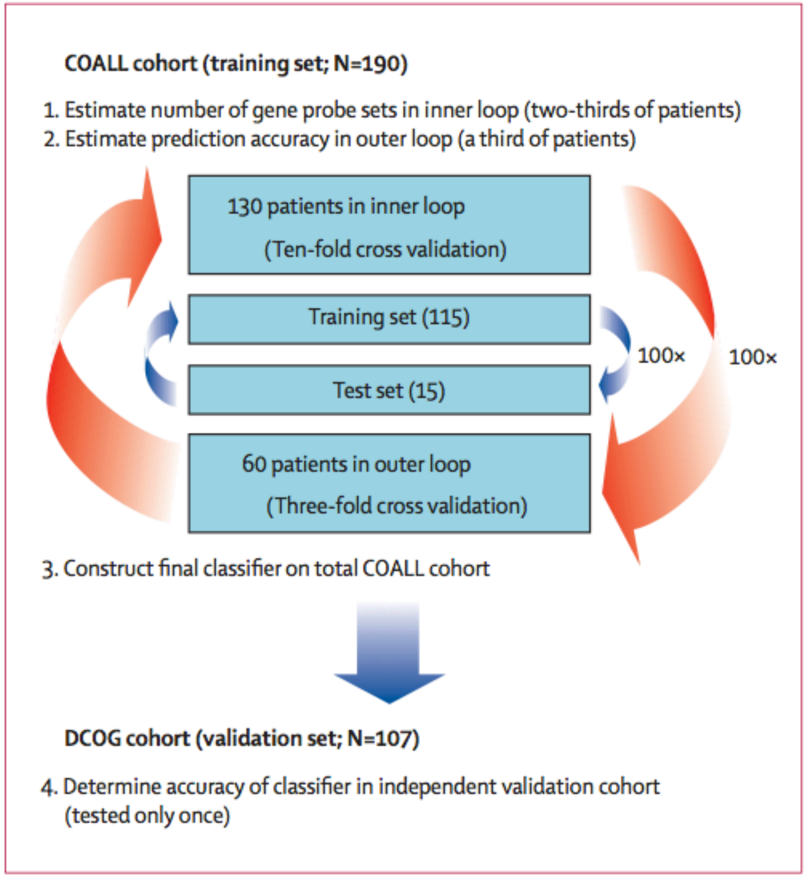


Fig. 3. **(A)** Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. **(B)** Genes distinguishing ALL 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.

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)

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).

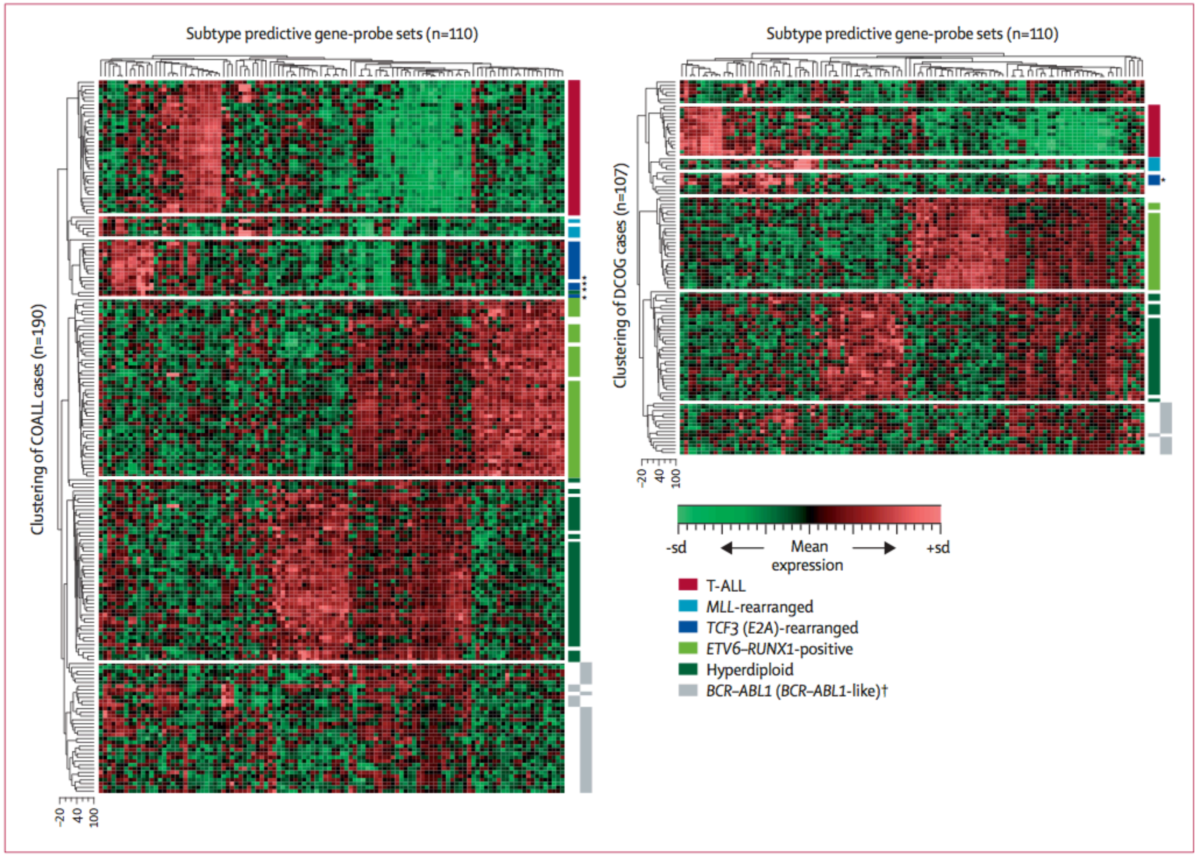


***Figure 2*: Clustering of ALL subtypes by gene-expression profiles**

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.

Den Boer 2009 - Accuracy of the classifier

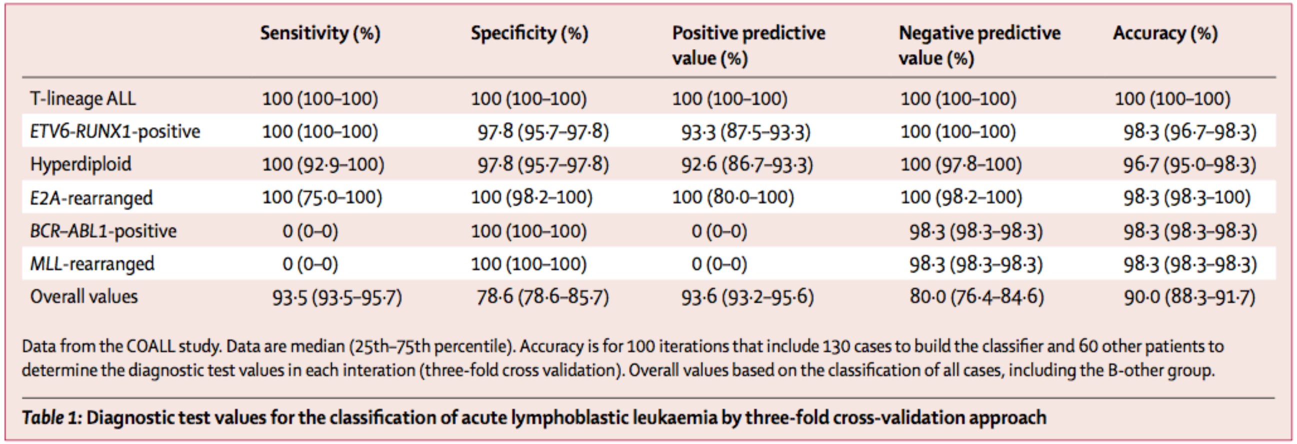
- The signature has an excellent diagnostic value: for the well-represented cancer types, the sensitivity and specificity are >90%.
- **Note:** accuracy is misleading some subtypes have 98% accuracy with 0% sensitivity.

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

$$Sn = TP / (TP + FN)$$

$$Sp = TN / (TN + FP)$$

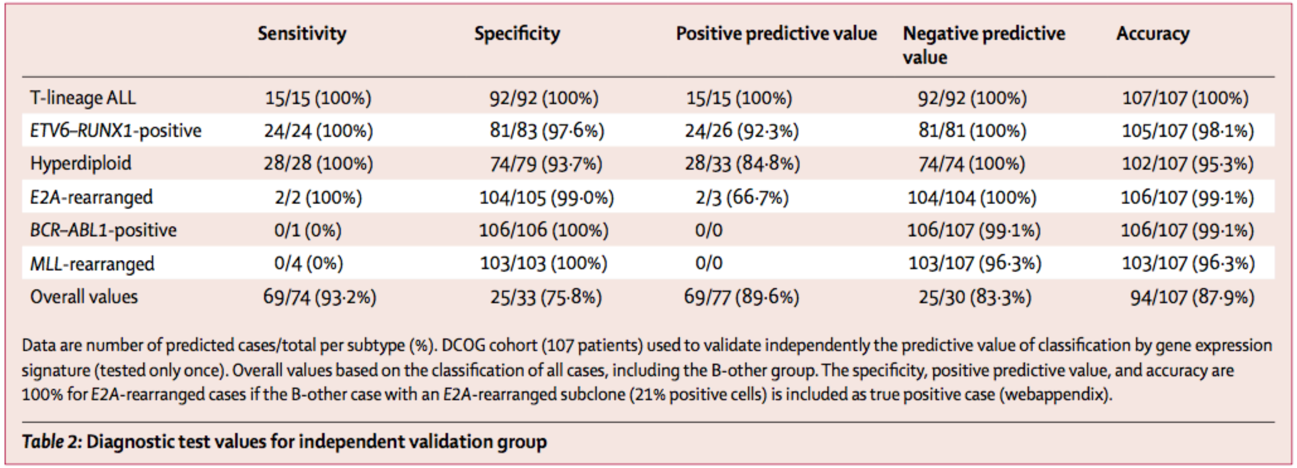
$$PPV = TP / (VP + FP)$$



	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Accuracy (%)
T-lineage ALL	100 (100–100)	100 (100–100)	100 (100–100)	100 (100–100)	100 (100–100)
ETV6-RUNX1-positive	100 (100–100)	97·8 (95·7–97·8)	93·3 (87·5–93·3)	100 (100–100)	98·3 (96·7–98·3)
Hyperdiploid	100 (92·9–100)	97·8 (95·7–97·8)	92·6 (86·7–93·3)	100 (97·8–100)	96·7 (95·0–98·3)
E2A-rearranged	100 (75·0–100)	100 (98·2–100)	100 (80·0–100)	100 (98·2–100)	98·3 (98·3–100)
BCR-ABL1-positive	0 (0–0)	100 (100–100)	0 (0–0)	98·3 (98·3–98·3)	98·3 (98·3–98·3)
MLL-rearranged	0 (0–0)	100 (100–100)	0 (0–0)	98·3 (98·3–98·3)	98·3 (98·3–98·3)
Overall values	93·5 (93·5–95·7)	78·6 (78·6–85·7)	93·6 (93·2–95·6)	80·0 (76·4–84·6)	90·0 (88·3–91·7)

Data from the COALL study. Data are median (25th–75th percentile). Accuracy is for 100 iterations that include 130 cases to build the classifier and 60 other patients to determine the diagnostic test values in each interation (three-fold cross validation). Overall values based on the classification of all cases, including the B-other group.

Table 1: **Diagnostic test values for the classification of acute lymphoblastic leukaemia by three-fold cross-validation approach**



	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
T-lineage ALL	15/15 (100%)	92/92 (100%)	15/15 (100%)	92/92 (100%)	107/107 (100%)
ETV6-RUNX1-positive	24/24 (100%)	81/83 (97·6%)	24/26 (92·3%)	81/81 (100%)	105/107 (98·1%)
Hyperdiploid	28/28 (100%)	74/79 (93·7%)	28/33 (84·8%)	74/74 (100%)	102/107 (95·3%)
E2A-rearranged	2/2 (100%)	104/105 (99·0%)	2/3 (66·7%)	104/104 (100%)	106/107 (99·1%)
BCR-ABL1-positive	0/1 (0%)	106/106 (100%)	0/0	106/107 (99·1%)	106/107 (99·1%)
MLL-rearranged	0/4 (0%)	103/103 (100%)	0/0	103/107 (96·3%)	103/107 (96·3%)
Overall values	69/74 (93·2%)	25/33 (75·8%)	69/77 (89·6%)	25/30 (83·3%)	94/107 (87·9%)

Data are number of predicted cases/total per subtype (%). DCOG cohort (107 patients) used to validate independently the predictive value of classification by gene expression signature (tested only once). Overall values based on the classification of all cases, including the B-other group. The specificity, positive predictive value, and accuracy are 100% for *E2A*-rearranged cases if the B-other case with an *E2A*-rearranged subclone (21% positive cells) is included as true positive case (webappendix).

Table 2: **Diagnostic test values for independent validation group**

- 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.