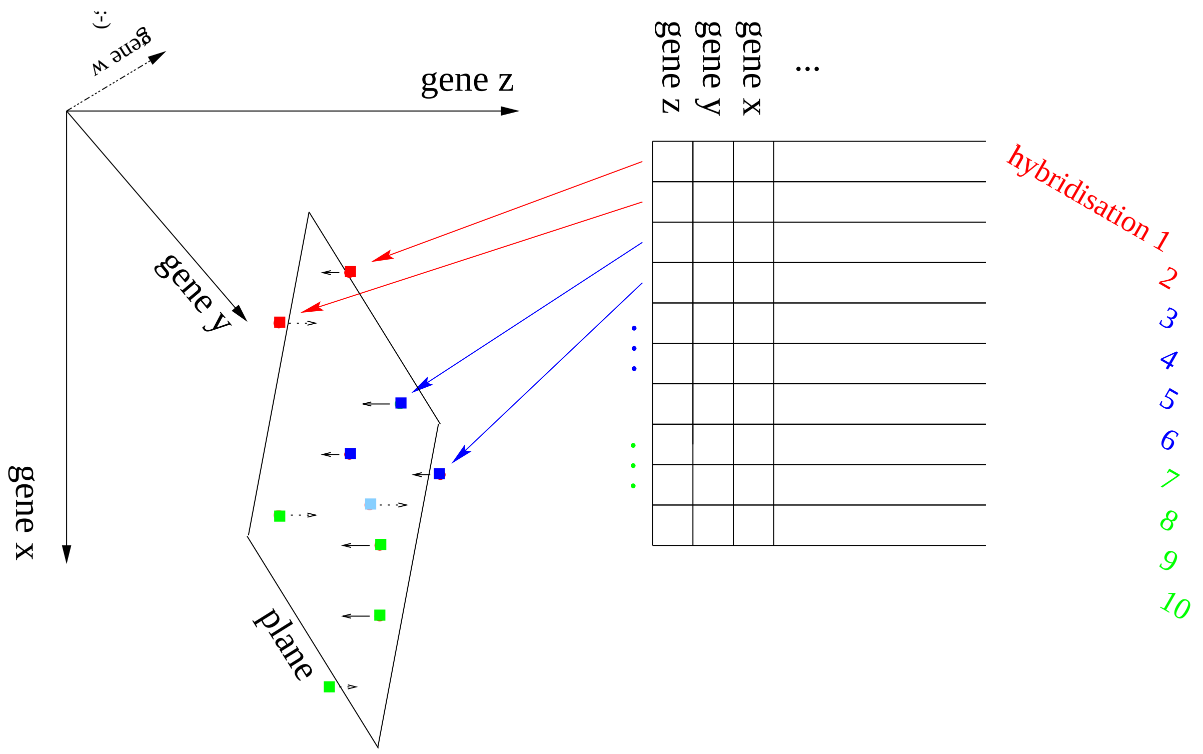
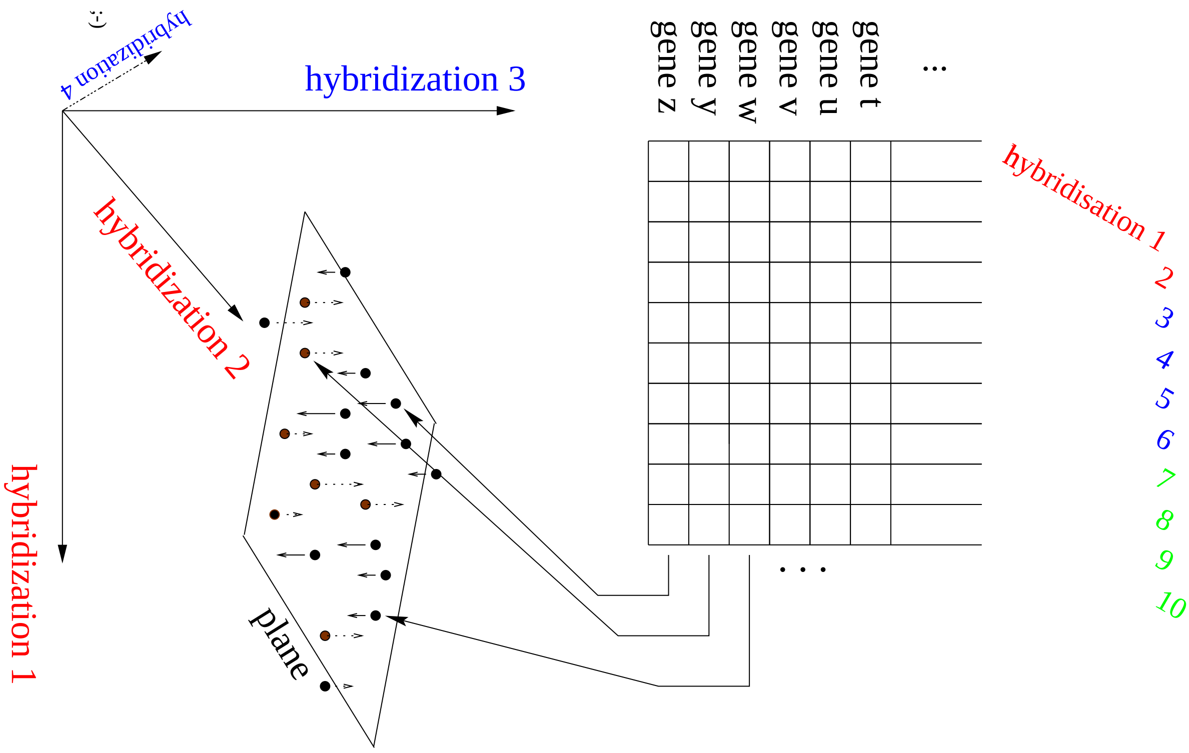
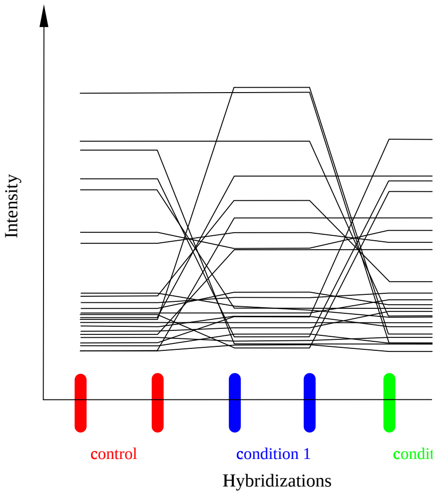


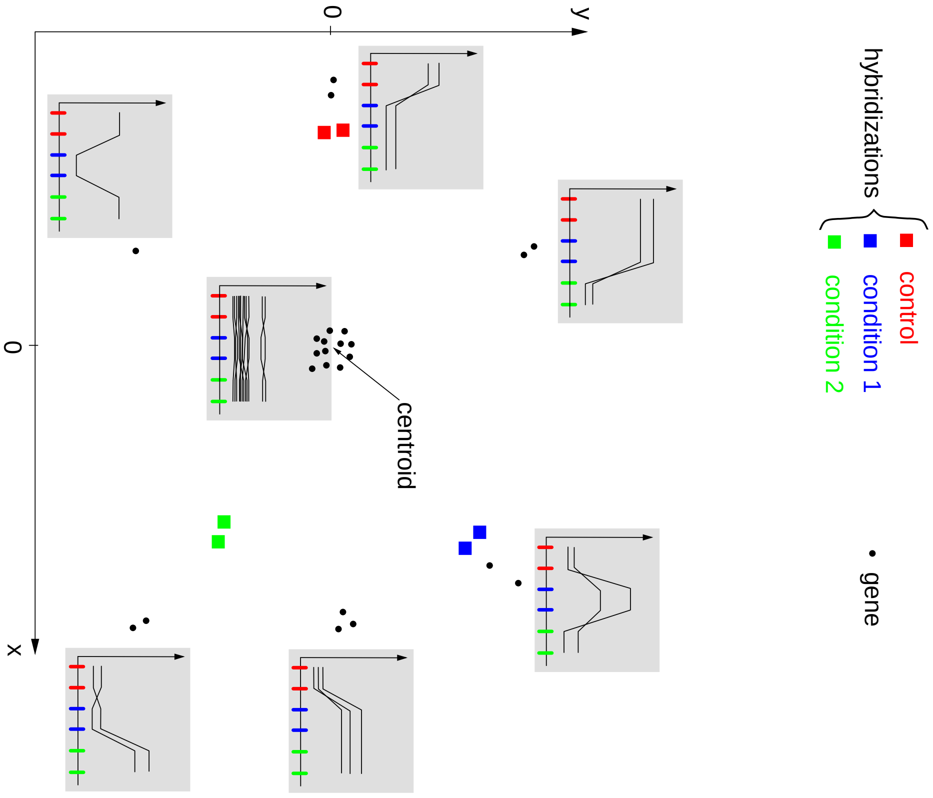
The m columns of a table of n genes × m hybridizations are represented in n-dimensional gene space (three dimensions are shown). n ranges from a few hundred to tenths of thousands. Most microarrays comprise several thousand elements. A plane is selected such that the distance of the hybridization vectors to the plane is minimal, thus conserving point-to-point distances among these vector points as well as possible.



Vice versa, one can regard the gene vectors in hybridization-dimensional space, projecting the genes as done above for the hybridizations.



To look into the interpretablility of such a plot, let me introduce a constructed (virtual) data example. It resembles real data in that the majority of the genes is lowly or not transcribed to a measurable amount. It comprises only 24 genes and differs from the real world in perfect reproducibility among the two hybridizations of each experimental condition.



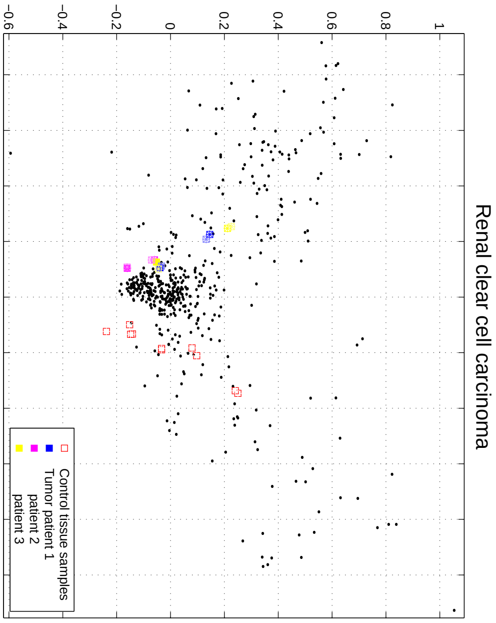
This is the expected output, demonstrating the properties of such projections more clearly than possible by showing a single plot of real data. Gene-clusters are shown together with the according gene profiles. The abscissa represents the first, the ordinate the second principal axis. Both axes are dimensionless. The following properties of such a plot are useful for its interpretation.

- Hybridizations showing high similarity in expression profile, for example because they belong to the same experimental condition, have a short distance in the 24-dimensional gene space, and therefore they will be neighbors in the projection as well.

- Genes with high intensities in a condition are located in the direction of this condition. The two genes located in the direction of the blue condition (upper right corner) are both upregulated particularly in the blue condition.

- Genes particularly downregulated under this condition are located at the opposite side of the centroid. One can regard this gene (lower left corner) as being downregulated in the blue condition. Another valid interpretation is, that it is located in the direction of the bisection line between the red and the green condition because it is equally abundant in these two conditions.

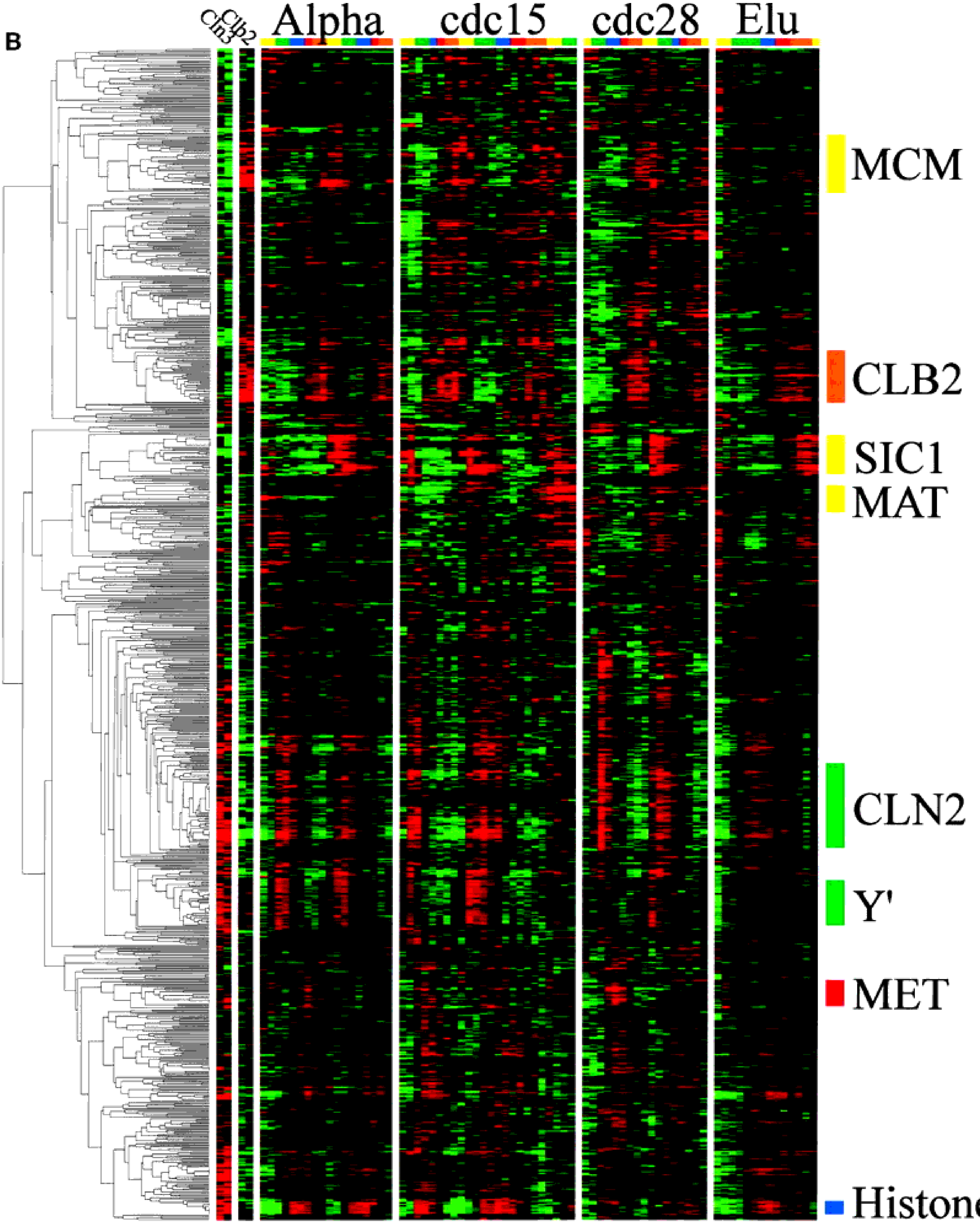
- All genes with unchanged expression, or those not expressed to a measurable amount in any of the conditions under study are located near the centroid. For experiments with comprehensive or complete gene sets, i.e. sets not particularly selected for high expression, the genes that are not detectable will be the majority. The CA plot will show a centric cloud of many genes lacking significantly changed expression throughout the experiment. The outer regions of the plot will contain the so-called 'differential' genes. Their distance to the centroid will reflect the significance of displaying differing expression from the 'average' ones in terms of χ^2 - statistics, which are placed at the center of the plot.

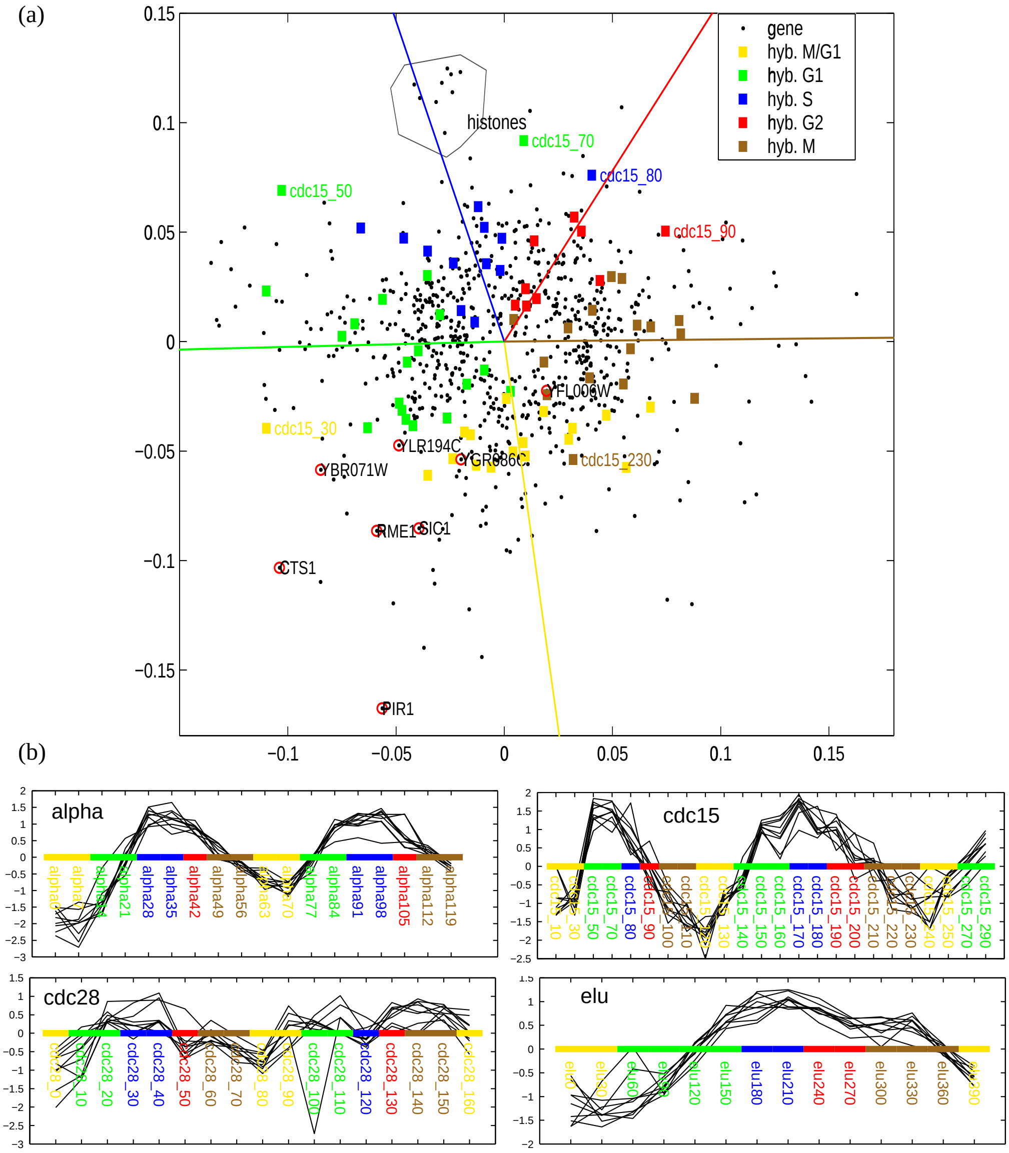


○ Tumor vers. normal (control) tissue biopsies of 3 patients
○ 2 repetitive hybridisations per tissue & patient
○ Double spotting: each spotset drawn as separate 'hybridisation' (dark / light color)
○ 501 most affected ESTs (out of 16128) depicted
● Genes upregulated in tumor are located in the left half of the plot
● Genes downregulated in tumor tissues are drawn in the right half

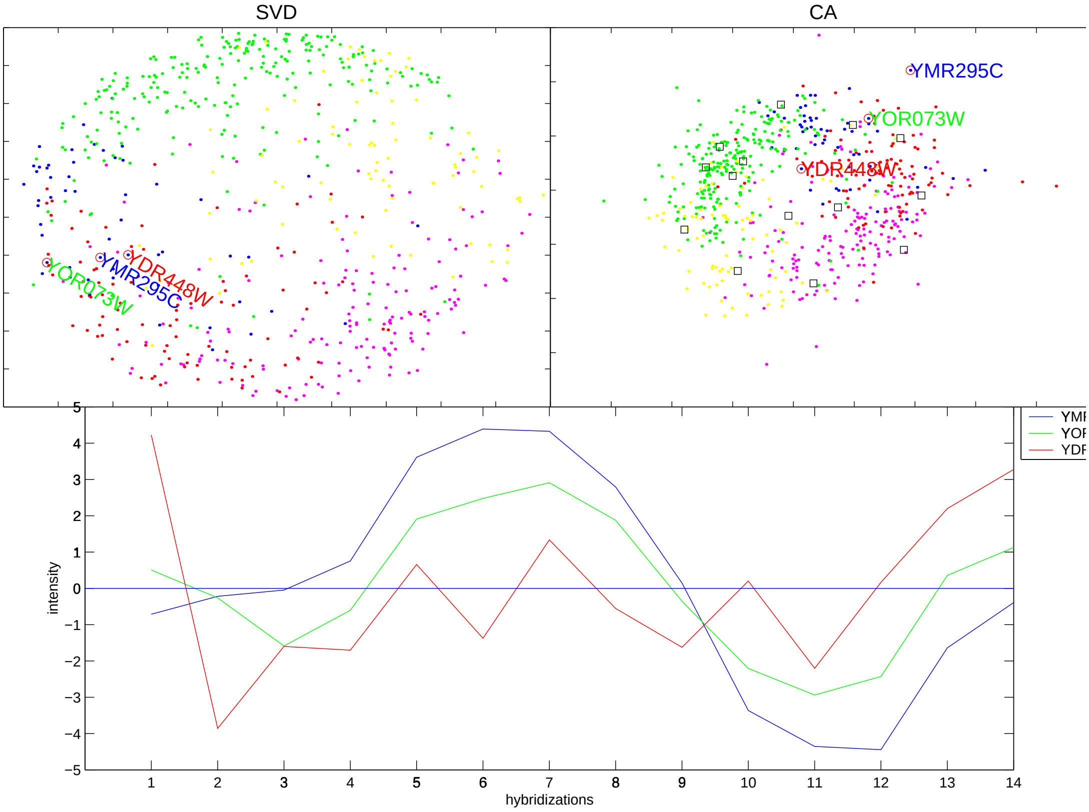
Now knowing at least roughly how to interpret such a plot, let's have a look on a simple real data example. Biopsies of both tumor tissue - the tumors being renal clear cell carcinoma - and normal tissue of the same patient (as a control) have been sampled and hybridised, imaged, normalized, quality filtered and projected.

The tumor tissues on the left half of the plot - tagged by three colors for three different patients - are separated from the normal tissue samples on the right. And all the genes upregulated in the tumor are located on the left side, all those downregulated in the tumor we find in the right half of the plot.


B
Clb2
Cln3
Alpha
cdc15
cdc28
Elu
MCM
CLB2
SIC1
MAT
CLN2
Y'
MET
Histone

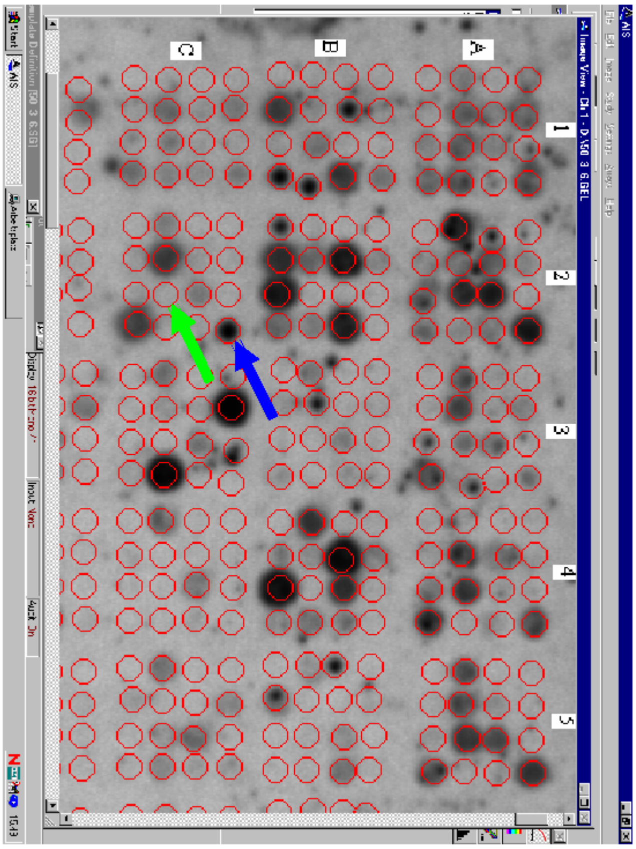


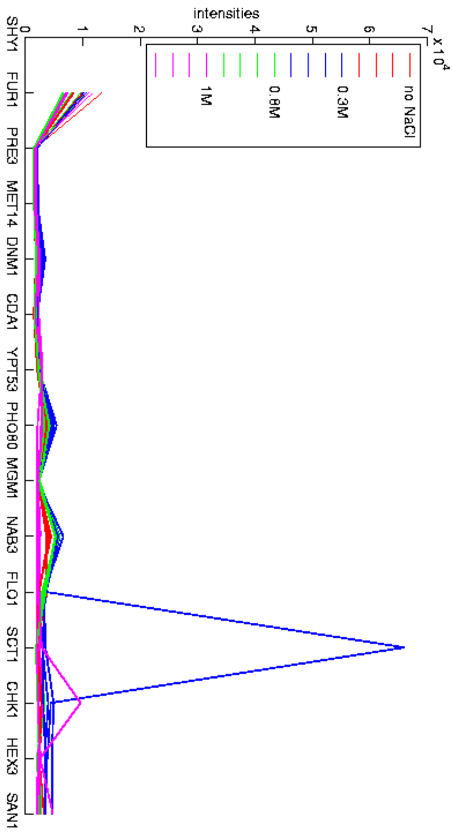
The planar embedding (of exactly the same data) produced by CA shows the hybridzations clearly separated according to their cell cycle phase. They are arranged in circular order of correct sequence. The lines denoting the direction of the hybridization medians emphasize



Alter *et al.* (O. Alter, P. O. Brown, and D. Bostein. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. U.S.A., 97:10101-10106, 2000.) successfully applied singular value decomposition to the analysis of the same data set. In CA plots, the distance of a given gene from the centroid represents the strength of its association with a hybridization lying in the same direction and vice versa. A direct comparison with phase and radius in the visualization of Alter *et al.*[2] shows that this is not necessarily the case in the singular value decomposition alone.

[2] as given e.g. at http://genome-www.stanford.edu/SVD/PNAS/Datasets/Sort_Elutriation.txt





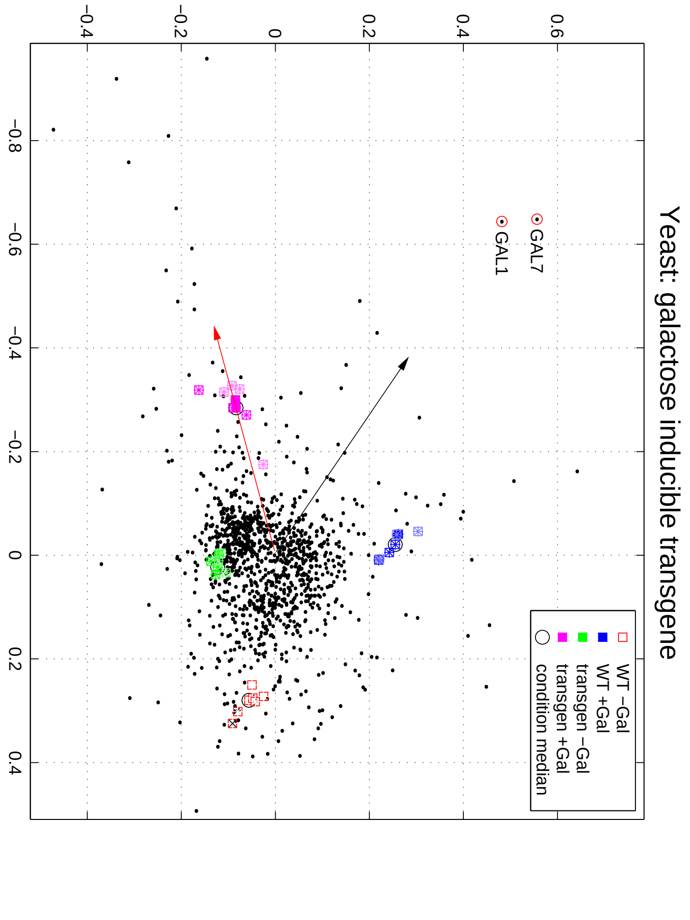
As shown above, it might be useful to visualize deviations of outlying measurements from the expected state. However, data sets frequently comprise severe outliers such as this one.

Wild type yeast was exposed to different concentrations of sodium chloride in the medium (see legend). Normalized transcription intensities of 14 genes are shown in a parallel coordinates plot, lines representing measurements and being color coded according to their particular experimental condition. The plot presents a typical subset of genes, representative with regard to the high number of genes not expressed to a measurable amount. Whereas the different conditions are reproducibly measured for most genes, SCT1 shows one far outlying signal for 0.3M NaCl (blue), which in this case is due to agglutinated label. In the corresponding image (below), the bright dots of unspecifically bound label are common to radioactively labeled targets, whereas the most severe outliers among multichannel data are frequently caused by highly flourescent dust (not shown). The ordinate shows arbitrary (machine dependent) intensity units.

For this reason, a thorough preprocessing is essential. Different normalization algorithms are applied to single and multichannel data for the different meaning of the particular raw intensities. Intensity-, ratio-, and reproducibility filters are applied to extract genes of marked expression for both types of data.

Genes with generally low reproducibility for most of the conditions under study are filtered out by the reproducibility filter. However, with increasing numbers of conditions, discarding all genes with low reproducibility in one of the conditions will leave no gene undiscarded. The same is true for the intensity filter. It is therefore reasonable to use these filters to discard only genes with low abundance or low reproducibility (often coinciding) in all the conditions under study. Thus, outliers as shown above have to be handled by other measures. Otherwise, they would seriously interfere with CA analysis, which in contrast to other methods is not similarity-driven but aims at displaying variance. Any difference to the default state (expected value) such as an outlier, will be regarded as important for the projection. The larger the difference, the more distinctly the corresponding point will be plotted.

We prevent this by choosing the principal axes according to the condition medians only (HMS).



○ WT yeast vers. trangenic strain, induced & uninduced

○ 3 to 4 repetitive hybridisations per genotype and growth condition

○ Double spotting: each spotset drawn as separate 'hybridisation' (dark / light color)

○ genewise median of condition drawn as 'hybridisation' marked by black circle

○ 1173 most affected genes (out of 6103) depicted

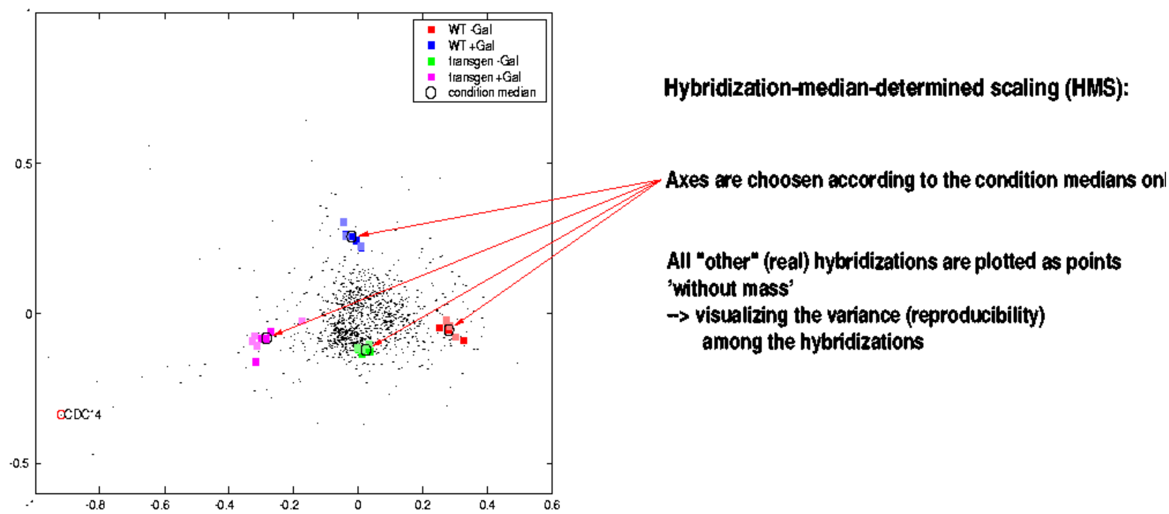
● Black arrow: genes upregulated by galactose both in transgenic and WT strain

● Red arrow: Genes upregulated only in the mutant strain

Let me first introduce a new data example. Here, a transgene was transfected into yeast cells under the control of a galactose inducible promoter:

Red empty boxes are WT yeast without galactose in the medium, blue WT with galactose, green the transgenic strain without galactose and pink transgenic with galactose and here we would expect the transgene to be induced and genes that follow the transgene. The aim was to separate those genes from all the genes which are upregulated by galactose in yeast anyway, that is: also in the WT strain.

The bisection line between WT and transgenic strain with galactose (black arrow) points to the genes induced by galactose both in the transgenic and in the WT strain. There we find genes like Gal 7 and Gal 1. The red arrow points to the genes upregulated specifically only in the transgenic strain - those are the genes the experimenter intended to look at with this experiment.

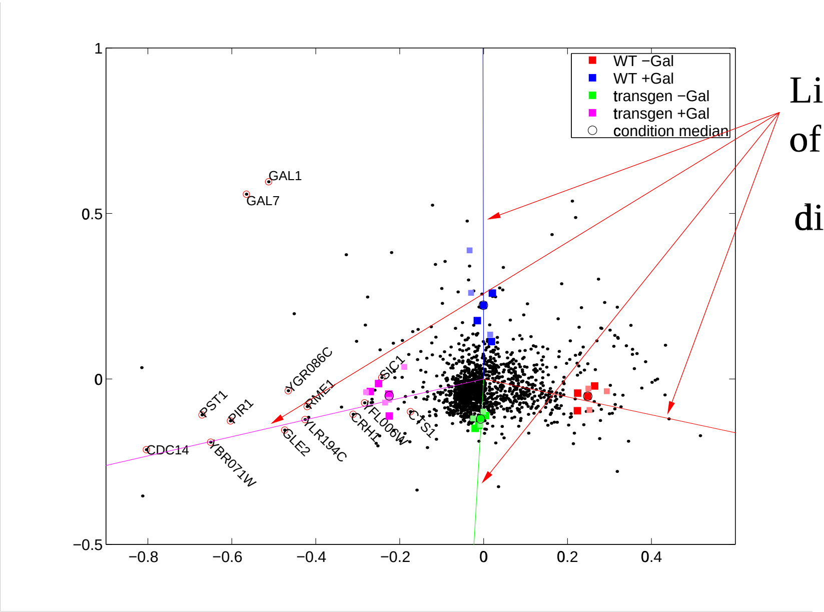


Hybridization-median-determined scaling (HMS):

Axes are choosen according to the condition medians onl

All "other" (real) hybridizations are plotted as points 'without mass'
--> visualizing the variance (reproducibility) among the hybridizations

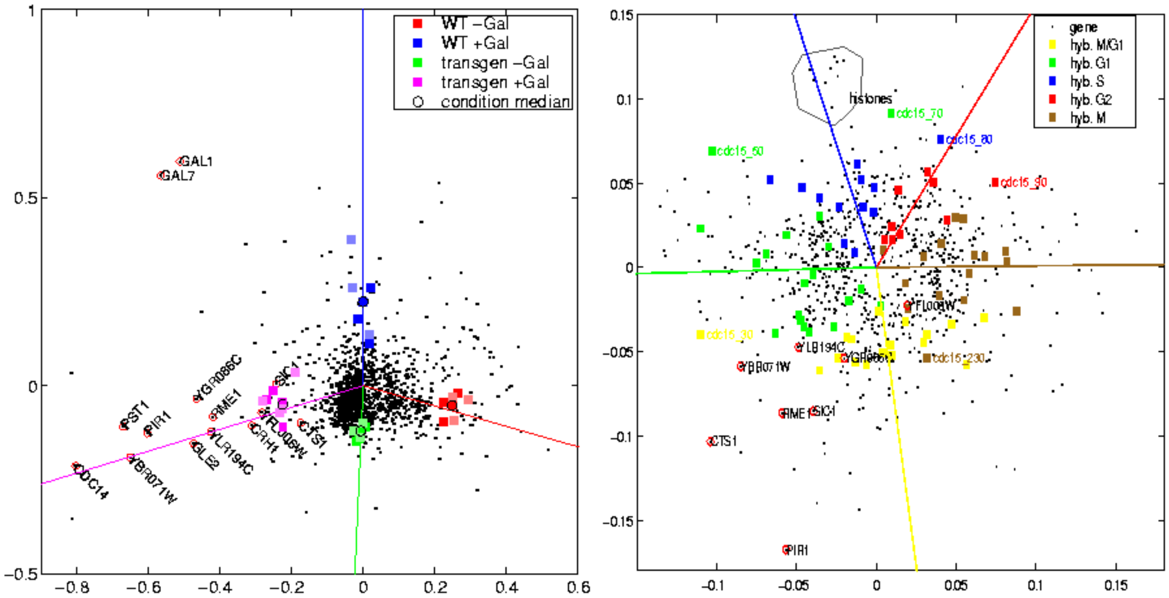
Typically, replicate hybridizations are performed for each condition under study leading to several values for one gene/condition pair. The number of such repeated hybridizations is often small. I therefore represent these values by their gene-wise median rather than their gene-wise average because the median is less sensitive to outliers. The need remains, though, to visualize also the original data and not only the median since they contain valuable information about experimental variance and quality of individual hybridizations. In fact, CA offers the possibility to reflect both aspects. To this end, CA is first effected by using the gene-wise medians, determining the coordinate system to embed the original hybridization intensities. These data points are then referred to as supplementary points or points without mass. Thus



Lines to standard coordinate
of condition-medians

directions of association

Due to all these precautions and given a sufficient number of repeated hybridizations, the variance explained by a CA plot will largely reflect biological changes, displaying the significance of differences both among the genes and among the hybridizations in terms of the χ^2-statistic. The power of the CA technique however is that it is able to show associations between genes and hybridizations. To fully exploit this property, it is necessary to examine the exact directions of gene-association with the experimental conditions. These are given by the standard coordinates of the according condition medians rather than by their principle coordinates. An experiments represented in standard coordinates can be viewed as a virtual gene having its entire mass (intensity) in this particular experiment. Thus, it is the gene of highest possible association with this experiment, able to "represent" the experiment in "gene-space". Plotting the standard coordinates directly would cause all principle coordinates to shrink into a small area in the middle of the plot. The introduction of lines representing the standard coordinates is of great help in the interpretation of the plots, relating genes and conditions to each other and circumventing direct plotting.

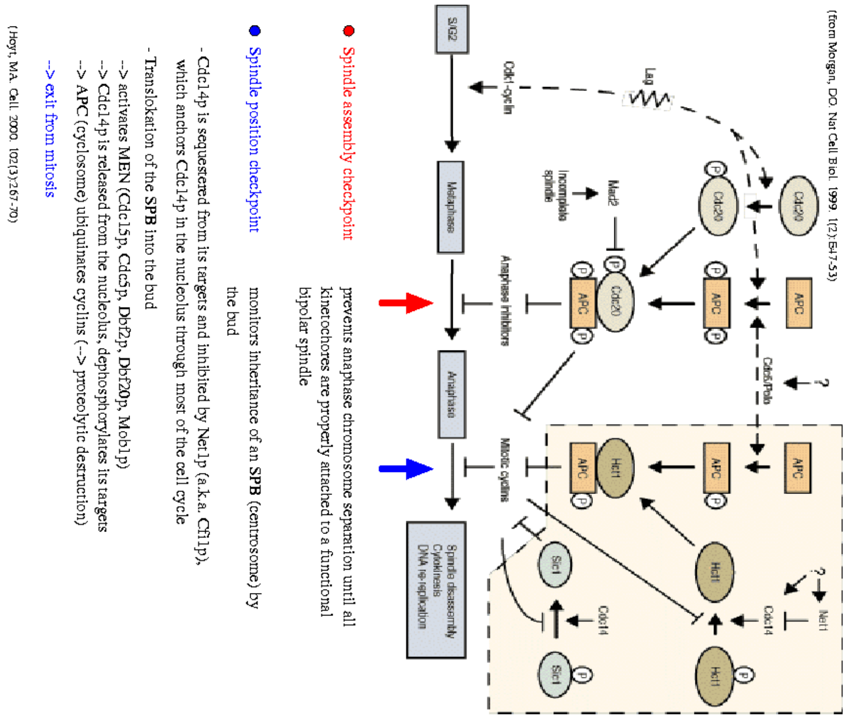


- **SIC1, known to be accumulated in a Cdc14p dependent fashion [1]**
- **CTS1, belongs to the cluster of SIC1 co-regulated genes [2]**
- **RME1, CRH1, PST1 known to be cell-cycle regulated with peaks in mitosis/G1 transition, G1 or late G1, respectively but have not yet been described in association with Cdc14p activity.**
- **YBR071W, PIR1, YGR086C, YLR194C, YFL006 not yet annotated to be cell-cycle regulated but in agreement with the data of ref. 2 (mitosis/G1 transition)**
- **GLE2 (nuclear pore protein): ? unknown function in Cdc14p activation context**

[1] D. O. Morgan. Regulation of the APC and the exit from mitosis. Nat. Cell Biol., 1(2):E47-53, 1999.
[2] P. T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast [...]. Mol. Biol. Cell, 9:3273-3297, 1998.

Some Biology: Genes in the direction of galactose induced transgenic yeast are those specifically upregulated upon *CDC14* induction as opposed to genes activated by galactose also in the WT strain, like *GAL1* and *GAL7*. This subtraction has been achieved purely computationally and is based on the provision of galactose activated genes in wild type as a separate condition. The set of genes associated specifically to the Cdc14p overproducing condition comprises *CDC14* itself as well as *SIC1*, known to be accumulated in a Cdc14p dependent fashion [1] and *CTS1* which belongs to the cluster of *SIC1* co-regulated genes [2]. *RME1*, *CRH1* and *PST1* are known to be cell cycle regulated with peaks in mitosis/G1 transition, G1 or late G1, respectively but have not yet been described in association with Cdc14p activity. *YBR071W*, *PIR1*, *YGR086C*, *YLR194C*, and *YFL006W* have not been annotated to be cell cycle regulated, but these results show that they are. This is in agreement with the data of Spellman *et al.* (right panel, genes marked by red circles), which also show these genes to be transcribed during mitosis/G1 transition. The role of the nuclear pore protein *GLE2* in a Cdc14p activation context remains unclear.

The biological context of CDC14 is sketched in the following transperancy.



(from Morgan, DO. Nat Cell Biol. 1999. 1(2):E47-53)

● Spindle assembly checkpoint — prevents anaphase chromosome separation until all kinetochores are properly attached to a functional bipolar spindle

● Spindle position checkpoint — monitors inheritance of an **SPB** (centrosome) by the bud

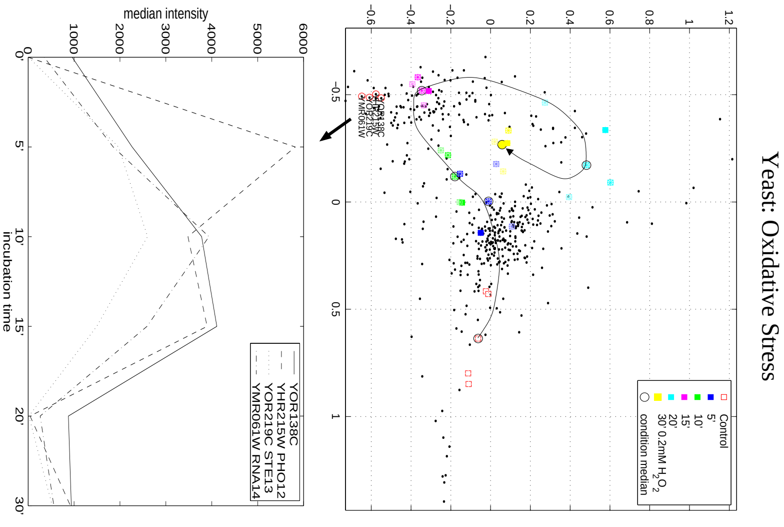
- Cdc14p is sequestered from its targets and inhibited by Net1p (a.k.a. Cfi1p), which anchors Cdc14p in the nucleolus through most of the cell cycle

- Translokation of the **SPB** into the bud

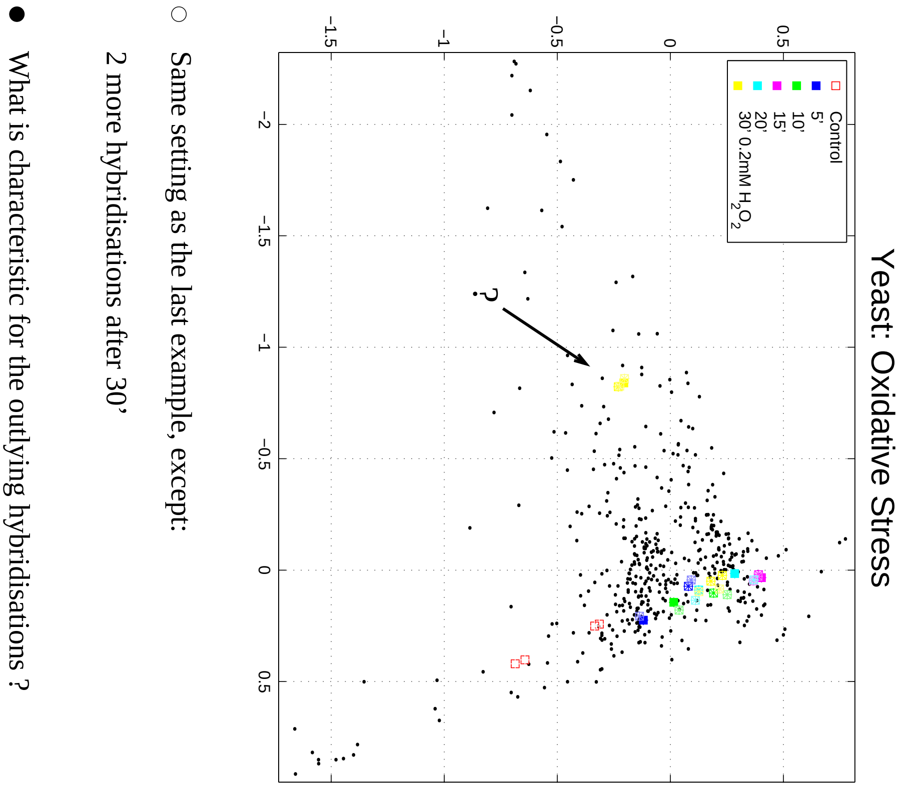
--> activates MEN (Cdc15p, Cdc5p, Dbf2p, Dbf20p, Mob1p)
--> Cdc14p is released from the nucleolus, dephosphorylates its targets
--> APC (cyclosome) ubiquinates cyclins (--> proteolytic destruction)

--> exit from mitosis

(Hoyt, MA. Cell. 2000. 102(3):267-70)



The last data example relates to the second part of the talk. A time course has been recorded for wild type *S. cerevisiae* cells under oxidative stress. The thin black arrow draws the chronological progression of the experiment. The cells responding to 0.2mM hydrogenperoxide in their medium show quite a leap in expression behaviour between 15 and 20 minutes that includes the downregulation of genes which had been switched on in the initial phase of the response. Four of those are flagged. Their gene profiles are plotted below. They are switched on initially and are being downregulated somewhere between 15 and 20 minutes.

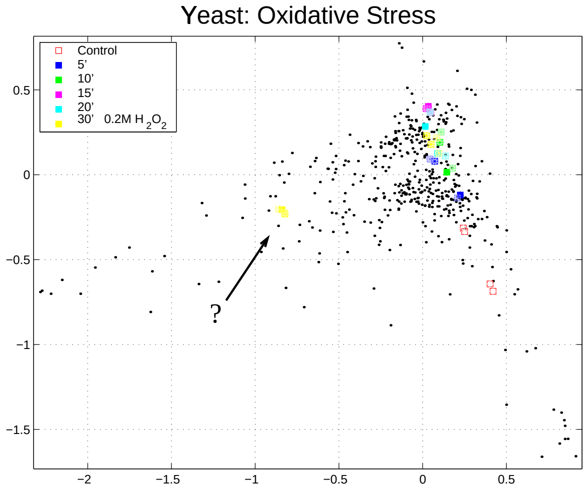

Yeast: Oxidative Stress

- ○ Same setting as the last example, except:

  2 more hybridisations after 30'
- ● What is characteristic for the outlying hybridisations ?

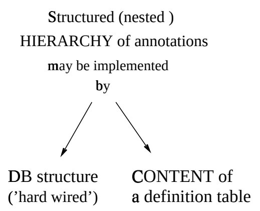
In the above example there is obviously something wrong. It is exactly the same experimental setting as before but now the yellow 30' condition is divided into two clusters located far away from each other and distorting the nice picture of the previous plot.

And we want to know why. What is wrong with the outliers? In other words: Can we find features in the experimental description which are characteristic for the outlying cluster? Are there annotation values overrepresented in the cluster? Or are there values missing or underrepresented in the cluster?



What is characteristic for the outlying hybridisations ?

In practise: Selecting these outliers, scanning for at least 2-fold over-or underrepresented annotation values results in values belonging to only 8 out of 111 annotations, listed in the next transperancy.


Structured (nested)
HIERARCHY of annotations
may be implemented
by

DB structure
('hard wired')

CONTENT of
a definition table

Given the tool of a relational DBMS, storage of e.g. experiment annotations can be implemented in different ways.

by STRUCTURE

annotations stored as attributes
(= columns, fields)

experiment	array_source	array_support	spotted_material
1	self_made	nylon	PCR
2	genome_systems	polypropylene	PCR
3	self_made	glas	colonies
⋮			

by CONTENT

... stored in table content

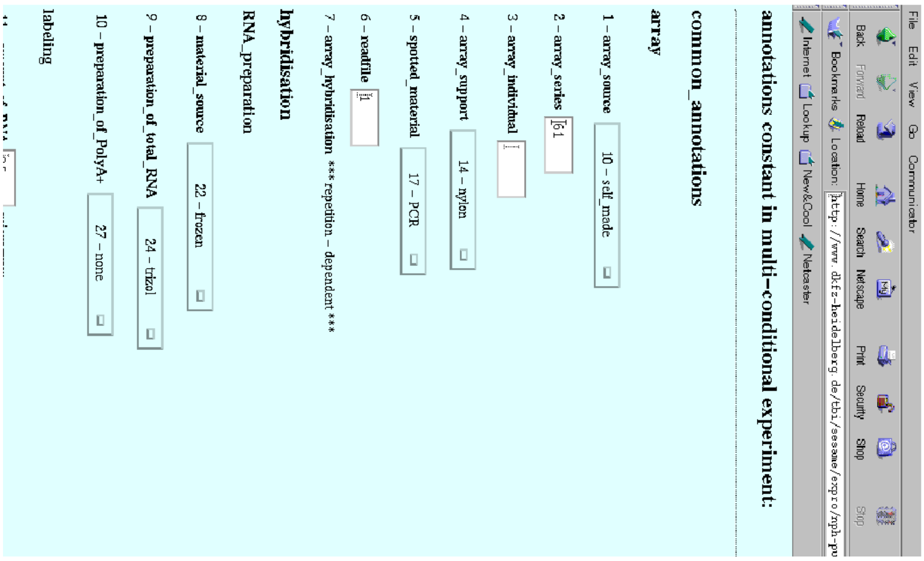
DEFINITIONS

annotation	defined value
array_source	self_made
array_source	genome_systems
array_support	glass
array_support	nylon
array_support	polypropylene
spotted_material	colonies
spotted_material	PCR

ANNOTATIONS

experiment	annotation	value
1	array_source	self_made
1	array_support	nylon
1	spotted_material	PCR
2	array_source	genome_systems
⋮		

The parameter names such as "array source", let us refer to them as "annotations", may become the attributes (column names) of a single table. Another possibility is to make them the content (the tuples) of a first table, whose only purpose is to define the annotations along with the values they may take. Here, a second table is needed to store the actual values taken in particular experiments.



File Edit View Go Communicator
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.dkfz-heidelberg.de/tbi/sesame/expro/nph-pu
Internet Lookup New&Cool Netcaster

annotations constant in multi–conditional experiment:

common_annotations

array

1 – array_source 10 – self_made

2 – array_series 61

3 – array_individual

4 – array_support 14 – nylon

5 – spotted_material 17 – PCR

6 – readfile 1

7 – array_hybridisation *** repetition – dependent ***

hybridisation

RNA_preparation

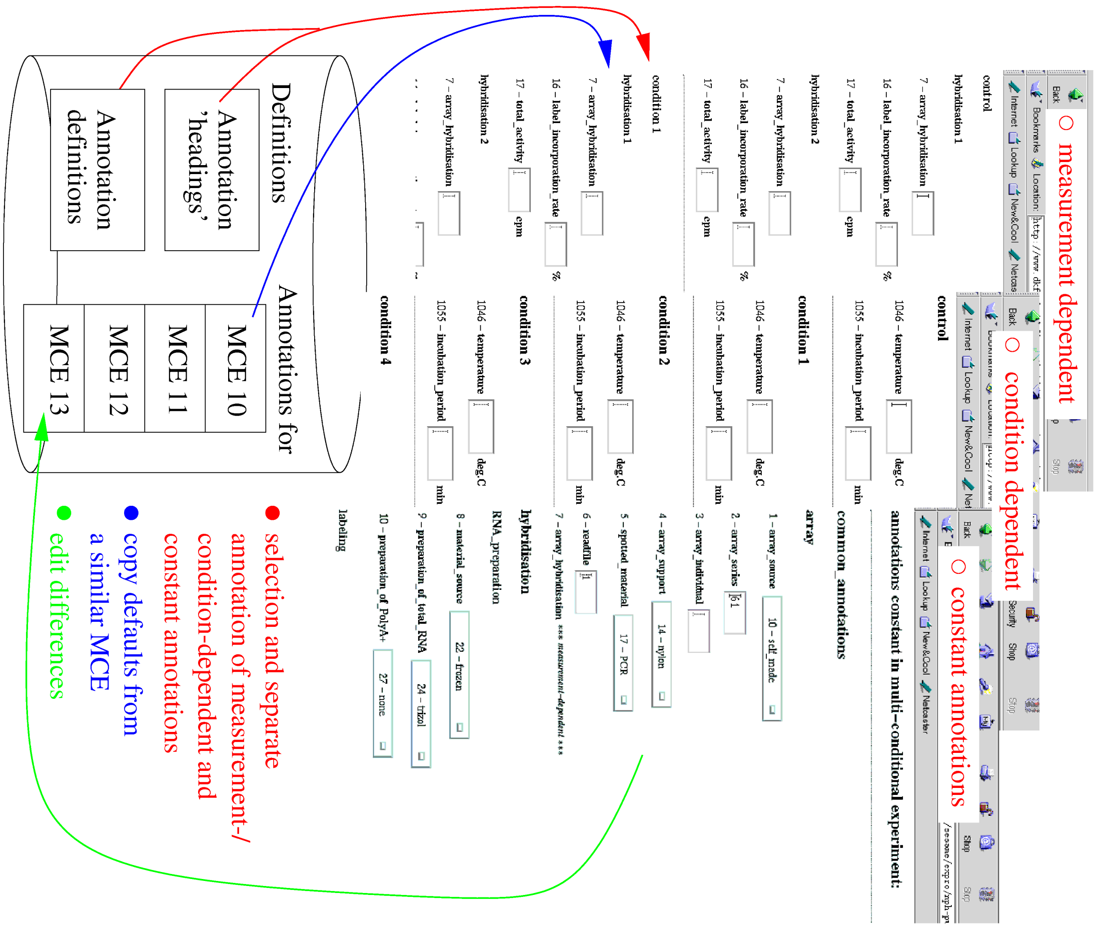
8 – material_source 22 – frozen

9 – preparation_of_total_RNA 24 – trizol

10 – preparation_of_PolyA+ 27 – none

labeling

The content of these definition tables serves as meta data to compile html forms used during the process of annotating an experiment.



control

hybridisation 1

7 – array_hybridisation

16 – label_incorporation_rate %

17 – total_activity cpm

hybridisation 2

7 – array_hybridisation

16 – label_incorporation_rate %

17 – total_activity cpm

condition 1

hybridisation 1

7 – array_hybridisation

16 – label_incorporation_rate %

17 – total_activity cpm

hybridisation 2

7 – array_hybridisation

control

1046 – temperature deg.C

1055 – incubation_period min

condition 1

1046 – temperature deg.C

1055 – incubation_period min

condition 2

1046 – temperature deg.C

1055 – incubation_period min

condition 3

1046 – temperature deg.C

1055 – incubation_period min

condition 4

annotations constant in multi–conditional experiment:

common_annotations

array

1 – array_source 10 – self_made

2 – array_series 61

3 – array_individual

4 – array_support 14 – nylon

5 – spotted_material 17 – PCR

6 – readfile 1

7 – array_hybridisation *** *measurement–dependent* ***

hybridisation

RNA_preparation

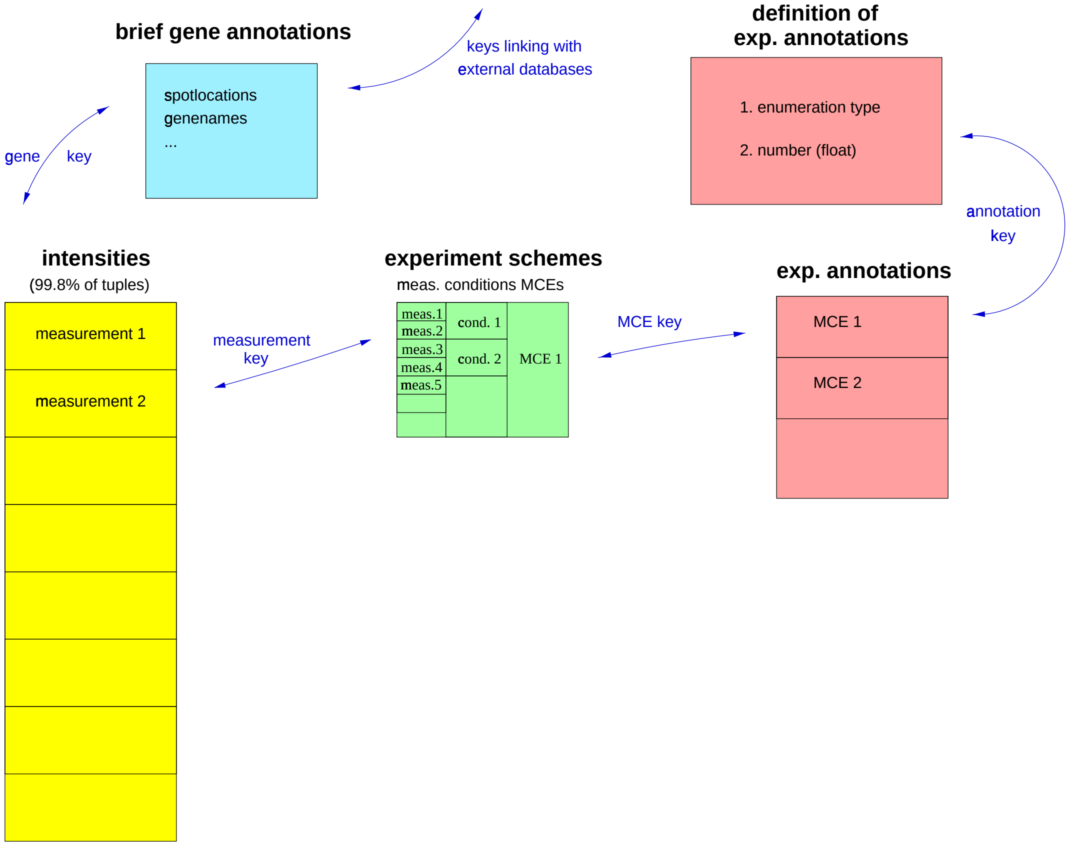
8 – material_source 22 – frozen

9 – preparation_of_total_RNA 24 – trizol

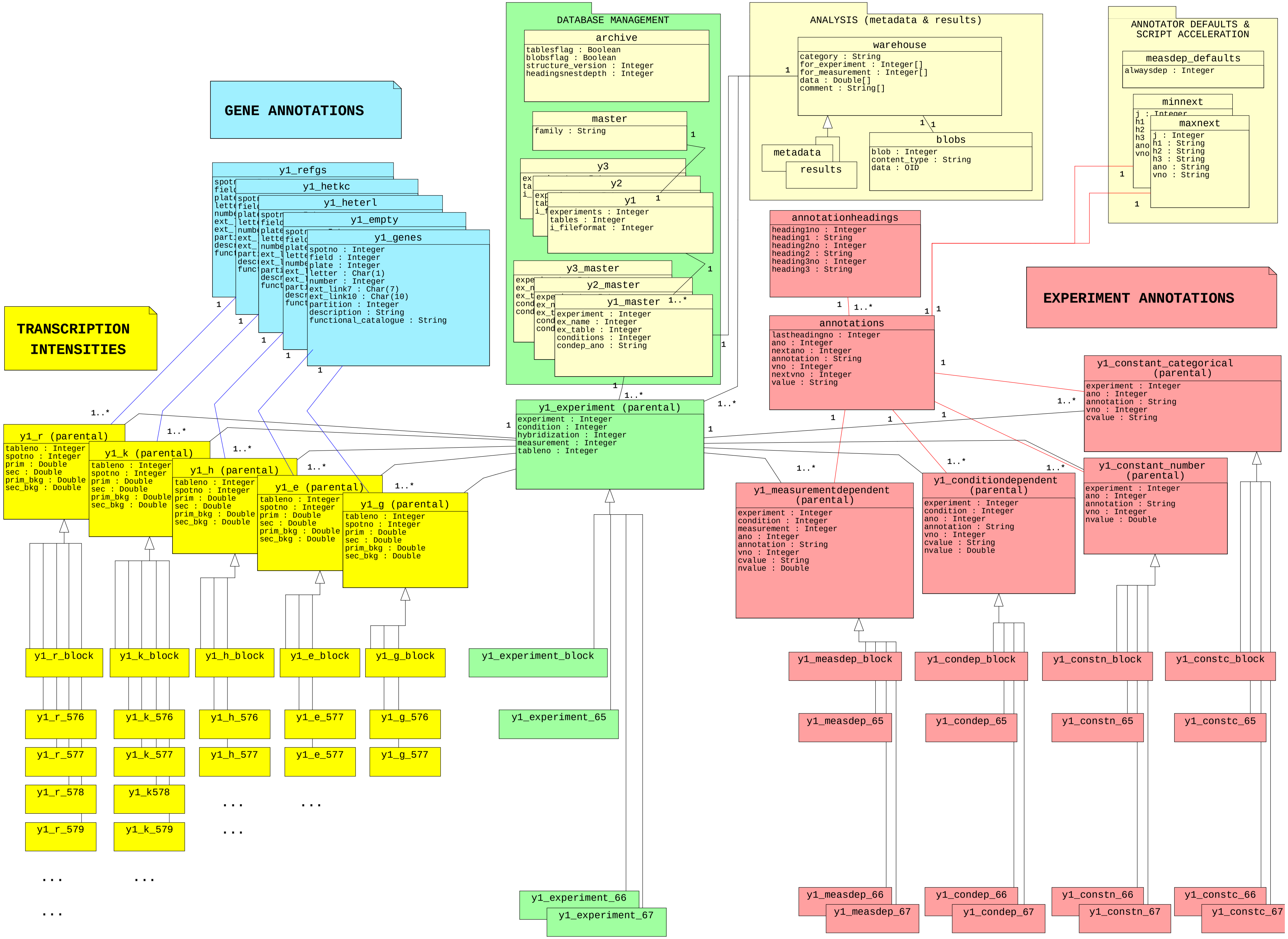
10 – preparation_of_PolyA+ 27 – none

labeling

A convenient way minimizing efforts in annotation. Every piece of information has to be entered only once. The annotation process may start with copying default values from the most



Above transparency shows them (lower right boxes) in database context. Gene annotations, signal intensities (please note the percentage!), and experiment annotations are displayed in blue, yellow and red, respectively, also in the next transparency. The gene annotations are linked both with the transcription intensities and with public external gene databases (e.g. GO) in order to enable explicit characterization of genes showing a particualar transcription behaviour. The intensities are stored as measurements. A measurement (i.e. a hybridization for radioactive or a single channel for multi-channel data) comprises a single value for each spot on the microarray. Experiment schemes record for each measurement which hybridization and experimental condition it belongs to, and which multiconditional experiment (MCE) this condition is contained in. The experiment schemes are the 'storekeepers' of the database, relating intensity data with experiment annotations, which allow for explicit characterization of measurements showing a particular expression pattern.


GENE ANNOTATIONS
TRANSCRIPTION INTENSITIES
DATABASE MANAGEMENT
archive
tablesflag : Boolean
blobsflag : Boolean
structure_version : Integer
headingsnestdepth : Integer
master
family : String
y3
y2
y1
experiments : Integer
tables : Integer
i_fileformat : Integer
y3_master
y2_master
y1_master
experiment : Integer
ex_name : Integer
ex_table : Integer
conditions : Integer
condep_ano : String
ANALYSIS (metadata & results)
warehouse
category : String
for_experiment : Integer[]
for_measurement : Integer[]
data : Double[]
comment : String[]
blobs
blob : Integer
content_type : String
data : OID
metadata
results
ANNOTATOR DEFAULTS & SCRIPT ACCELERATION
measdep_defaults
alwaysdep : Integer
minnext
maxnext
j : Integer
h1 : String
h2 : String
h3 : String
ano : String
vno : String
y1_refgs
y1_hetkc
y1_heterl
y1_empty
y1_genes
spotno : Integer
field : Integer
plate : Integer
letter : Char(1)
number : Integer
ext_link7 : Char(7)
ext_link10 : Char(10)
partition : Integer
description : String
functional_catalogue : String
EXPERIMENT ANNOTATIONS
annotationheadings
heading1no : Integer
heading1 : String
heading2no : Integer
heading2 : String
heading3no : Integer
heading3 : String
annotations
lastheadingno : Integer
ano : Integer
nextano : Integer
annotation : String
vno : Integer
nextvno : Integer
value : String
y1_constant_categorical (parental)
experiment : Integer
ano : Integer
annotation : String
vno : Integer
cvalue : String
y1_constant_number (parental)
experiment : Integer
ano : Integer
annotation : String
vno : Integer
nvalue : Double
y1_conditiondependent (parental)
experiment : Integer
condition : Integer
ano : Integer
annotation : String
vno : Integer
cvalue : String
nvalue : Double
y1_measurementdependent (parental)
experiment : Integer
condition : Integer
measurement : Integer
ano : Integer
annotation : String
vno : Integer
cvalue : String
nvalue : Double
y1_experiment (parental)
experiment : Integer
condition : Integer
hybridization : Integer
measurement : Integer
tableno : Integer
y1_r (parental)
y1_k (parental)
y1_h (parental)
y1_e (parental)
y1_g (parental)
tableno : Integer
spotno : Integer
prim : Double
sec : Double
prim_bkg : Double
sec_bkg : Double
y1_r_block
y1_k_block
y1_h_block
y1_e_block
y1_g_block
y1_experiment_block
y1_measdep_block
y1_condep_block
y1_constn_block
y1_constc_block
y1_r_576
y1_r_577
y1_r_578
y1_r_579
y1_k_576
y1_k_577
y1_k578
y1_k_579
y1_h_576
y1_h_577
y1_e_577
y1_e_577
y1_g_576
y1_g_577
...
y1_experiment_65
y1_experiment_66
y1_experiment_67
y1_measdep_65
y1_measdep_66
y1_measdep_67
y1_condep_65
y1_condep_66
y1_condep_67
y1_constn_65
y1_constn_66
y1_constn_67
y1_constc_65
y1_constc_66
y1_constc_67
1
1..*