Statistics & Data Analysis for High-Density Oligonucleotide Microarray Data

Alexander Ploner

Medical Epidemiology & Biostatistics
Karolinska Institutet

April 7, 2005

Overview

1. Technology & Workflow

2. Low-level analysis

   - extracting expression values
   - normalizing expression values
   - quality control & gene selection

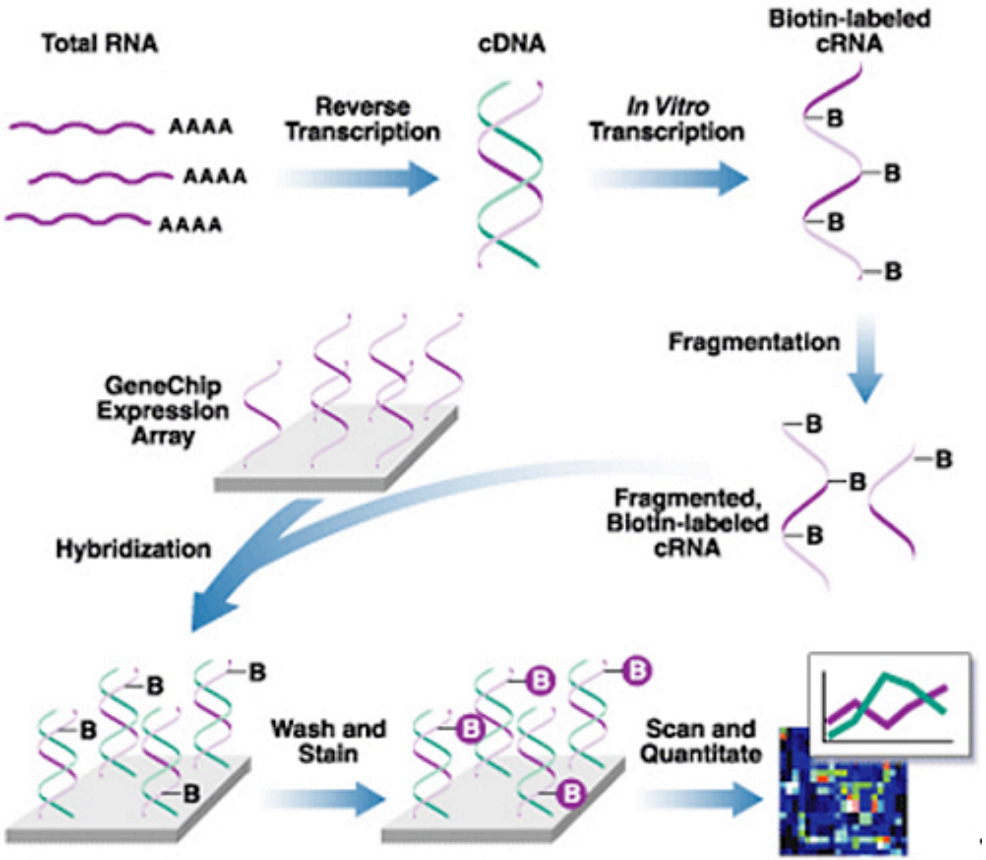
3. High-level analysis

   - class discovery
   - class prediction
   - gene discovery

4. Summary

The Affymetrix Technology

- Each target sequence is probed by multiple(10-20) 25-mers
- Oligonucleotides are photolithographically synthesized in situ on each chip
- mRNA from one source is hybridized to each chip (as opposed to two-color systems)
- Currently:
  - up to 1.3×10^6 oligos corresponding to 47000 target sequences
  - minimum amount of 5μg mRNA



Source: Affymetrix

Analysis of Affymetrix Data

Wet-lab $\Rightarrow$ Image processing $\Rightarrow$ Low-level analysis $\Rightarrow$ High-level analysis

Low-level analysis: Extract a measurement of relative mRNA abundance that is reasonably free of technical variation

High-level analysis: Relate the relative mRNA abundance to the biology of interest

The high-level analysis of Affymetrix data is **identical** to the analysis of two-color microarray data collected in a reference design!

Low-level Analysis

Extracting Expression Values

Several processing steps are required to get an expression value that is **representative** of relative mRNA abundance and **comparable** between chips:

- Background correction
- Normalization
- Signal adjustment
- Signal extraction
- Normalization

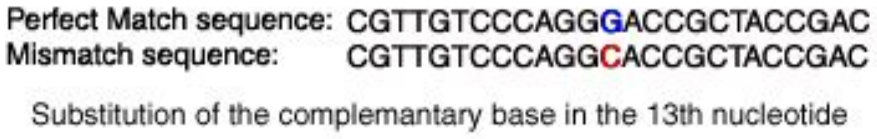
Unfortunately, there is no agreement how to do this correctly.

Signal Extraction & Expression Measures – The Problem

Each 'gene' (target sequence) is covered by 11-20 **probe pairs** consisting of

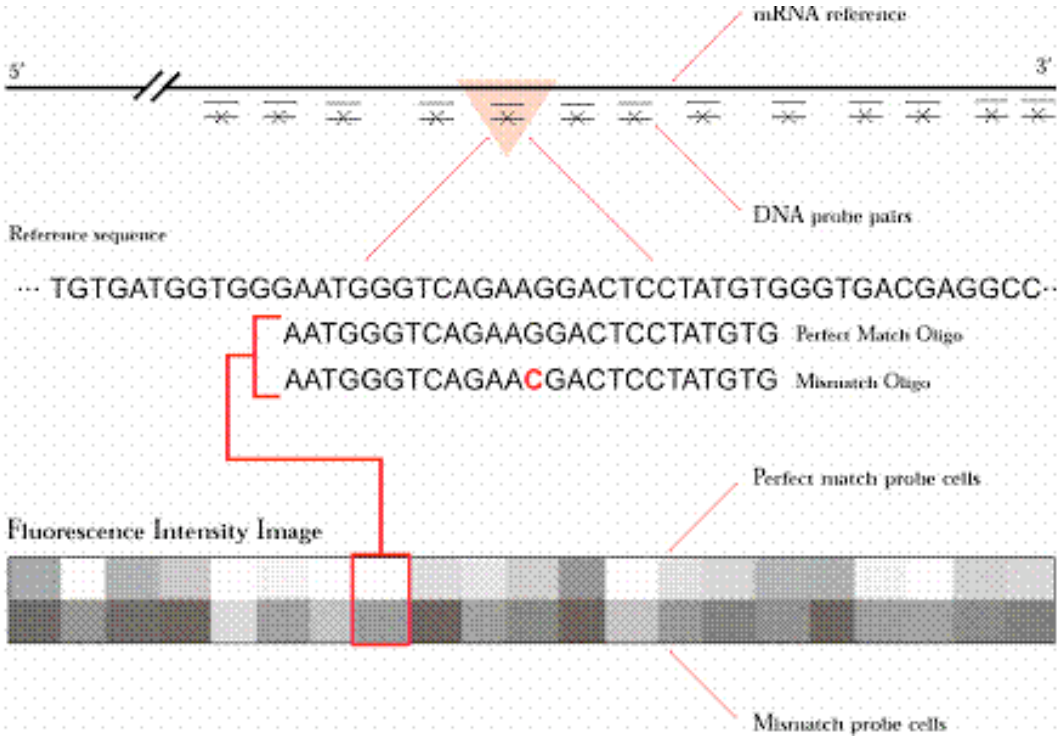
- perfect match (PM): exact substring of the target sequence,
- mismatch (MM): the PM with one central base altered.

. . . like this:



Source: Affymetrix

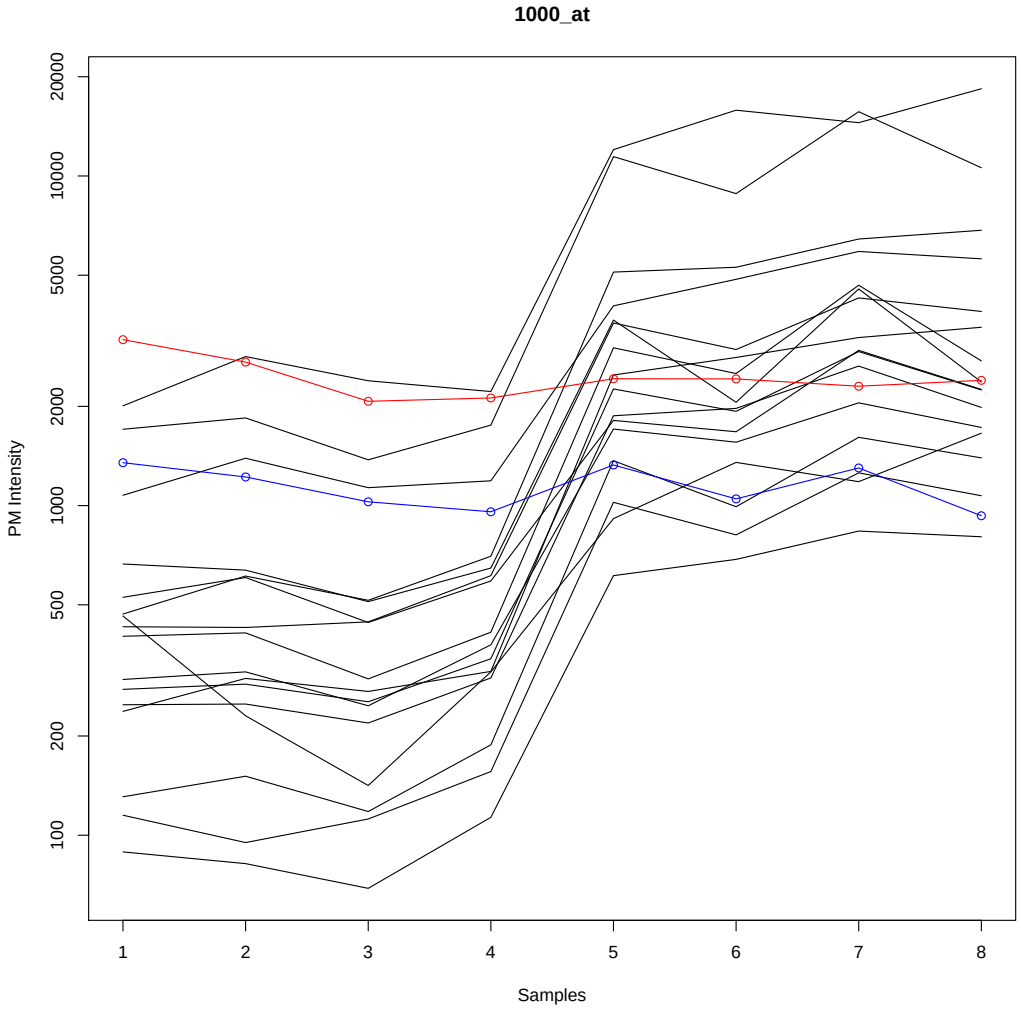
Each target sequence is represented by a **probe set** of 22-40 measurements on each chip!



Source: [15]

An expression measure combines the multiple PM/MM measurements into one number representative of relative mRNA abundance of the target sequence on each chip

Example: Cell Line Data (8 chips)



Expression Measures – Methods

Single chip methods compute expression values separately for each chip

Multichip methods use the full set of available chips to compute expression values

Probe pair data is is highly variable:

- inherent noise of mRNA data,
- inherently different base intensities of individual probes, mostly due to different C/G content

$\Rightarrow$ some **robust averaging** across probe pairs is usually applied

MAS5: single chip method, current standard suggested by Affymetrix. For each probe set on each chip,

- compute an adjusted difference PM-MM for each probe pair (guaranteed to be positive),
- compute a robust average ($\approx$ median) across the probe pairs.

Values are usually logarithmized or otherwise transformed [10]

RMA: Robust Multichip Average [11], for each probe set

- use all arrays $i = 1, \ldots n$ and all probe pairs $j = 1, \ldots k$,
- compute $y_{ij} =$ background-adjusted $\log 2$ of the PM intensities,
- robustly fit the model

$$y_{ij} = a_i + p_j + \epsilon_{ij}$$

- the fitted a_i are the expression values

RMA is originally logged, but can be further transformed.

MBEI: Model-Based Expression Index [14], implemented in dChip. For each probe set

- use all arrays $i = 1, \ldots n$ and all probe pairs $j = 1, \ldots k$,
- compute either $y_{ij} = \text{PM}_{ij}$ or $y_{ij} = \text{PM}_{ij} - \text{MM}_{ij}$,
- robustly fit the model

$$y_{ij} = a_i p_j + \epsilon_{ij}$$

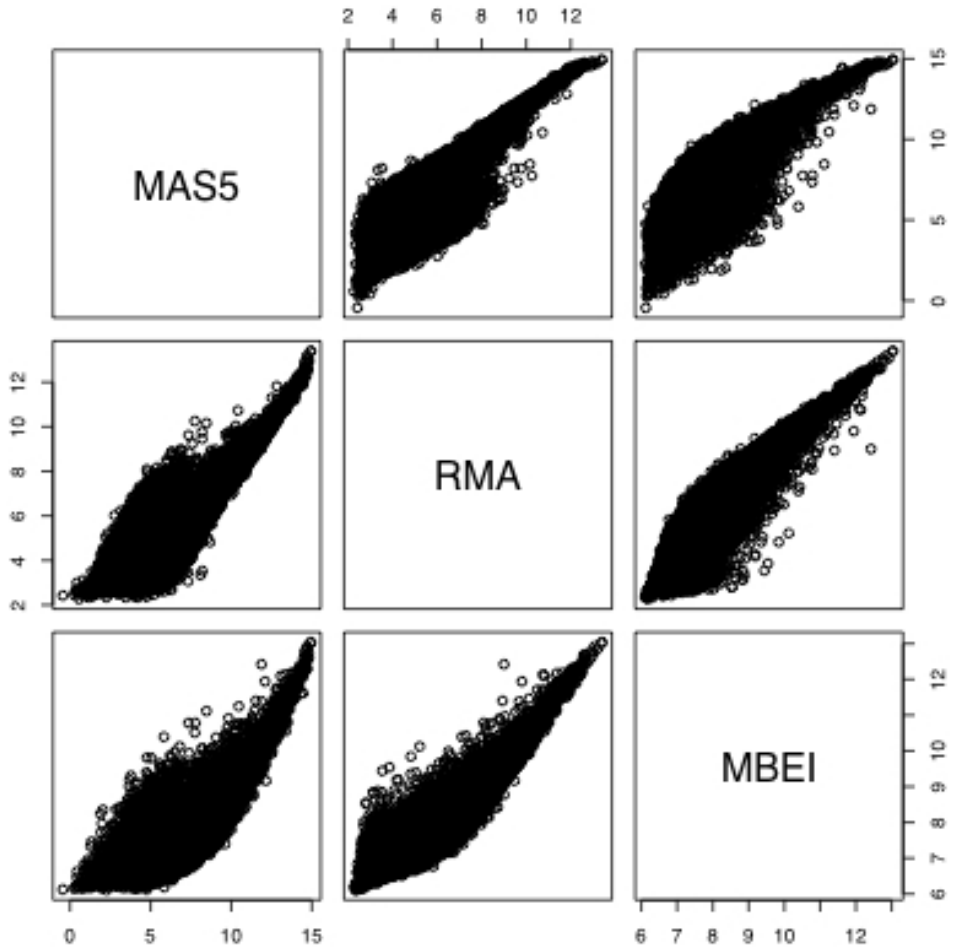
- the fitted a_i are the expression values

MBEI is commonly logged. It requires 10-20 chips for realistic results.

Alternatives: e.g.

- gcRMA: and extension of RMA that incorporates the different G/C ratio of the the individual probes in a probe set
- PDNN: positional dependent nearest neighbor, tries to estimate the effect of non-specific binding [22]
- Skip expression measure if you are interested only in differential expression/gene lists [13]

Expression measures: 1 chip, all probe-sets



Normalization

Both the overall intensity level and the distribution of intensity values vary between chips due to technical reasons:

- mRNA amount
- sample processing
- scanner calibration

In order to be able to compare expression values between chips gainfully, the effect of this technical variability should be reduced or eliminated: **Normalization**.

This requires the assumption that some aspect of the distribution of intensity values remains constant between chips; different normalization procedures make different assumptions.

Normalization can be done

- **before** summarizing the the probes through an expression measure: probe-level normalization
- **after** summarizing the the probes through an expression measure: probeset-level normalization

Usually a normalization procedure is associated with a specific expression measure, but in principle all combinations are possible (if not always useful). [3, 4]

Global mean normalization: assumes constant average intensity on all chips (ie little or balanced differential expression). Usually on probe-set level, usually with MAS5

Quantile normalization: assumes constant: distribution (ie histogram) of intensities on all chips (ie very little differential expression). Usually on probe level, usually with RMA

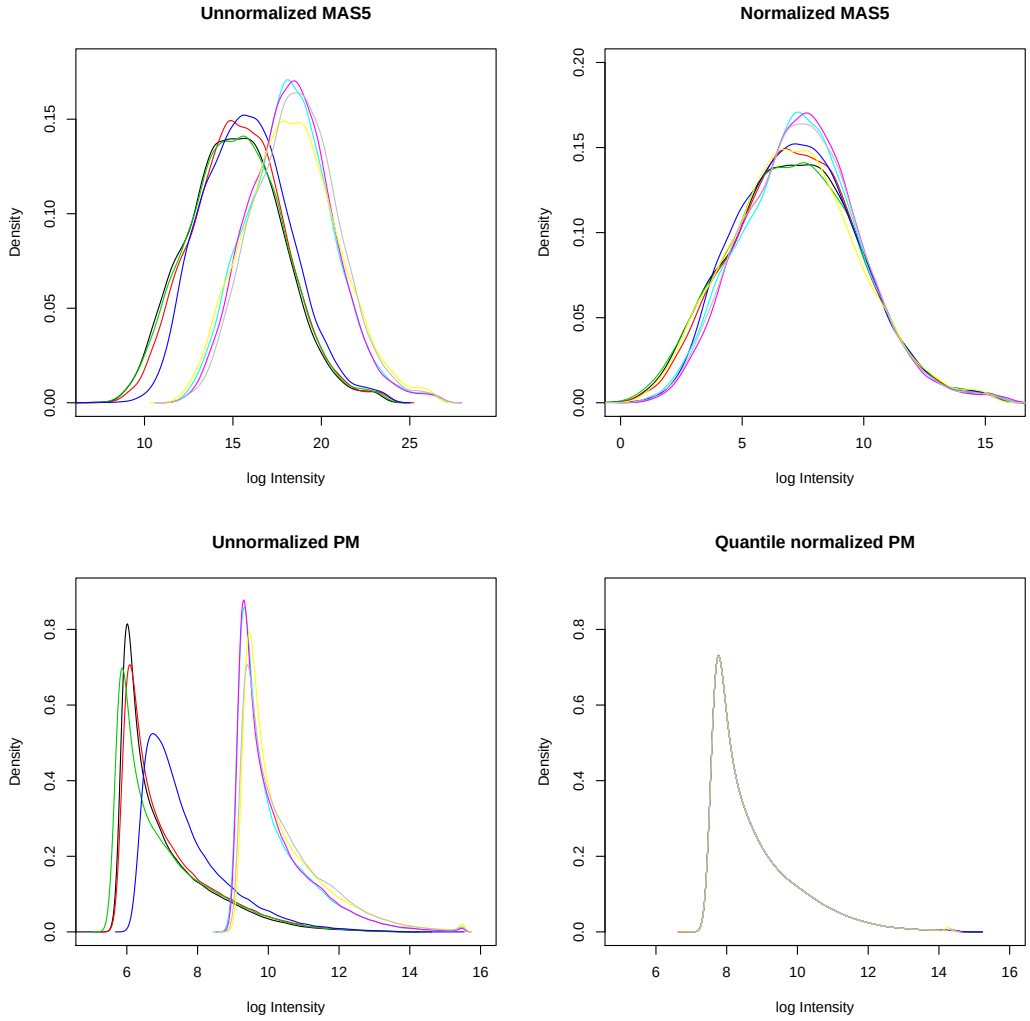
Invariant set normalization: assumes constant intensities on a subset of genes

- identify a set of genes with small change in rank between chips,
- iteratively reduce this set until convergence,
- fit a smoothing curve to these genes between each chip and an artificial reference chip (median chip),
- use the curve to normalize all genes.

Usually on probe level, usually with MBEI

Alternative: use pairwise normalization scheme like for two-dye systems (loess-normalization), either iteratively or to a artificial reference chip (mean/median chip). Usually on the probe-set level – can be combined with most expression measures [4]

Normalization: Cell lines



Does the choice of low-level analysis matter?

Yes! [9, 2, 3, 4] Choice of summary measure and normalization has serious impact on all subsequent high-level analysis.

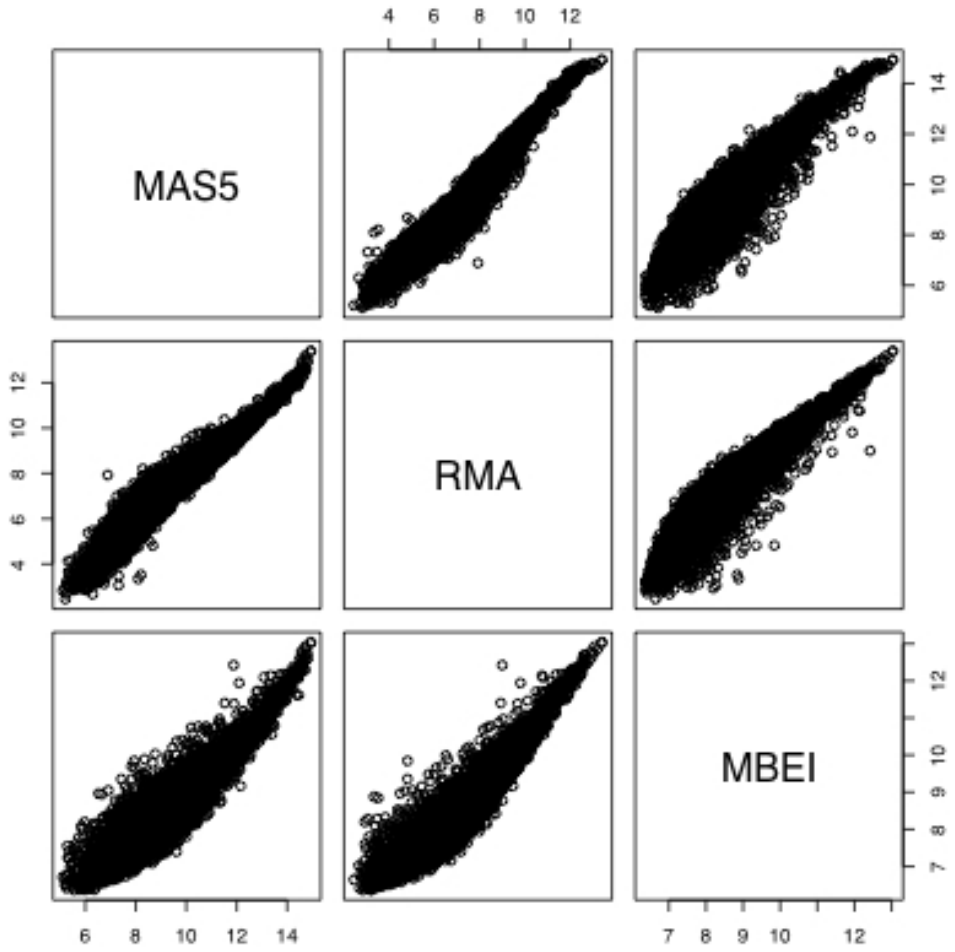
Evaluation & comparison of different low-level approaches are generally based on artificial reference data (spike-in, dilution), trying to establish general superiority of a strategy. Results so far suggest that no single approach is uniquely superior for all possible data sets (e.g. clinical vs. experimental data).

An alternative is to assess the quality of the low-level analysis for a specific data set, based on studying correlation of random pairs of genes [17].

Quality Control

- Graphical:
  - Digestion plots
  - Chip images and residual plots
  - MA-plots only of limited use
    - too many pairs of chips
    - average MA-plots (between groups) with little QC information
- Numerical: Affymetrix presence/absence calls
  - based on p-value for Wilcoxon test comparing PM and MM
  - can be Present, Absent, Middling
  - not optimal, but helpful

Expression measures: 1 chip, only present calls



High-level Analysis

The most common ways of analysing expression data can be summarized as follows:

Class discovery: find subclasses of samples based on the expression pattern of a large number of probesets: e.g. molecular subclasses of cancer

Class prediction: given a biological or clinical grouping of the samples, find a small number of genes whose expression pattern can be used to reliably predict the group membership of new samples: e.g. treatment response in patients

Gene discovery: given a biological or clinical grouping of the samples, identify the genes that are differentially expressed between these groups and estimate their fold changes: e.g. cell line response to treatment over time

Roughly speaking, CD and CP are generally used more for clinical and observational studies than for biological and experimental studies. GD is suitable for both situations, but most successful for experimental studies.

Methods for analysing expression data can be classified in several ways:

- statistical vs. machine learning
- exploratory vs. inferential
- supervised vs. unsupervised learning

The first distinction is somewhat cultural; the second distinction is used more in statistics, the third more in machine learning.

All this is very similar or identical between oligonucleotide/one-dye systems and many cDNA/two-dye systems (i.e. those with a common reference)

Class Discovery

Unsupervised learning or exploratory analysis - we try to identify groups of samples with expression patterns that are similar within groups and dissimilar between groups.

This begs the question of what we mean with similarity. This is an important question for any kind of class discovery: What kind of distance measure do we use?

A distance measure defines for two samples s_i and s_j how similar their expression pattern is:

$$d_{ij} = \text{dist}(s_i, s_j) = f((x_{1i}, \ldots x_{gi}), (x_{1j}, \ldots x_{gj}))$$

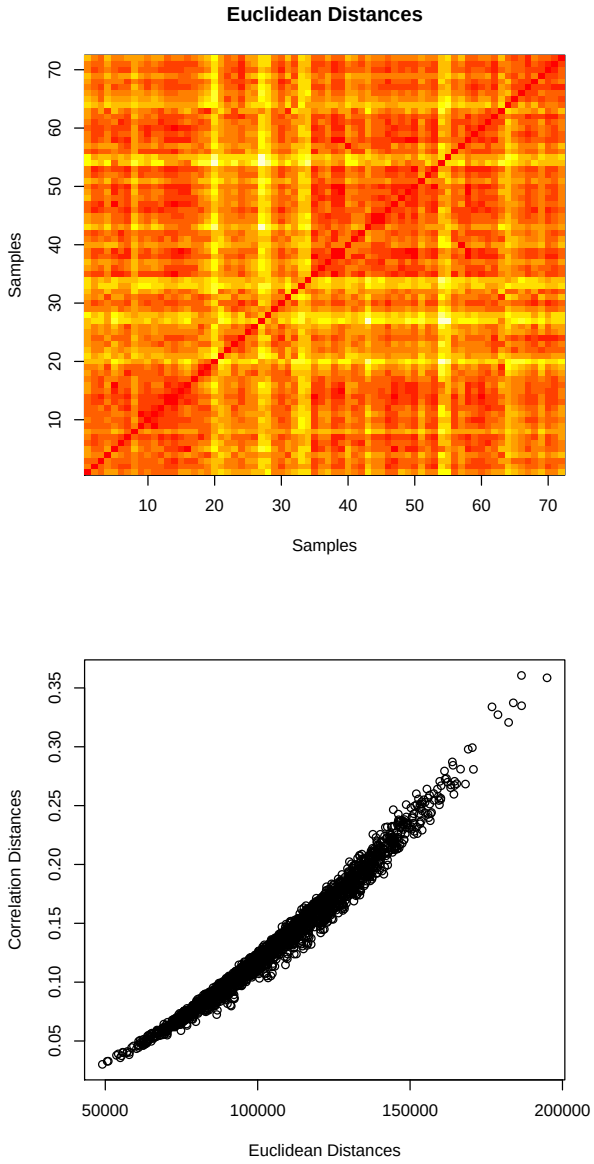
Different distance measures are in use [18], most commonly e.g.:

- Euclidean distance: $d_{ij} = \sqrt{\sum_{k=1}^{g}(x_{ki} - x_{kj})^2}$
- Correlation disctance: $d_{ij} = 1 - corr((x_{1i}, \ldots x_{gi}), (x_{1j}, \ldots x_{gj}))$

Note that the Euclidean distance uses the actual level of expression, whereas the correlation distance only uses the pattern of expression across the genes!

Example: 72 lymphomas with 7129 genes [8]

- red = close, white = distant
- distance matrix is symmetric with zero diagonal
- some lymphomas seems to have a large distance from everybody else
- rather good agreement between Euclidean and correlation distance, but the small distances are proportionally smaller for correlation distances


Euclidean Distances
Samples
10
20
30
40
50
60
70
Samples
Correlation Distances
Samples
10
20
30
40
50
60
70
Samples
Correlation Distances
0.05
0.10
0.15
0.20
0.25
0.30
0.35
50000
100000
150000
200000
Euclidean Distances

Some Methods for Class Discovery

Hierarchical clustering: based on the idea of successively combining the closest observations into clusters. This results in a hierarchy of possible groupings, from all observations in separate singleton clusters to one cluster with all observations. Ideally, we will find an obvious solution when looking at the corresponding dendrogram.
There are different variants of hierarchical clustering, depending on how to calculate the distance between non-singleton clusters. (e.g. [18, 5])

Non-hierarchical clustering: Many different procedures. These usually require the specification of the required number of clusters beforehand, or alternatively several runs of the same method with different numbers of clusters specified. [5]

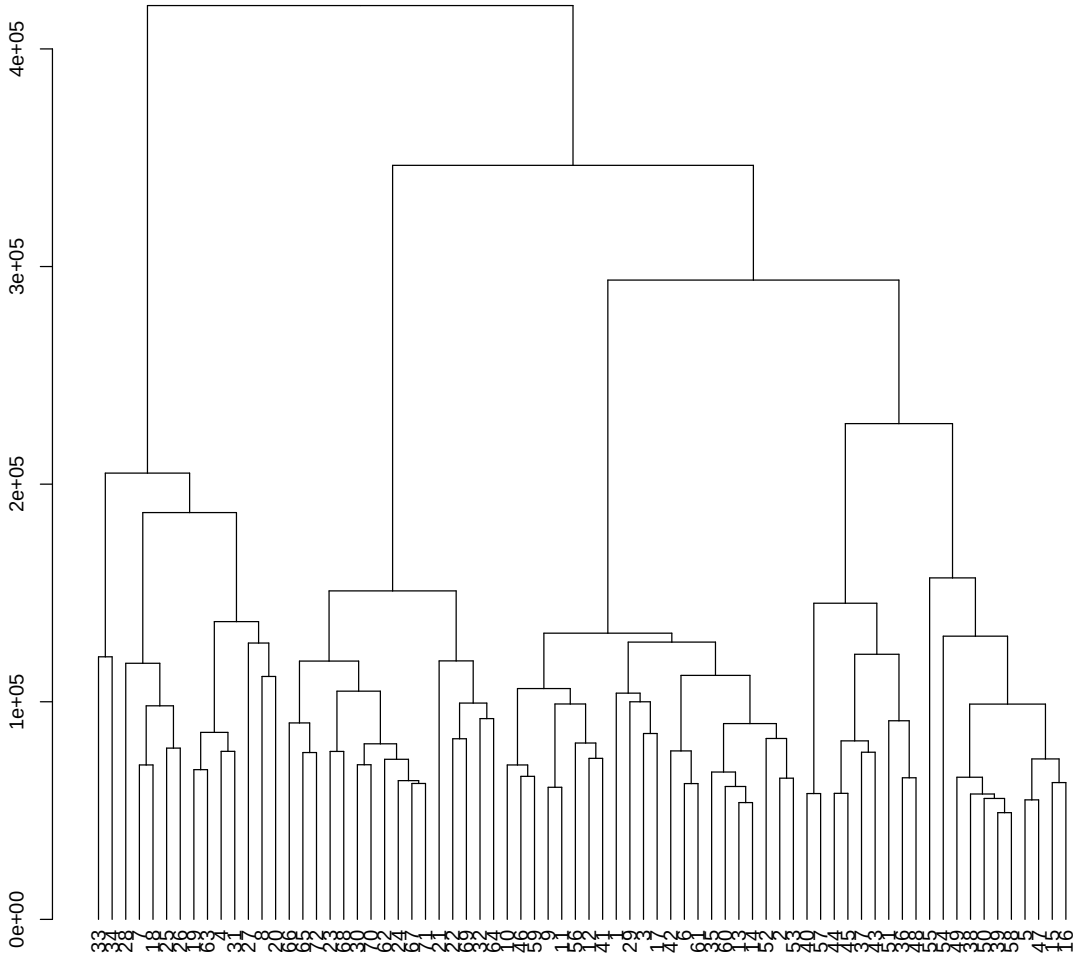
Multidimensional scaling (MDS): tries to build a set of points in 2D or 3D space where a) each point corresponds to one sample and b) the geometrical distances

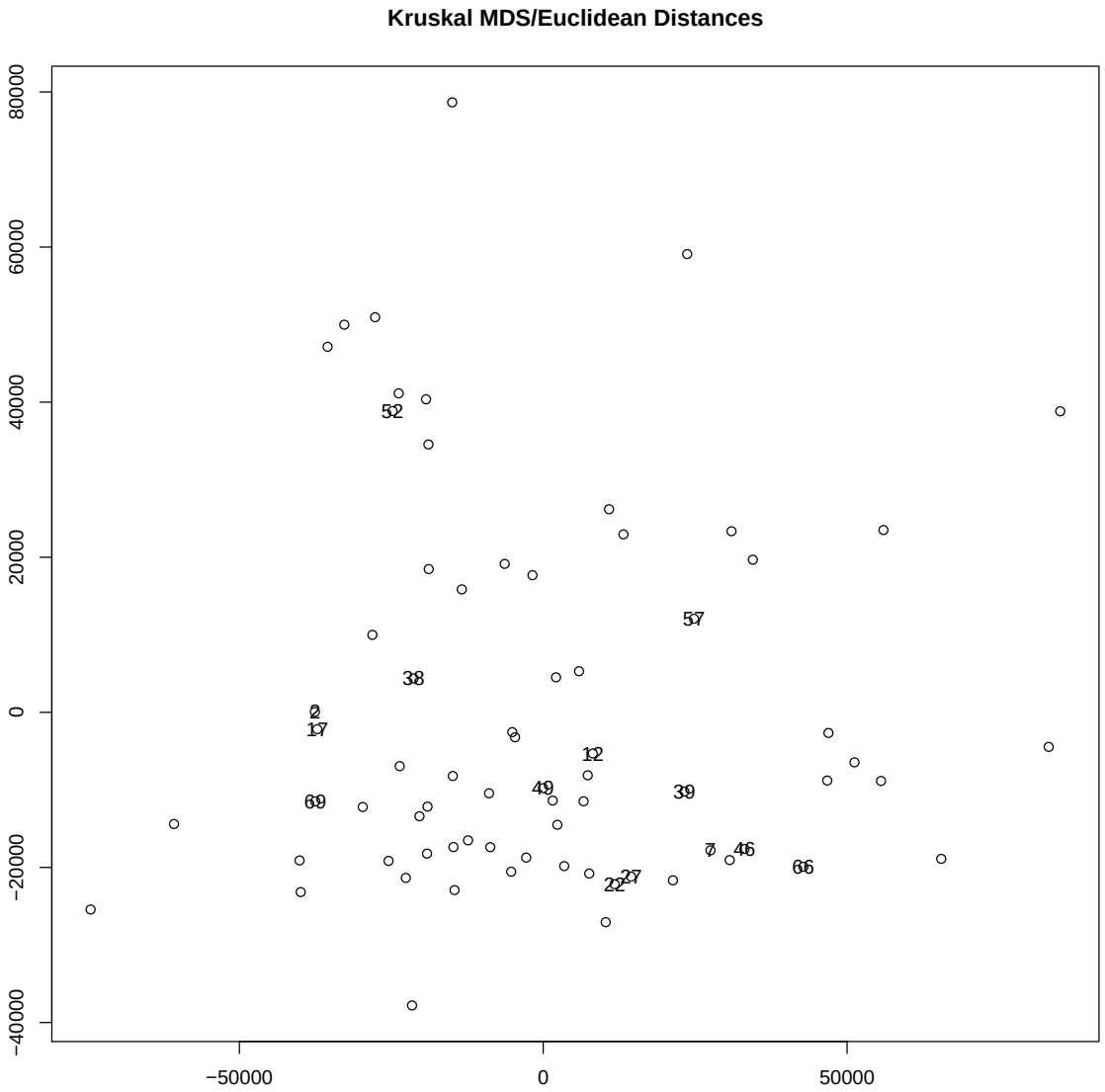
between points are as close as possible to the actual distances between samples as measured by the chosen metric.
Different MDS methods exist. Usually some robust variant is preferable.

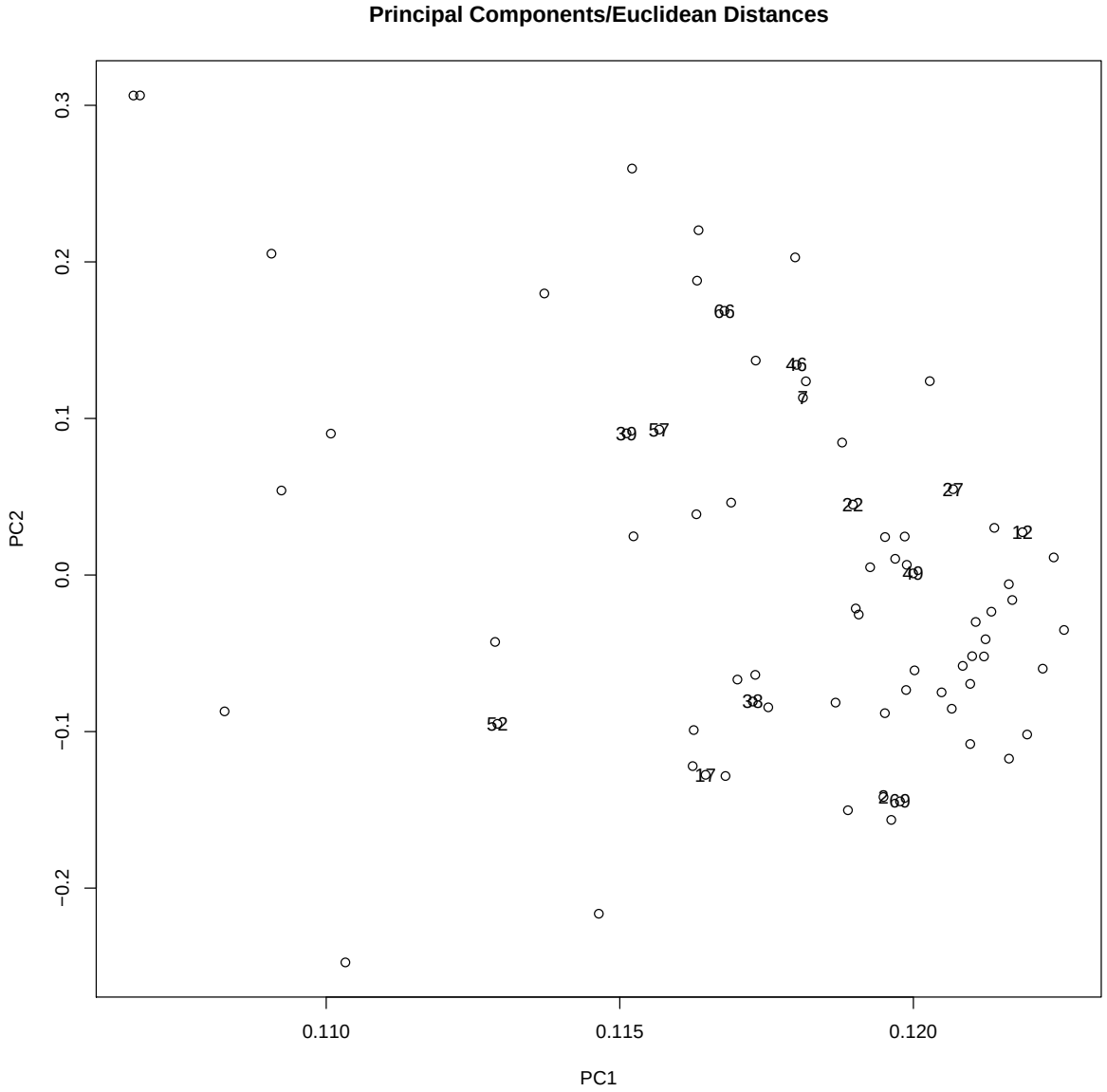
Principal component analysis (PCA) is an exploratory statistical technique with a wide range of applications. Briefly, it constructs a set of new variables (as many as samples in the set) that are weighted sums of the original vectors of gene expression. These new variables are uncorrelated and explain successively as much of the variability in the data as possible (i.e. the first variable explains most, the second less and so on). If we limit ourselves to specific variants of the Euclidean and correlation distance, the PCA can be used in a similar manner as the MDS, by looking at scatter plots of the first two or three variables (i.e. those with highest information content) and trying to identify groupings visually.

Example: Lymphoma

- dendrogram shows three to four clear clusters
- MDS and PCA show similar picture, though rotated; separation less clear


Ward's Method/Euclidean Distances
4e+05
3e+05
2e+05
1e+05
0e+00


Kruskal MDS/Euclidean Distances
80000
60000
40000
20000
0
−20000
−40000
−50000
0
50000


Principal Components/Euclidean Distances
PC2
PC1
0.3
0.2
0.1
0.0
−0.1
−0.2
0.110
0.115
0.120

Class Validation

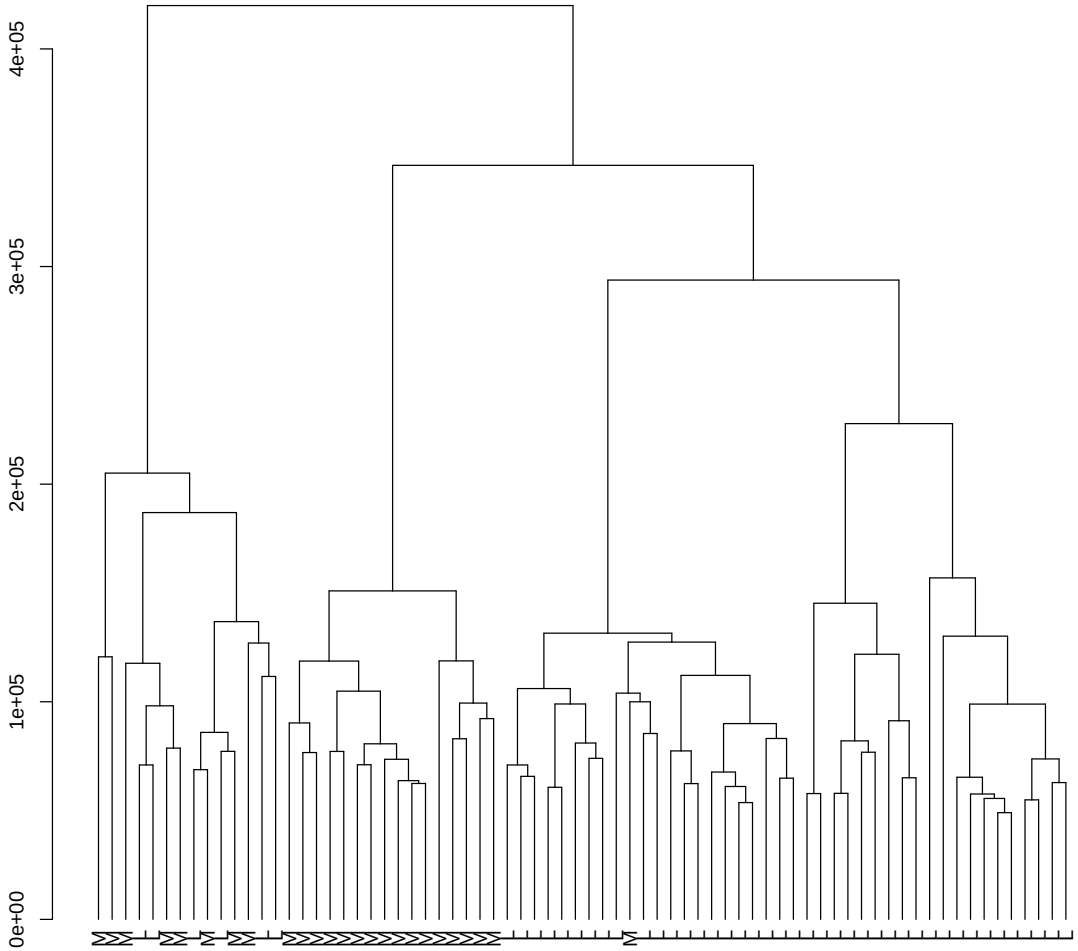
Once reasonably well separated classes have been established, they need to be characterized, confirmed and validated. This can and should be done in two ways:

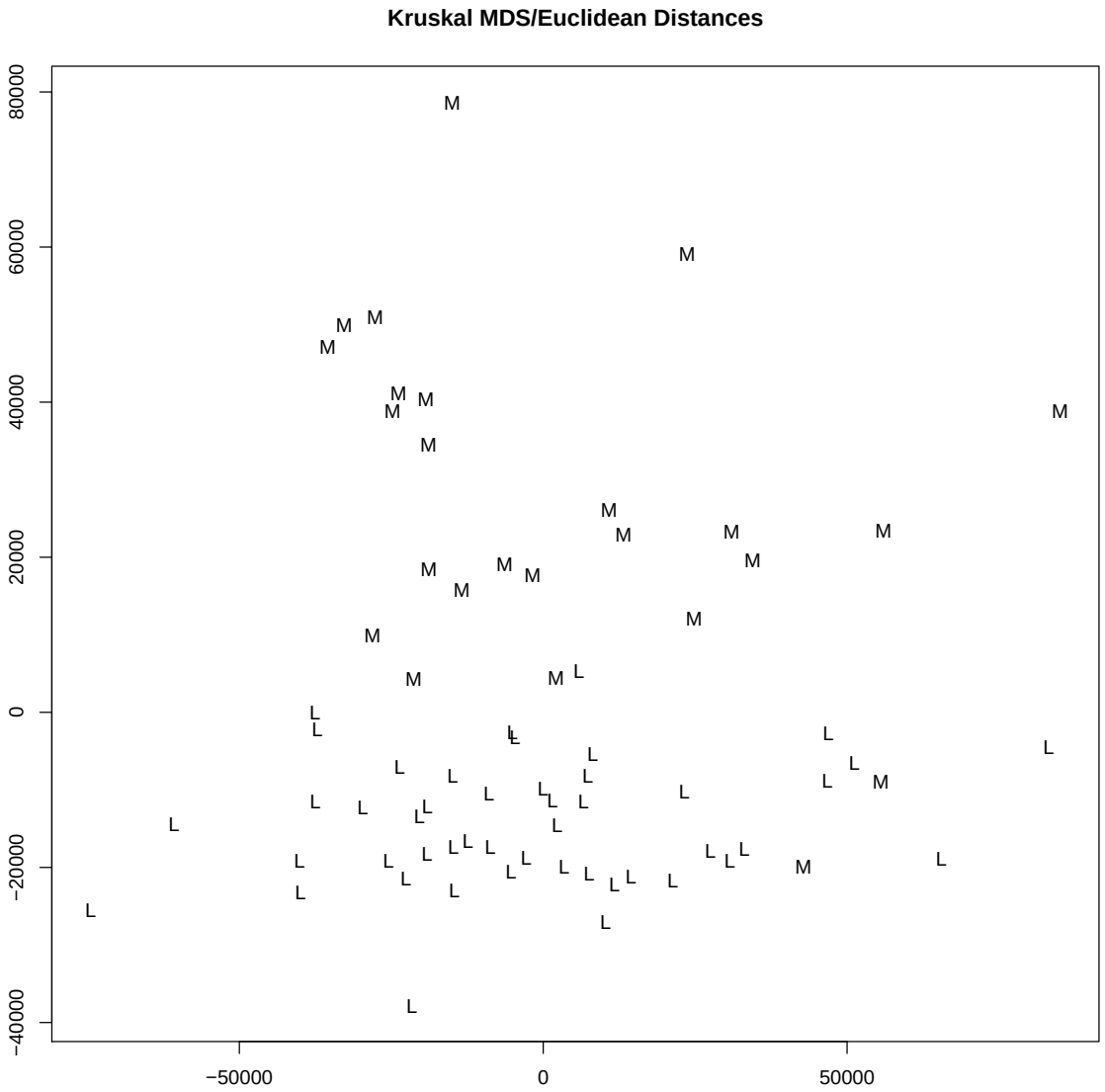
- by relating the newfound classes to already established clinical or biological properties of the samples: e.g. treated vs. untreated

- by applying the same procedure to an independent data set

Example: Lymphoma classification as ALL or AML

- dendrogram has one ALL, one AML, one mixed cluster; ALLs might be further subdivided.
- MDS separates ALL and AML


Ward's Method/Euclidean Distances
4e+05
3e+05
2e+05
1e+05
0e+00


Kruskal MDS/Euclidean Distances
80000
60000
40000
20000
0
−20000
−40000
−50000
0
50000

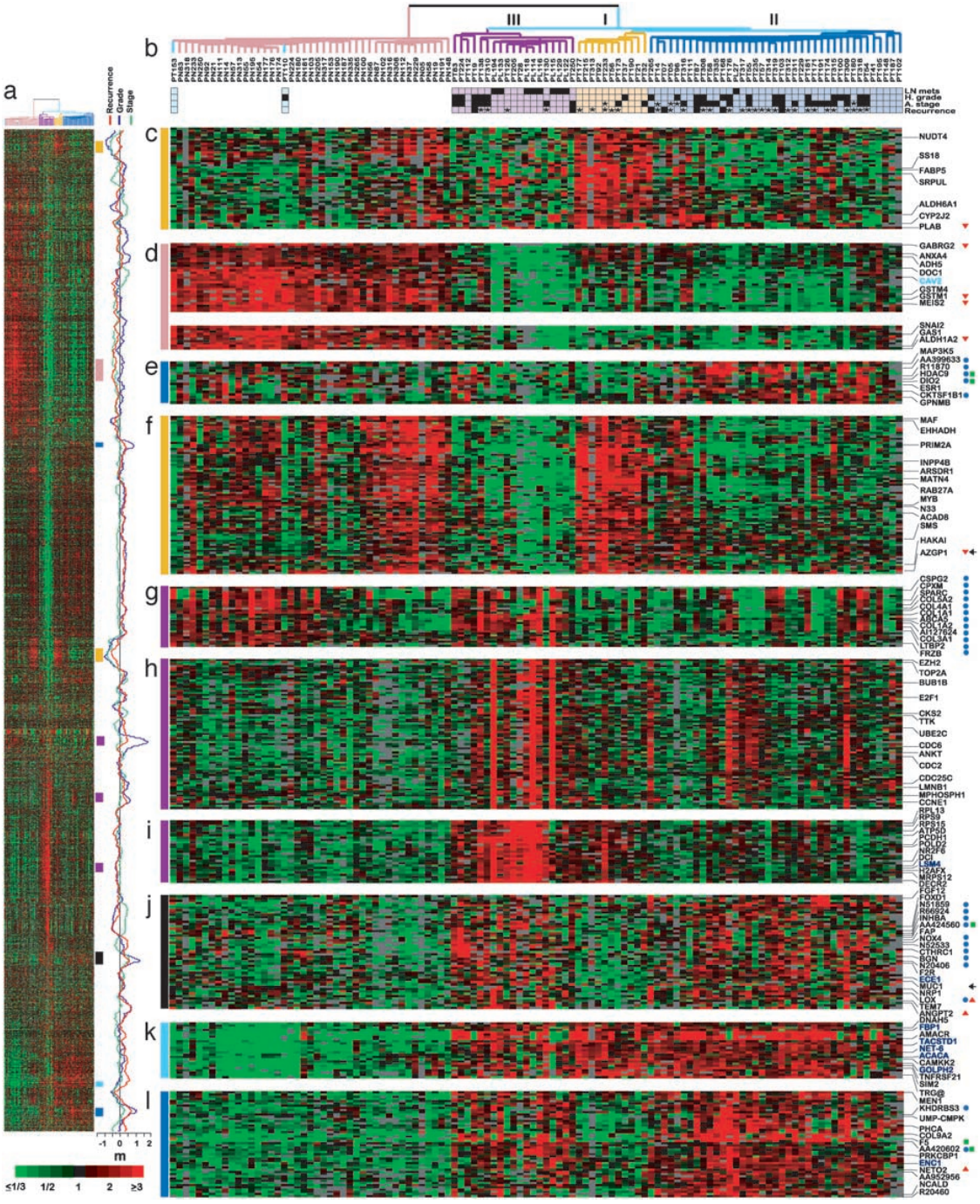
Other Applications

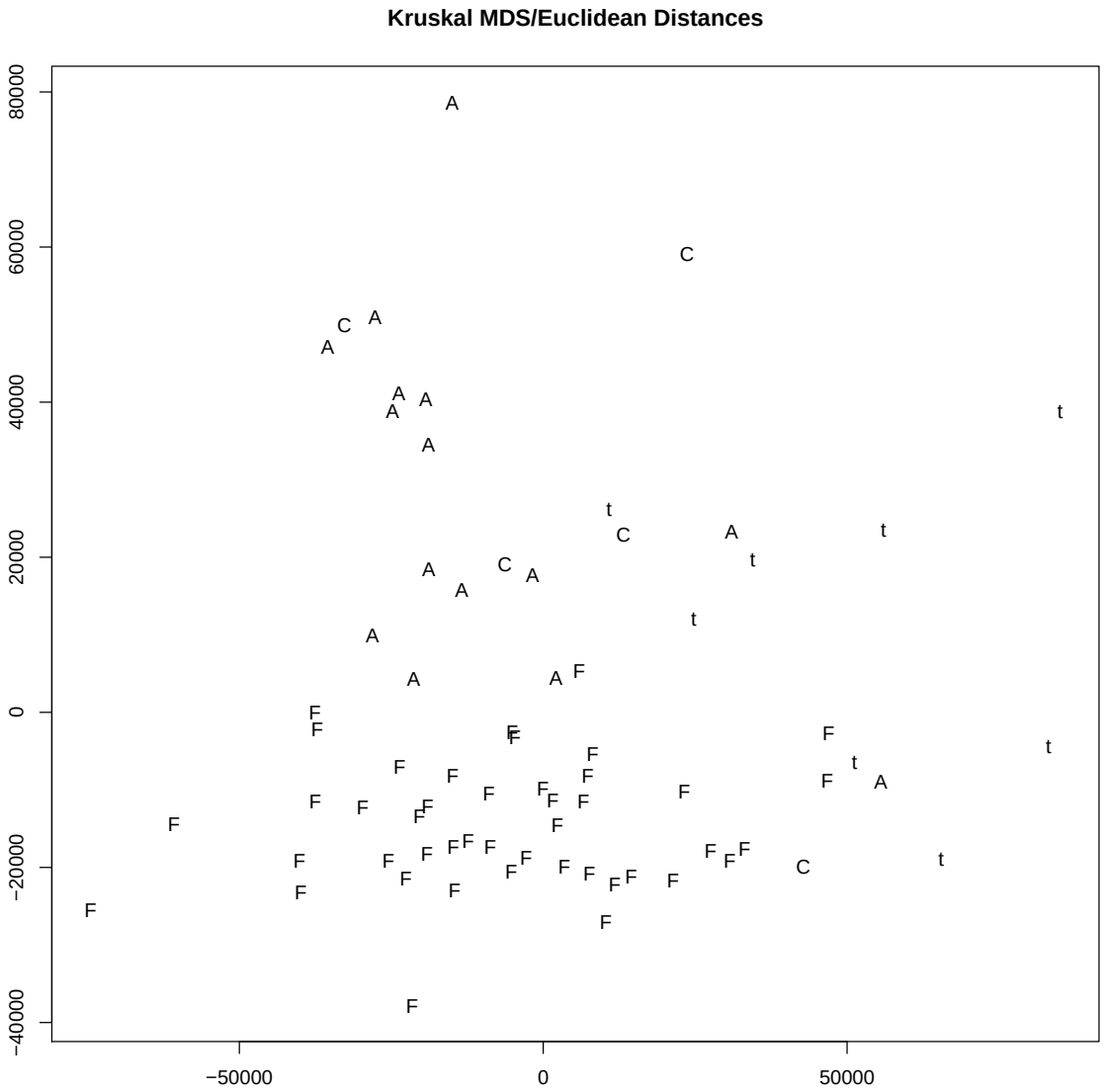
In exactly the same manner as the samples, we can also try to **identify groups of genes** with similar expression pattern across samples. This is mostly done via clustering, as the more graphical methods tend to break down with too many genes, but it still requires the specification of a distance measure and a clustering algorithm. This are usually taken to be the same as for the samples, but that is not required.
Often the sample and gene clustering are displayed together with the expression values in a heatmap.

Class discovery can usefully be employed for simple **quality control** by studying whether a grouping in the data corresponds to potential confounders like array batches, processing dates, operators, source of material etc.

Example: Prostate cancer [12] shows a clustered and annotated heatmap for cDNA arrays

Example: Lymphoma – samples are labeled with codes showing the sample source. This is obviously confounded with AML/ALL – design?!


a
b
c
d
e
f
g
h
i
j
k
l
III
I
II
Recurrence
Grade
Stage
LN mets
H. grade
A. stage
Recurrence
NUDT4
SS18
FABP5
SRPUL
ALDH6A1
CYP2J2
PLAB
GABRG2
ANXA4
ADH5
DOC1
CAV2
GSTM4
GSTM1
MEIS2
SNAI2
GAS1
ALDH1A2
MAP3K5
AA399633
R11870
HDAC9
DIO2
ESR1
CKTSF1B1
GPNMB
MAF
EHHADH
PRIM2A
INPP4B
ARSDR1
MATN4
RAB27A
MYB
N33
ACAD8
SMS
HAKAI
AZGP1
CSPG2
CPXM
SPARC
COL3A2
COL4A1
COL1A1
LUCAN
COL1A2
AI127624
COL3A1
LTBP2
FRZB
EZH2
TOP2A
BUB1B
E2F1
CKS2
TTK
UBE2C
CDC6
ANKT
CDC2
CDC25C
LMNB1
MPHOSPH1
CCNE1
RPL13
RPS9
RPS15
ATP5D
PCDH1
POLD2
NR2F6
DCI
LIM4
H2AFX
MEC912
PECR2
FGF12
FOXD1
N51859
R66924
INHBA
AA424560
FAS
NOX4
N52533
CTHRC1
BGN
N20406
F2R
ECE1
WIG1
NRP1
LOX
TEM7
ANGPT2
DNAH5
FBP1
AMACR
TACSTD1
NET-6
ACACA
CAMKK2
GOLPH2
TNFRSF21
SIM2
TRGB
MEN1
KHDRBS3
UMP-CMPK
PHCA
COL9A2
F5
AA420602
PRKCBP1
ENC1
NETO2
AA952956
NCALD
R20480
m
-1
0
1
2
≤1/3
1/2
1
2
≥3

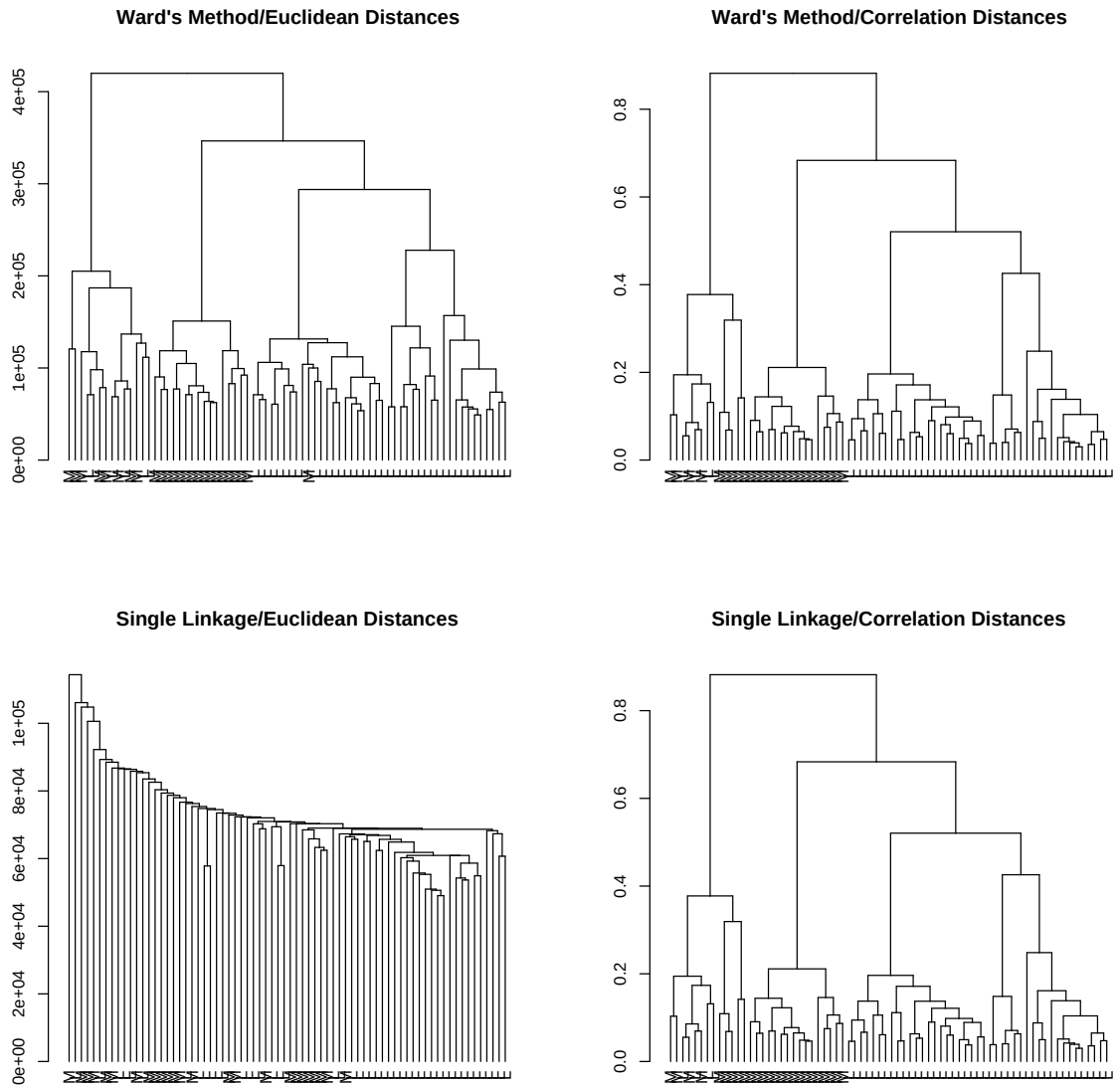

Kruskal MDS/Euclidean Distances
80000
60000
40000
20000
0
−20000
−40000
−50000
0
50000

Caveats

Results depend sometimes to a large degree on the choice of method and distance measure.

Results also depend crucially on the list of genes employed: quality or signal filtering of genes is ok, but selecting genes based on their association with a property of interest is an incredibly bad idea.

Example: Lymphoma – four different groupings, depending on distance/method.


Ward's Method/Euclidean Distances
0e+00
1e+05
2e+05
3e+05
4e+05
Ward's Method/Correlation Distances
0.0
0.2
0.4
0.6
0.8
Single Linkage/Euclidean Distances
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Single Linkage/Correlation Distances
0.0
0.2
0.4
0.6
0.8

Class Prediction

This is supervised learning – we are interested in one specific property of the samples that separates them into two or more classes: e.g. treatment response, tumor grade, distant metastasis.

The goal is to identify a rather small set of genes that allows a correct classification of a new sample. A reasonable biological interpretation of these genes is nice, but secondary. In a way, this approach tries to by-pass proper biological understanding of expression signatures in favor of immediate (usually clinical) black-box prediction.

Methods for class prediction come in two flavors:

- classical statistical procedures or variations thereof, like discriminant analysis or logistic regression,
- machine learning methods that are based on data mining procedures like nearest neighbors or support vector machines.

Misclassification Rate

The misclassification rate for new samples of any method can be estimated in different ways:

- By looking at the misclassification rate in the original data set that was used to fit the model or train the procedure. This is traditional in statistics when we have many more samples than variables (e.g. R^2), though it's tricky even there. In case of expression data with many more variables (genes) than samples this is useless and dangerous as it underestimates the true misclassification rate severely.

- By using a leave-one-out crossvalidated misclassification rate: each sample is in turn removed from the data set and the model is fit to/the procedure is trained on the $n-1$ left over samples; the model is used to predict the class of the removed sample. By going through the whole sample and comparing

the leave-one-out prediction and the actual class, we get a much more realistic estimate of the true misclassification rate.

- This approach can be extended to leaving out k samples at a time and predicting them with the rest of the data.

- In extremis, this leads to the trainings/test data approach: the data is split randomly in a trainings data set, where the model is fit, and a test data set, where the model is applied for prediction. The misclassification rate on the test data is a realistic estimate of the true misclassification rate, but using too little data in the trainings set can reduce the power of the model to predict correctly.

Note that even this last approach is often not the gold standard of independent replication: any systematic error in collecting, preparing or processing the samples will be shared by trainings and test set.

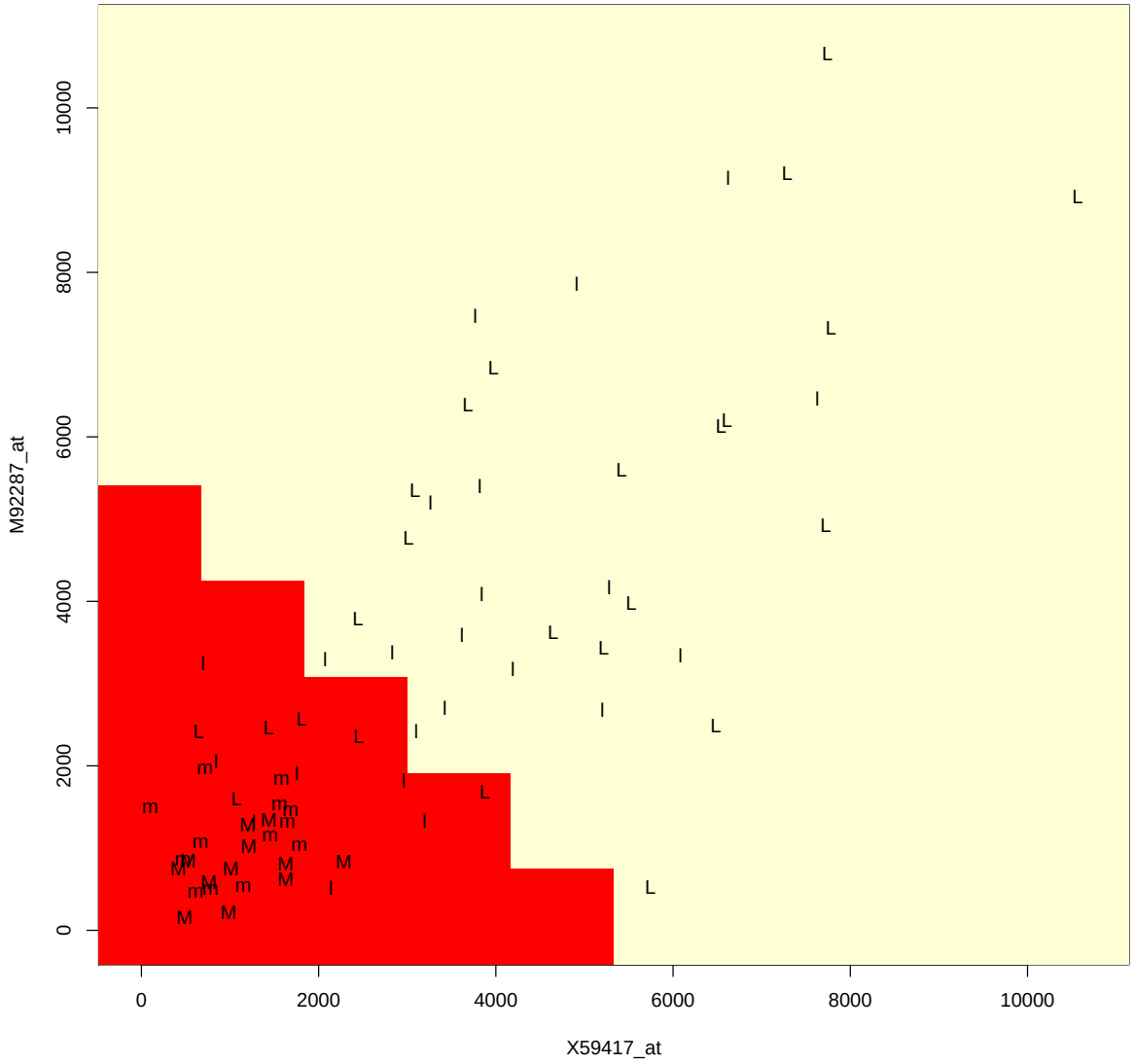
Example: Diagonal Linear Discriminant Analysis

This procedure is a simplification of a classical statistical procedure (LDA) that disregards correlations between genes.Basically, we assign each new sample to the group with the closest group center, where the group center is just the mean expression signature over all genes, and close is according to a weighted Euclidean distance:

$$C(x_{1r}, \ldots x_{gr}) = \text{argmin}_k \sum_{i=1}^{g} \frac{(x_{ir} - \bar{x}_{ik})^2}{s_k^2}$$

Example: Lymphoma, using two genes, predicting AML/ALL

- M/L = trainings sets, m/l = test set
- red = predicted AML, yellow = predicted ALL
- Misclassification rate: trainings set: 17%, test set: 25%


M92287_at
X59417_at
0
2000
4000
6000
8000
10000

Example: k-nearest neighbor classification

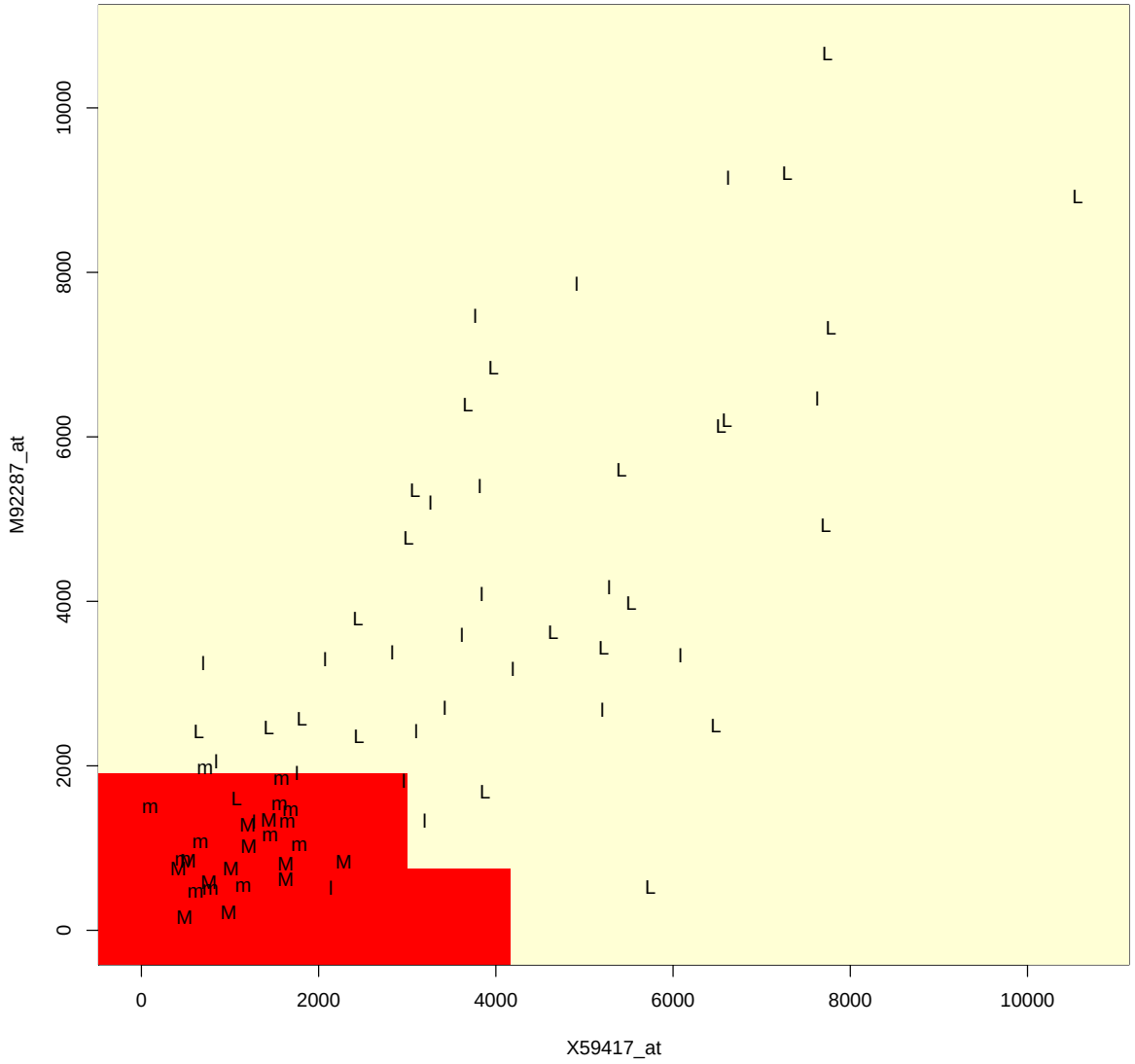
This is a simple machine learning procedure that classifies new samples according to the majority vote among the k nearest neighbors.

This approach has two parameters that determine its properties: the choice of the distance measure and the size of the local neighbourhood. The latter is often chosen via leave-one-out cross-validation.

Note that for $k = n$, every new sample is assigned to the most frequent class.

Example: Lymphoma, using two genes, predicting AML/ALL

- Misclassification rate using $k = 1$: trainings set: 0%, test set: 8%


M92287_at
X59417_at
0
2000
4000
6000
8000
10000

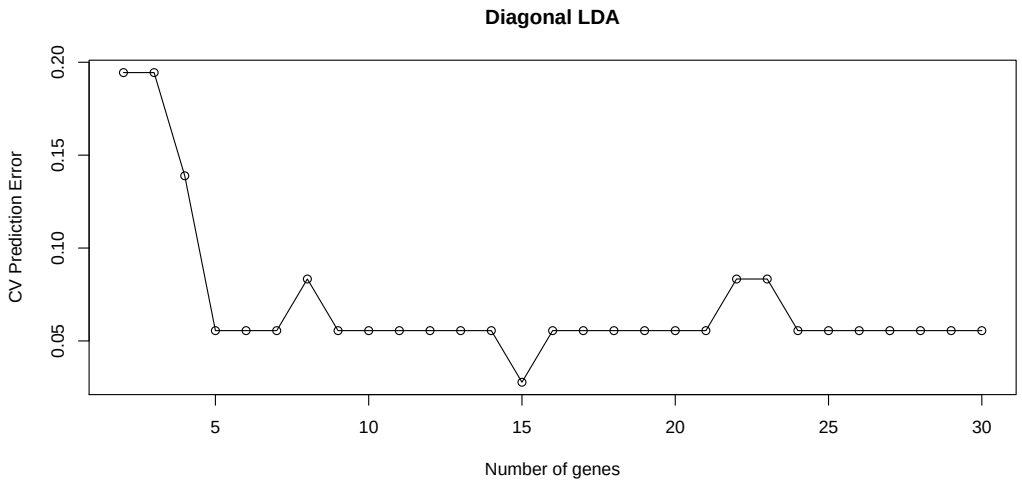
Finding a Small Set of Predictors

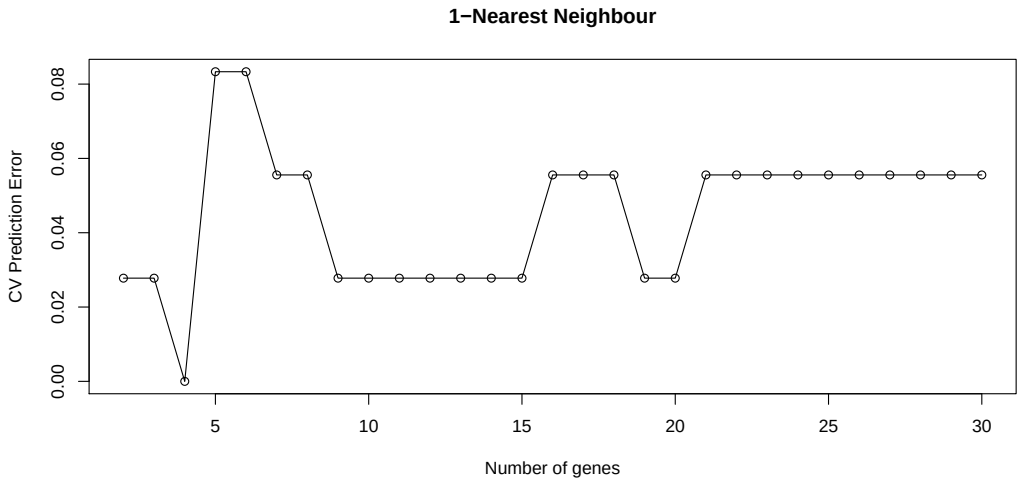
Note that it is often impractical to use the full gene set to predict a classification - this would imply that the full microarray would have to be run on every new sample for classification. The goal is rather to identify a small subset of genes that do the same work in comparable quality. These can then be fit on a custom chip or measured in a different manner.

For this purpose, genes are often ranked by the strength of the association with the classification, e.g. the absolute value of their t-statistic. The procedure begins the prediction with one or two top genes and adds successively more genes while tracking the leave-one-out cross-validated misclassification error. The final model can then be applied to a test set.

Example: Lymphoma, predicting ALL/AML with no more than 30 genes

Name	Score
X59417_at	8.89
M92287_at	8.83
M31523_at	8.73
U05259_rna1_at	8.44
X95735_at	8.42
M84371_rna1_s_at	8.12
M11722_at	7.99
M31211_s_at	7.99
X17042_at	7.89
J05243_at	7.87
M89957_at	7.73
S50223_at	7.43
D88270_at	7.35
U29175_at	7.32
Y08612_at	7.30
⋮	





Comparison:

No. of genes	DLDA Cv	DLDA Test	KNN Cv	KNN Test
2	0.19	0.28	0.03	0.11
3	0.19	0.19	0.03	0.08
4	0.14	0.31	0.00	0.11
5	0.06	0.08	0.08	0.08
10	0.06	0.08	0.03	0.08
15	0.03	0.08	0.03	0.08
20	0.06	0.08	0.03	0.11
25	0.06	0.08	0.06	0.11
30	0.06	0.06	0.06	0.11

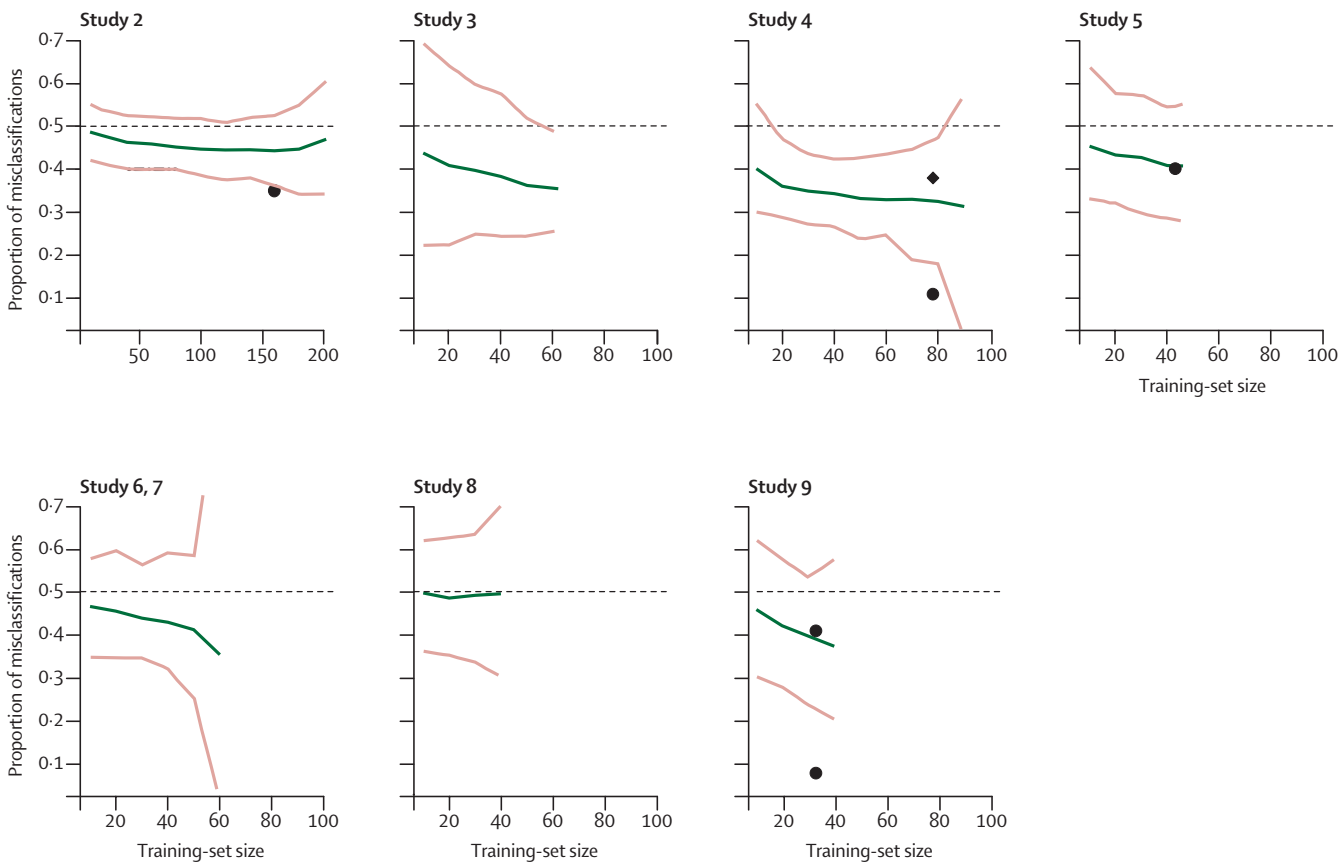
Note that the cross-validate error is still somewhat optimistic for KNN!

Caveats

A lot of people get it wrong and report unrealistic low misclassification rates, e.g.[20].

There is an element of randomness in the random split between trainings and test set that can bias the results. It is more realistic to repeat this process multiple times. [16]

Many different methods are available. Most of them are more or less reasonable, and it is rare that one outperforms the others completely.


Study 2
Study 3
Study 4
Study 5
Study 6, 7
Study 8
Study 9
Proportion of misclassifications
0·7
0·6
0·5
0·4
0·3
0·2
0·1
50
100
150
200
20
40
60
80
100
Training-set size

Source: [16]

Differential Expression

More cautious: identify genes that are differentially expressed (**DE**) and try to build some testable biology on that. This is a very reasonable use of microarrays – basically as a **screening tool** and for hypothesis generation. However, it often ends with users staring at too short or too long gene lists.

The main problem is multiplicity. Very roughly we can distinguish three approaches:

- classical tests with some kind of multiplicity correction [6]. These tests usually ignore the relationships between genes.
- modified tests that recover some of the information from the relationships between genes; these tests are often somewhat Bayesian and may or may not be built around a multiplicity correction scheme (e.g. [19]
- data reduction approaches, where we try to reduce the number of hypothesis that is tested simultanously. These may be data-driven (non-specific filtering)

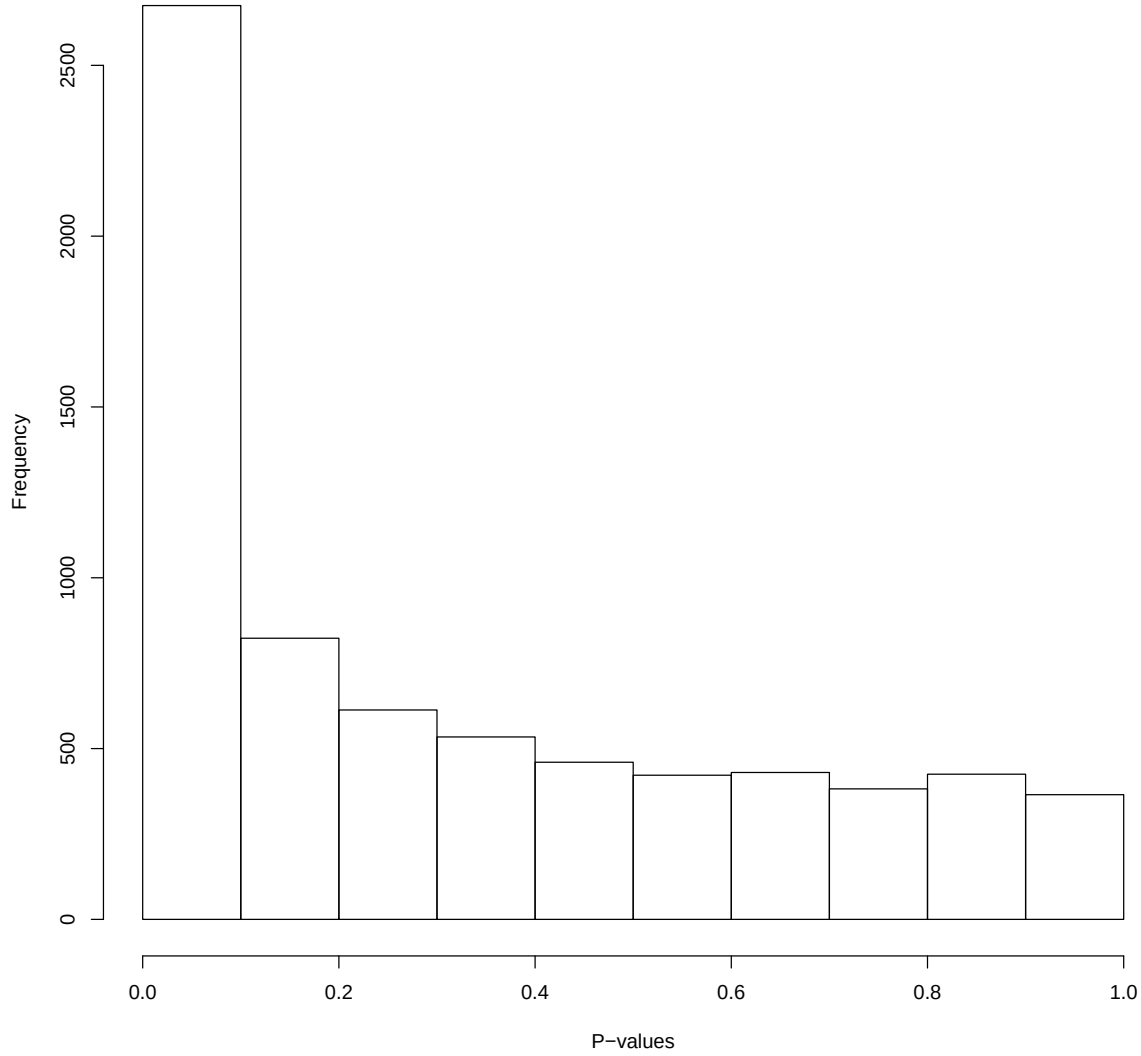
or biology-driven (pathways, ontologies), see [21].

Example: Lymphoma – running parallel t-tests, we find that 2079 or almost 30% of all genes have p-value below 0.05.

Example: Lymphoma – we apply multiplicity corrections. Note that the false discovery rate (FDR.BH, [1]) assumes that there is only a weak kind of dependency between genes.

Example: Lymphoma – the SAM approach finds varying number of DE genes at different levels of FDR, e.g. 731 at 4%.

Example: Lymphoma – [21] find a significant excess of regulated genes on chromosome 7.

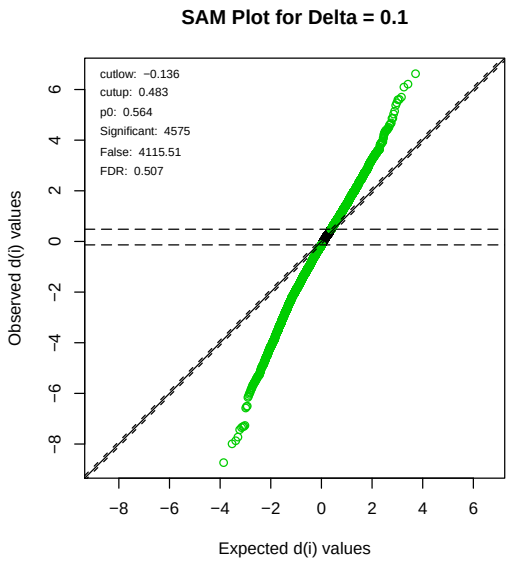

Frequency
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
P−values

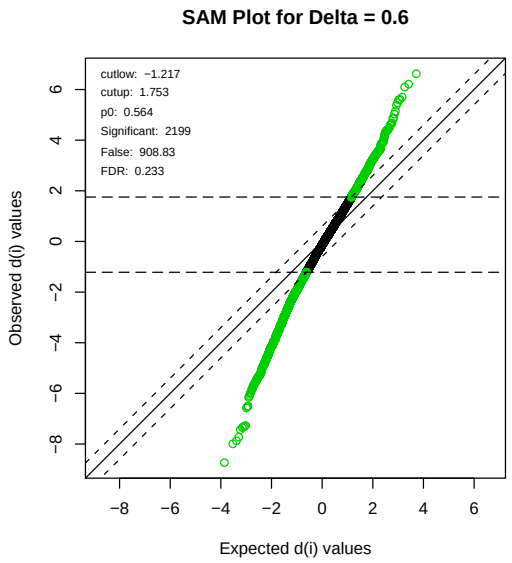
Several multiplicity corrections and FDRs based on the raw p-values when comparing ALL and AML yield the following number of regulated probe-sets:

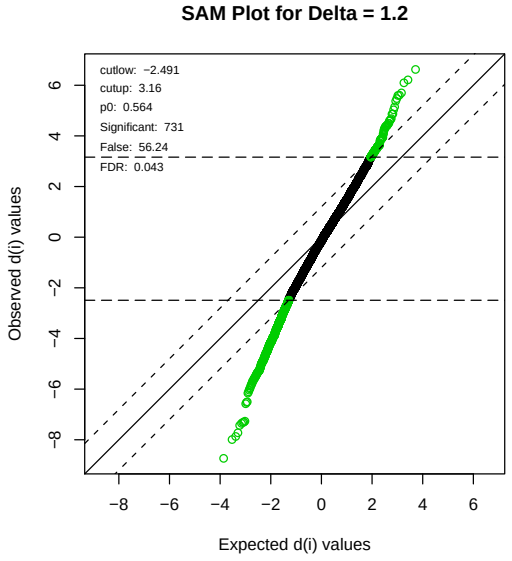
Cutoff	Raw p-values	Bonferroni	Holm	FDR.BH	FDR.BY
< 0.10	2675	192	192	1530	591
< 0.05	2079	164	165	1106	435
< 0.01	2675	192	192	1530	591

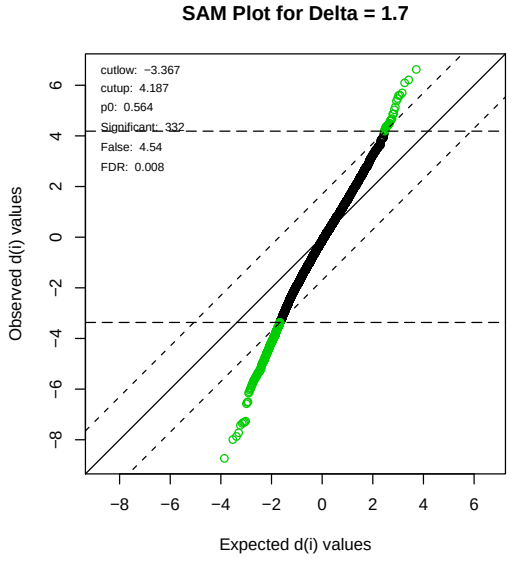
This table shows the trade off between specificity and FDR when running SAM:

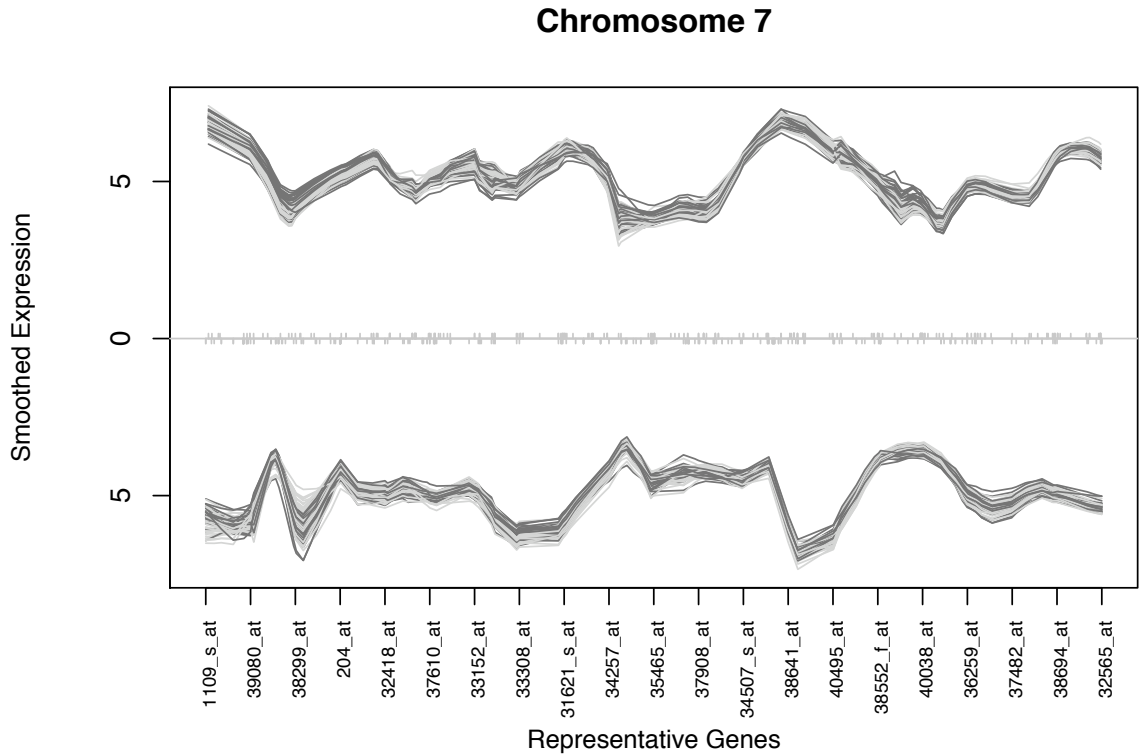
	Delta	p0	False	Called	FDR
1	0.1	0.5640	4115.5	4575	0.5073
2	0.6	0.5640	908.8	2199	0.2331
3	1.2	0.5640	56.2	731	0.0434
4	1.7	0.5640	4.5	332	0.0077
5	2.2	0.5640	0.2	144	0.0008


SAM Plot for Delta = 0.1
cutlow: −0.136
cutup: 0.483
p0: 0.564
Significant: 4575
False: 4115.51
FDR: 0.507
Observed d(i) values
Expected d(i) values


SAM Plot for Delta = 0.6
cutlow: −1.217
cutup: 1.753
p0: 0.564
Significant: 2199
False: 908.83
FDR: 0.233
Observed d(i) values
Expected d(i) values


SAM Plot for Delta = 1.2
cutlow: −2.491
cutup: 3.16
p0: 0.564
Significant: 731
False: 56.24
FDR: 0.043
Observed d(i) values
Expected d(i) values


SAM Plot for Delta = 1.7
cutlow: −3.367
cutup: 4.187
p0: 0.564
Significant: 332
False: 4.54
FDR: 0.008
Observed d(i) values
Expected d(i) values


Chromosome 7
Smoothed Expression
5
0
5
1109_s_at
39080_at
38299_at
204_at
32418_at
37610_at
33152_at
33308_at
31621_s_at
34257_at
35465_at
37908_at
34507_s_at
38641_at
40495_at
38552_f_at
40038_at
36259_at
37482_at
38694_at
32565_at
Representative Genes

Source: [21]

Summary

Low-level analysis is an active research area. While a strong signal will show up more or less regardless of the approach, a suitable choice will improve the power of the analysis.

While the distinction between class discovery, gene discovery and class prediction are in practice not rigid, it may be a good idea to position your research interest somewhere in the triangle.

Class prediction is hard. Prepare for a large sample size and repeat the trainings/test set split.

Detecting DE is hard unless you have a very strong fold change or many regulated genes.

Simple methods can be competitive, plus we understand them.

If you're going to be seriously involved with microarray expression analysis, have a look at Bioconductor [7].

References

[1] Y. Benjamini and Y. Hochberg. Controling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

[2] B.M. Bolstad. Comparing the effects of background, normalization and summarization on gene expression estimates. http://stat-www.berkeley.edu/users/bolstad/, 2002.

[3] B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

[4] Sung Choe, Michael Boutros, Alan Michelson, George Church, and Marc Halfon. Preferred analysis methods for Affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16, 2005.

[5] Susmita Datta and Somnath Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, 2003.

[6] Y. Ge, S. Dudoit, and T.P. Speed. Resampling-based multiple testing for microarray data analysis. *TEST*, 12(1):1–44, 2003.

[7] Robert Gentleman, Vincent Carey, Douglas Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[8] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

[9] R. Hoffmann, T. Seidl, and M. Dugas. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology*, 3(7):0033.1–0033.11, 2002.

[10] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:96–104, 2002.

[11] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[12] Jacques Lapointe, Chunde Li, John P Higgins, Matt van de Rijn, Eric Bair, Kelli Montgomery, Michelle Ferrari, Lars Egevad, Walter Rayford, Ulf Bergerheim, Peter Ekman, Angelo M DeMarzo, Robert Tibshirani, David Botstein, Patrick O Brown, James D Brooks, and Jonathan R Pollack. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A*, 101(3):811–6, Jan 2004.

[13] W.J. Lemon, A. Liyanarachchi, and M. You. A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biology*, 4(10):R67, 2003.

[14] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS*, 98(1):31–36, 2001.

[15] R.J. Lipshutz, S.P.A. Fodor, T.R. Gingeras, and D.J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21:20–24, 1999.

[16] Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458):488–492, 2005.

[17] Alexander Ploner, Lance Miller, Per Hall, Jonas Bergh, and Yudi Pawitan. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics*, 6(1):80, Mar 2005.

[18] J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

[19] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–5121, 2001.

[20] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[21] Anja von Heydebreck, Wolfgang Huber, and Robert Gentleman. Differential expression with the bioconductor project. *Bioconductor Project Working Papers*, 2004.

[22] Li Zhang, Michael F Miles, and Kenneth D Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotech*, 21(7):818–821, 2003.