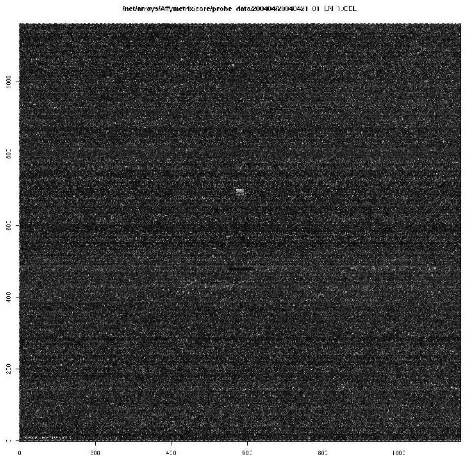
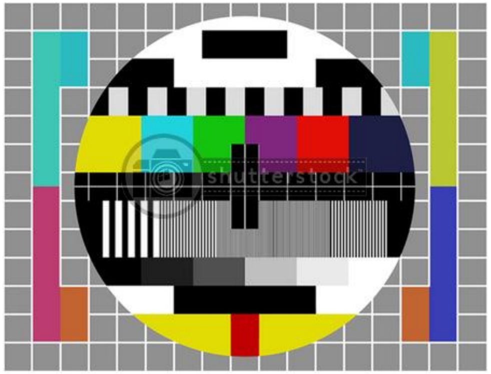
Microarray data analysis, intensive course 3.-5.9.2007.


/net/arrays/Affymetrix/core/probe_data/200404/20040421_01_LN_1.CEL
1000
800
600
400
200
0
200
400
600
800
1000



Signal? or... No signal?

Jukka Corander,Department of Mathematics, Åbo Akademi University
Laura Elo, BTK, Turun Yliopisto

Outline:

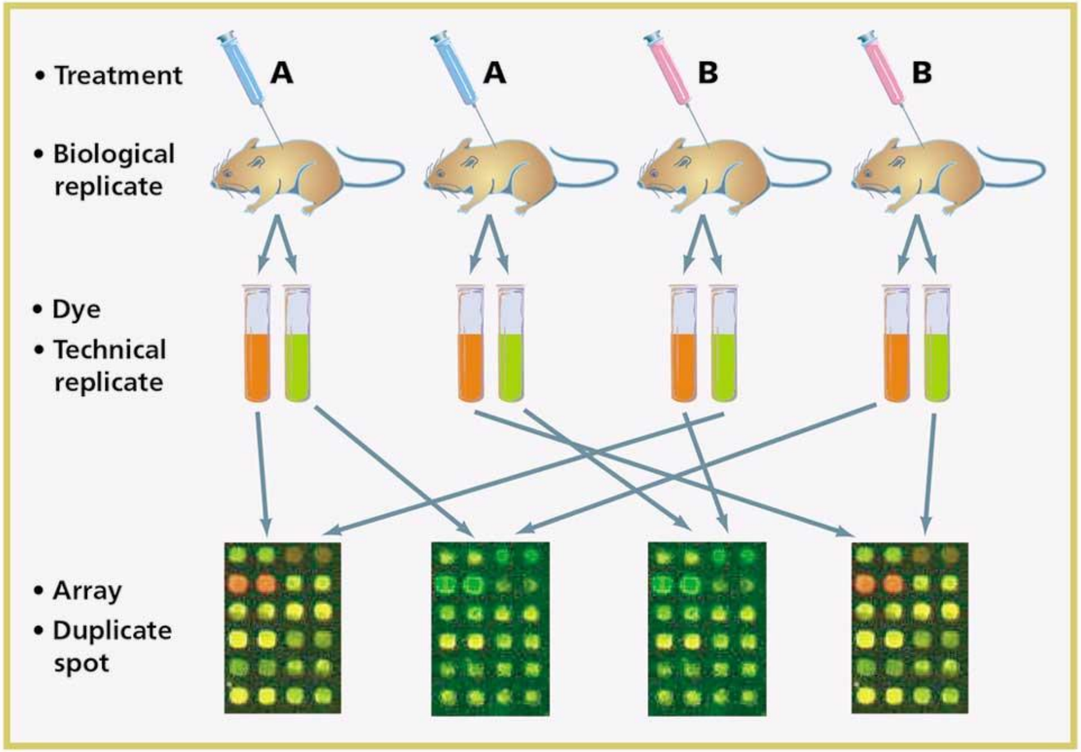
- Biology and some array basics have already been covered...

1. Basic design issues
2. Array image analysis
3. Pre-processing array data

1. Basic issues in designing a microarray experiment

- Replication of the biological samples – this is essential for drawing conclusions from the experiment.
- Technical replicates (two or more RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups.
- Spots of each cDNA clone or oligonucleotide are present at least as duplicates on the microarray slide, to provide a measure of technical precision in each hybridization (# probe replicates has been variable over Affymetrix array generations).
- Replication will be considered more in detail later, in the context of statistical modeling of array data.

- It is critical that information about the sample preparation and handling is discussed in order to help identify the independent units in the experiment as well as to avoid inflated estimates of significance.



Source: Churchill, Nature

Technical limitations of microarrays I

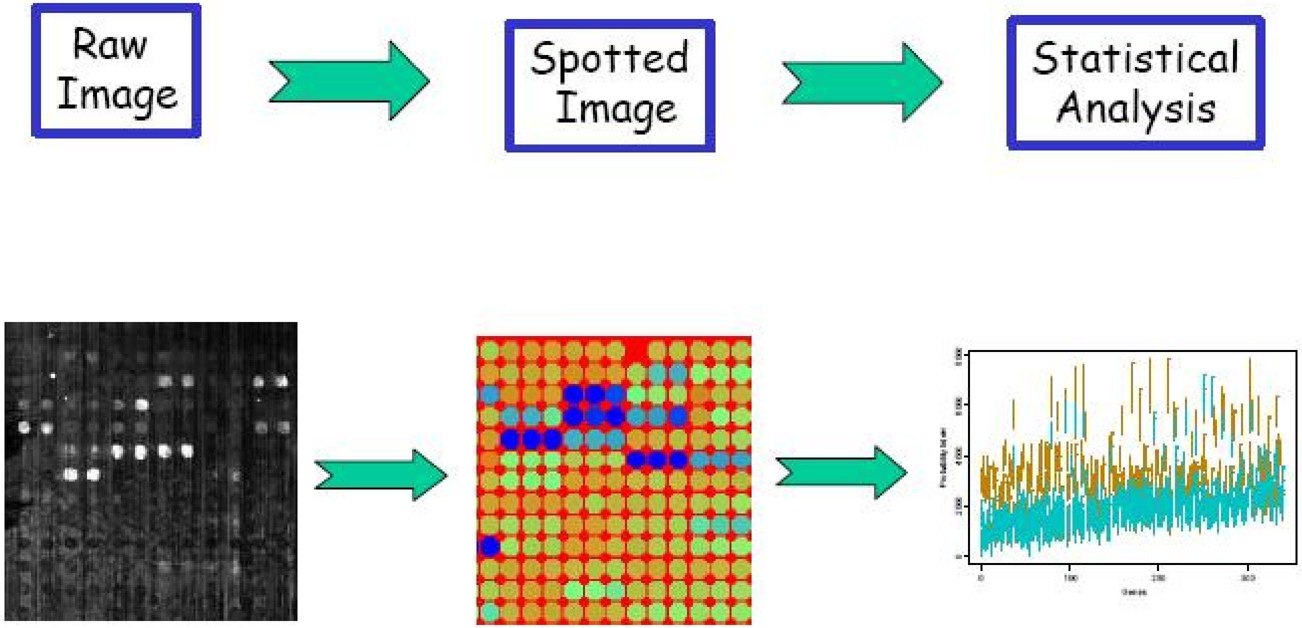
- Apart from the other limiting factors, such as pricing, access to high quality tissue/RNA samples, there are certain intrinsic factors that limit what can learned from microarray experiments.
- Expression does not necessarily lead to gene products (e.g. downstream regulation).
- Many cell processes change rapidly under certain conditions and it is difficult to track the changes through arrays (even with multiple arrays over time).
- Incorrectly labeled probe sets (probe/gene pairs do not match in reality.
- Errors in the oligos due to restrictions in the biochemical production process (e.g. can make some MM probes meaningless).

Technical limitations of microarrays II

- Errors in the cDNA probes, e.g. due to PCR amplification.
- Vague and complex connection between measured expression levels and the amounts of corresponding gene products in the cell.
- Complexity of gene networks makes it very difficult to infer associations and causalities between genes through expression patterns, even if they were completely unbiased by any factors!!
- The challenge of learning gene networks even with the best thinkable data is nicely discussed by Geier et al (BMC Syst Biol 2007).
- Replicability issues, measured expression levels can be highly variable over a set of samples (e.g. external factors in the experiment).
- Nevertheless, what alternative is out there?

2. Image analysis

Basic workflow for acquiring the data from an array.



Raw Image

Spotted Image

Statistical Analysis

Main steps in the conversion of raw image data to a spotted image

- Gridding
- Segmentation
- Quantification

However, the process may be more complicated than this, e.g. multiple imaging scans may be done for the same array, which means that the raw image data are not treated in isolation. More about this later...

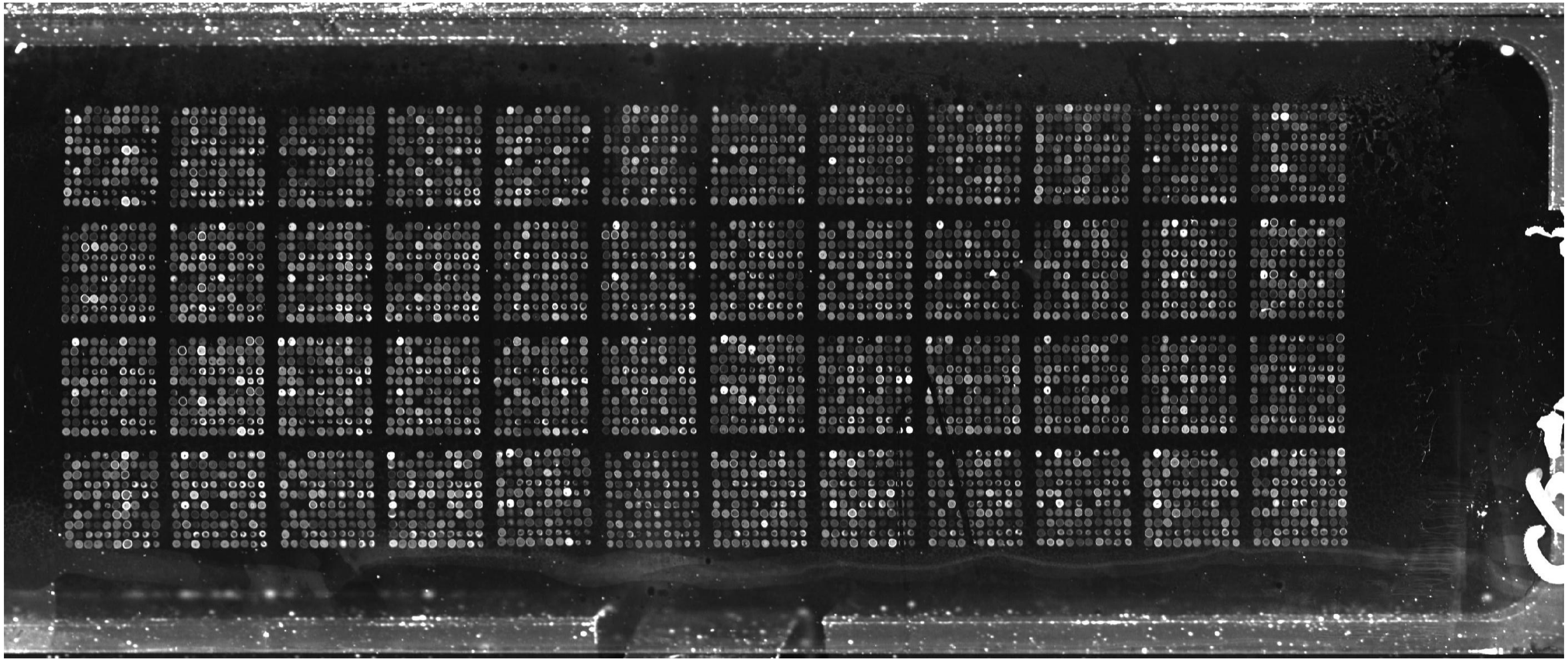
Gridding

- What is a microarray image?
- When the image of the slide is produced by a laser scanner, an array of *m* x *n* (e.g. 3000 x 1500) grayscale pixel values is obtained.
- Each pixel is typically a 16-bit or 24-bit intensity measurement, in the former case values range from 0 to 65535.
- If the image would be converted into 8-bit format, the range is just from 0 to 255, which would induce a considerable loss of information.
- In a 2-channel experiment two images are produced for the same array using different wavelengths of the laser, leading to images typically labeled by ”red” and ”green”, respectively.
- The main target of the image analysis is to estimate an expression level value from the pixel intensities representing the same gene.
- In cDNA arrays the ’genes’ are represented as ’spots’ along a rectangular grid and in oligo arrays the genes are squares placed next to each other (see the examples below).
- The estimation task is not exactly trivial...

A reminder

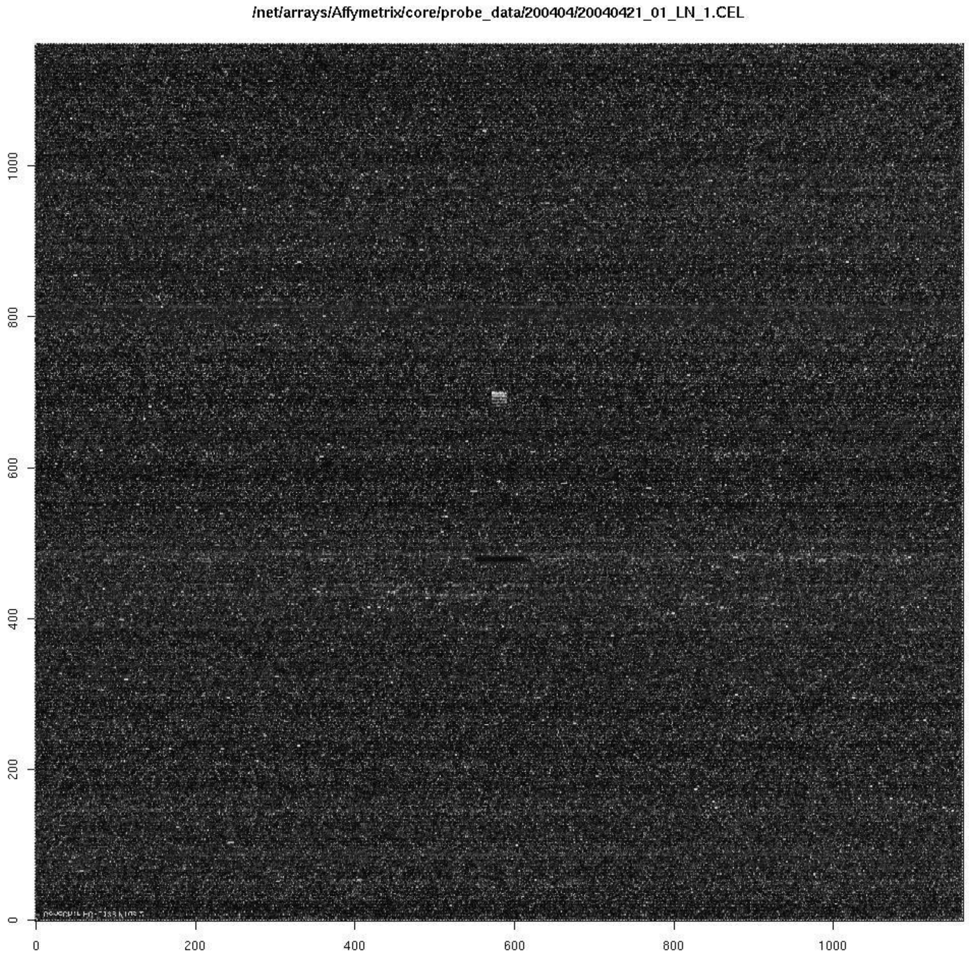
- To detect which cDNA (or cRNA for oligonucleotide arrays) binds where, the samples are labeled with reporter molecules, fluorophores that emit certain type of light when exposed to a given wavelength.
- The number of fluor molecules that label each cDNA depends on its length and sequence composition.
- This is one reason why fluorescent intensities for different cDNAs cannot be directly compared.
- However, identical cDNAs will still be comparable as long as the samples are prepared analogously.

Typical example of a cDNA raw image

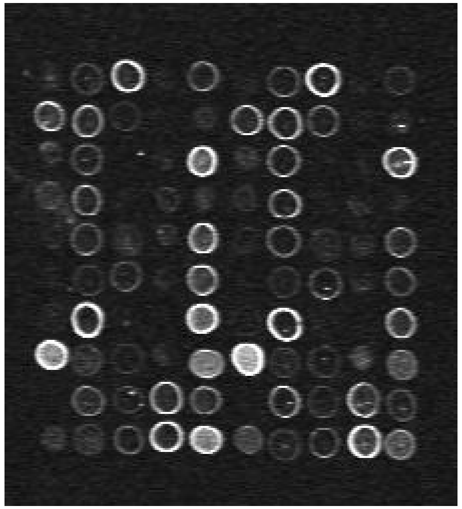


Not exactly a price winner for a crisp, smudge-free panorama landscape shot...

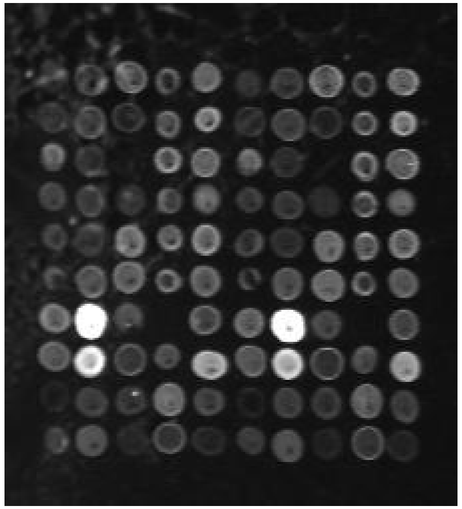
Typical example of an Affy raw image


/net/arrays/Affymetrix/core/probe_data/200404/20040421_01_LN_1.CEL
0
200
400
600
800
1000
0
200
400
600
800
1000

Lets have a closer look at one of the squares for both channels (the cDNA array example)



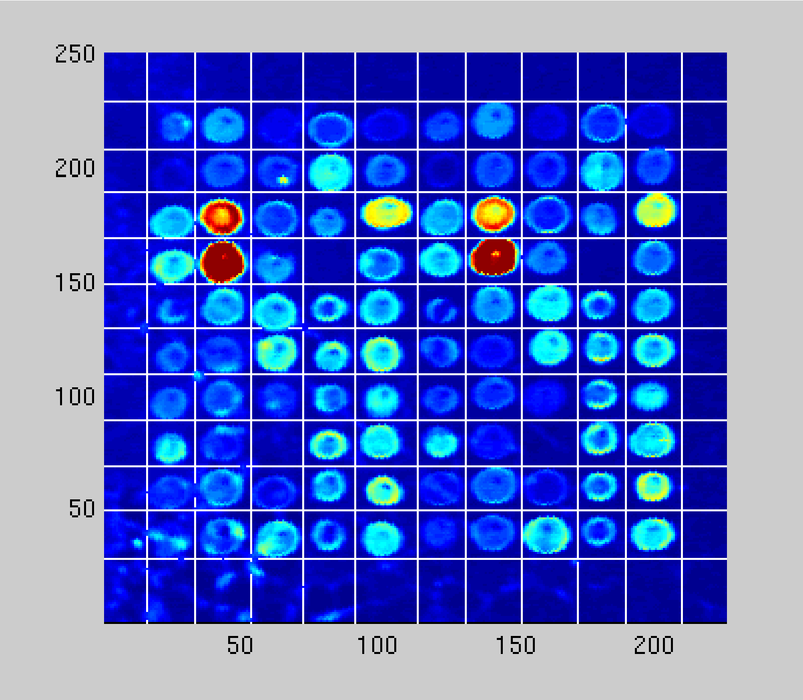
Red



Green

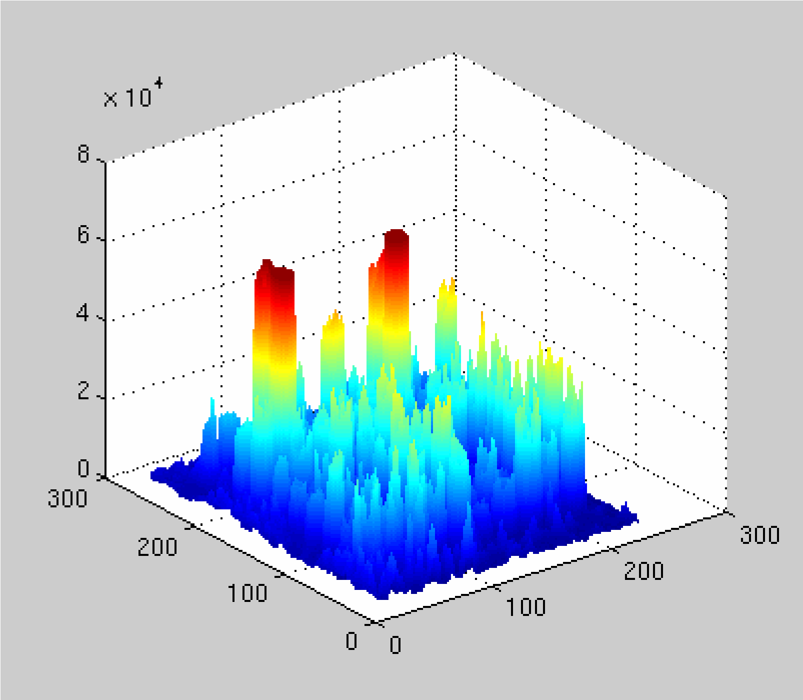
Notice that the ’spots’ are most often not really spots, but ring-like shapes with varying sizes and distances from each other.

By gridding we end up with a locally adapted grid, e.g. by using first Gaussian kernel smoothing on the image raw intensity values.



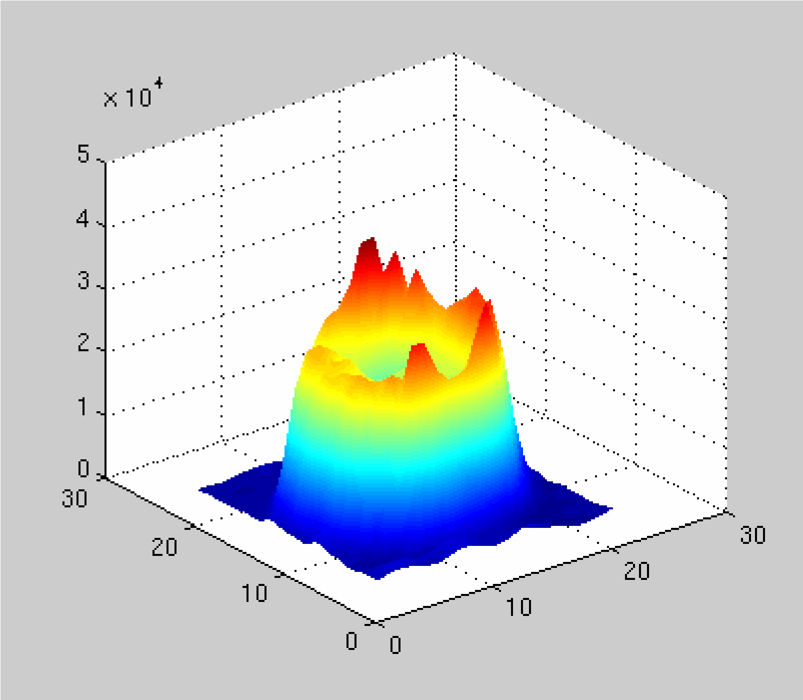
Segmentation

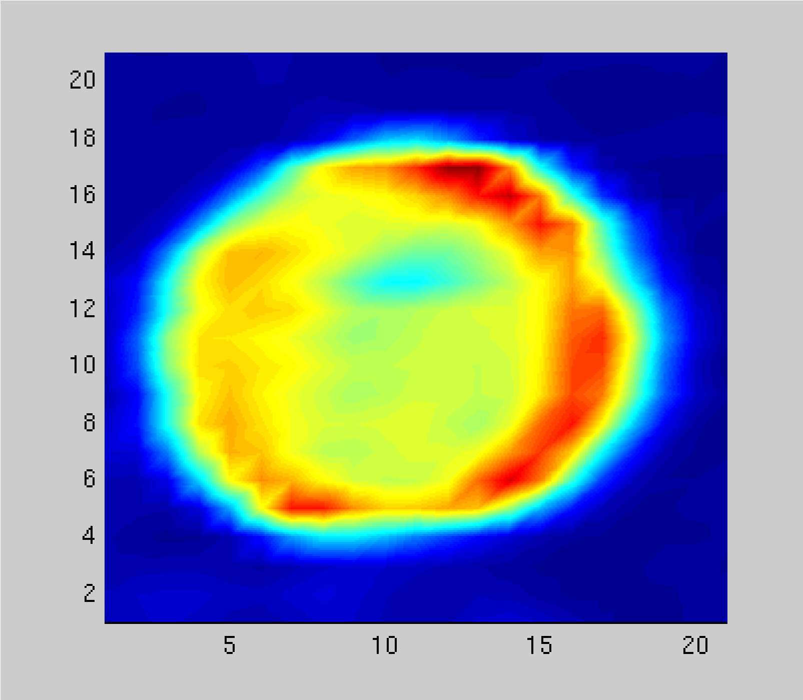
- Segmentation refers to the separation of the ’spots’ from the background.
- Here again locally adaptive methods are necessary for achieving reliable results.
- The picture below shows better the intensity variation across the investigated region.



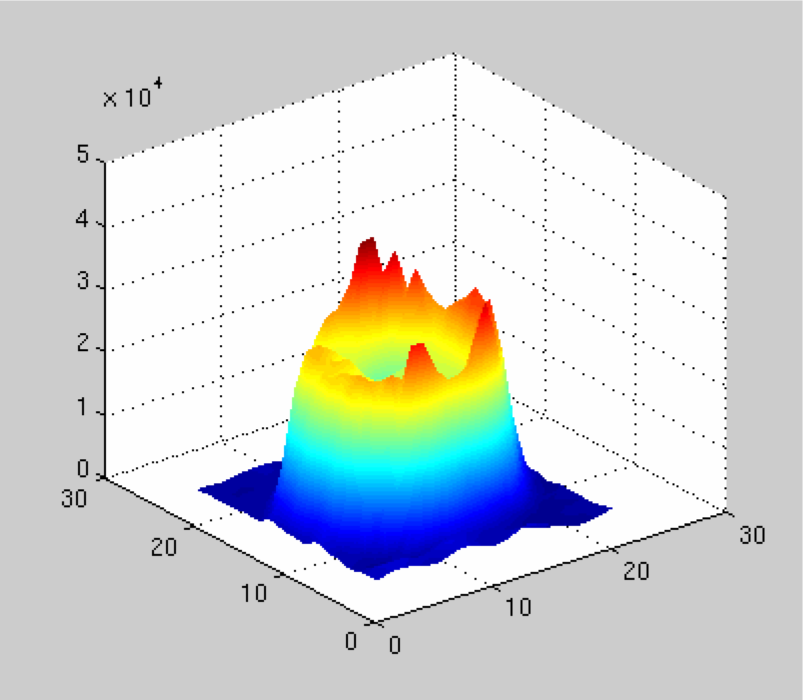
Quantification of expression level

Lets have a closer look at one of the spots in the square...



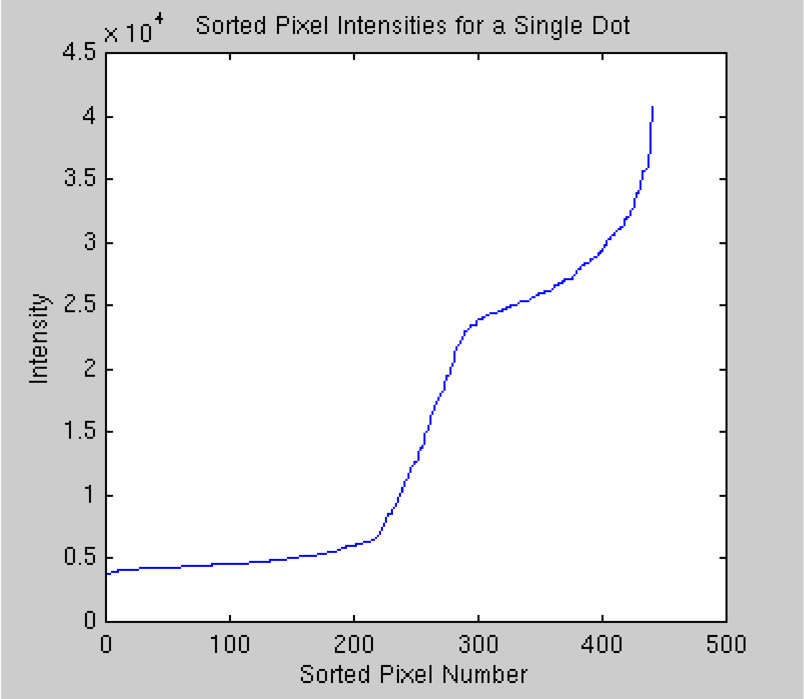


The ’spot’ characteristics suggest the following:



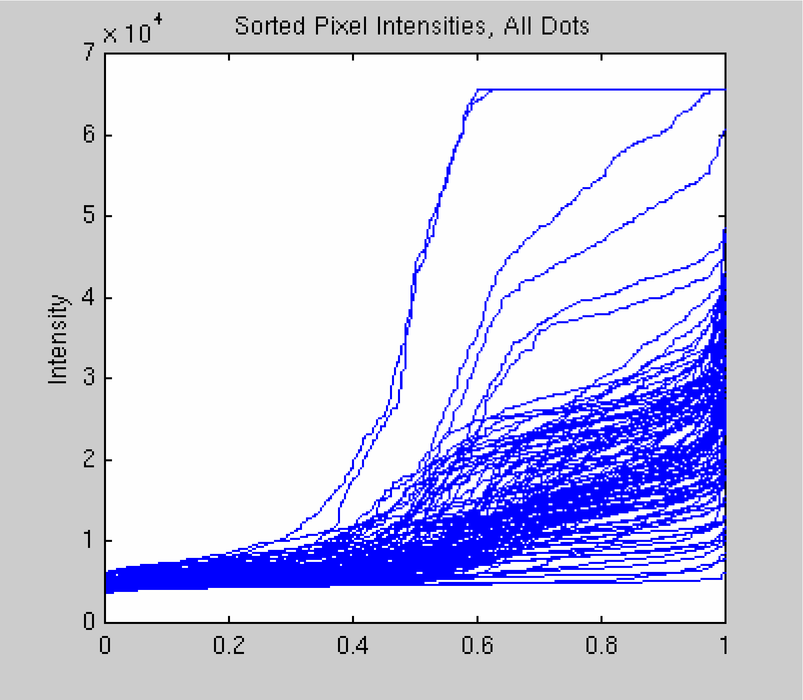
- It is not trivial to estimate the intensity of the ’spot’.
- Suggestions include a.o. mean, median and the total intensity over the pixels associated with the ’spot’.
- Nevertheless, the ’local background’ should be used as a reference, because the background intensities vary over the array parts.

How would the estimates behave?



- The picture above has 441 pixels.
- The sorted intensities are shown to the left.
- Notice that nearly half of the pixel have background level intensities, rendering median estimate pretty much useless.
- This suggests that a single number alone poorly captures all necessary information about spot intensities.
- Therefore, various spot intensity statistics are used.

Lets look at many spots simultaneously



- There are many different shapes of intensity distributions.
- Some spots have saturated signals.
- This means that the actual intensity would have been higher than, e.g. the upper limit determined by the 16-bit image.
- Saturated signals can lead to biased ratios of expression values.

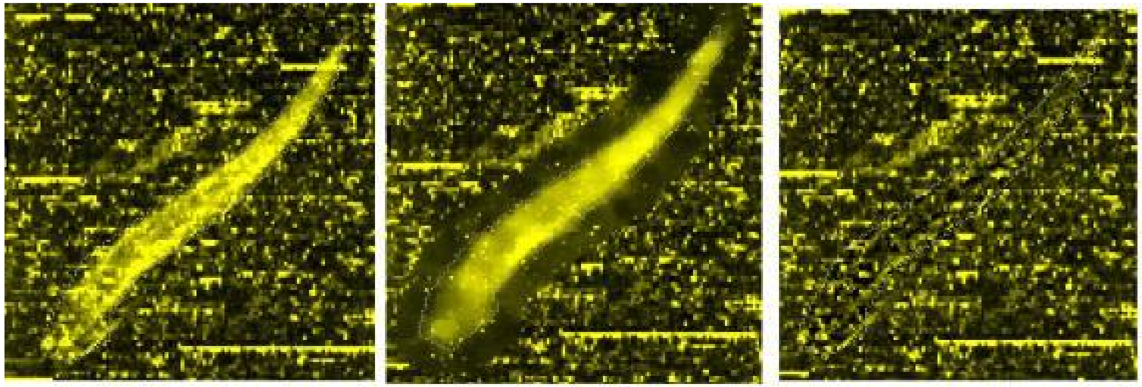
Quality assessment and adjustment of intensity estimates

- Once the spotted image is acquired and the related statistical characterizations of the intensities are available, it is advisable to assess the quality for the individual spots as well as for the whole array.
- Abnormalities in the intensity values due to scratches, dust, etc, should be removed/corrected.
- Otherwise they may bias the statistical modeling done later.
- High-resolution graphics are an efficient means for quality assessment.

Spatially organized defects in the array

- A smudge covering a substantial part of the background image, which shows higher or lower intensities compared to the rest of the image.
- Vertical, horizontal or diagonal strips in the background that show higher or lower intensities.
- A gradient in the background intensities across the array.
- A row or column effect.
- Bleeding, i.e. series of consecutive spots blurred together.

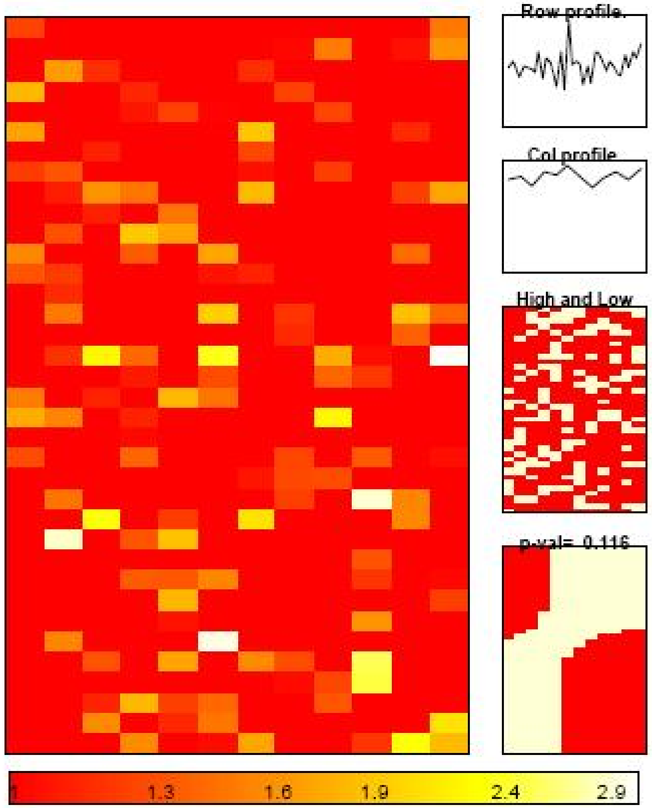
Example of defects in an Affy image



- There are computational/statistical methods for correcting contaminated regions in the array, e.g. dChip software contains implementations of such approaches.

- The methods use adaptively specified regions to derive corrected intensity values for the background and signal.

Spatial randomness of the intensities of an array can be examined in various ways.



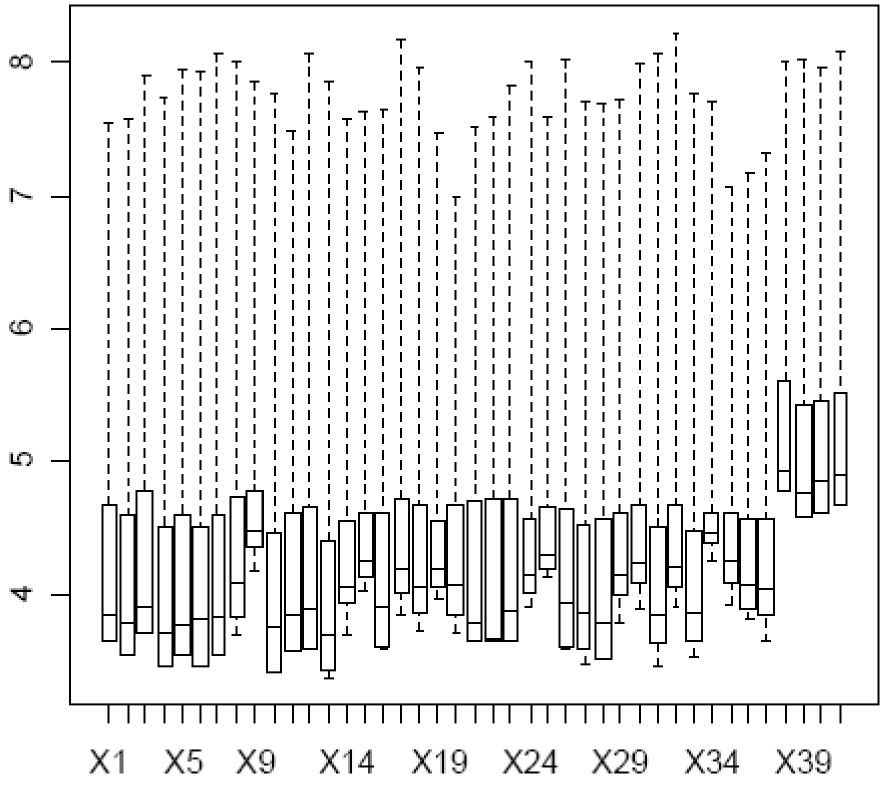
Row average spot intensities.

Column average spot intensities.

Binarization into high vs low spot intensities.

Discriminant analysis of the binary intensities using spot coordinates (white = incorrect predictions). P-value for the proportion of correctly predicted spots can be obtained by randomly permuting the spot intensities a large number of times.

Quality control of arrays

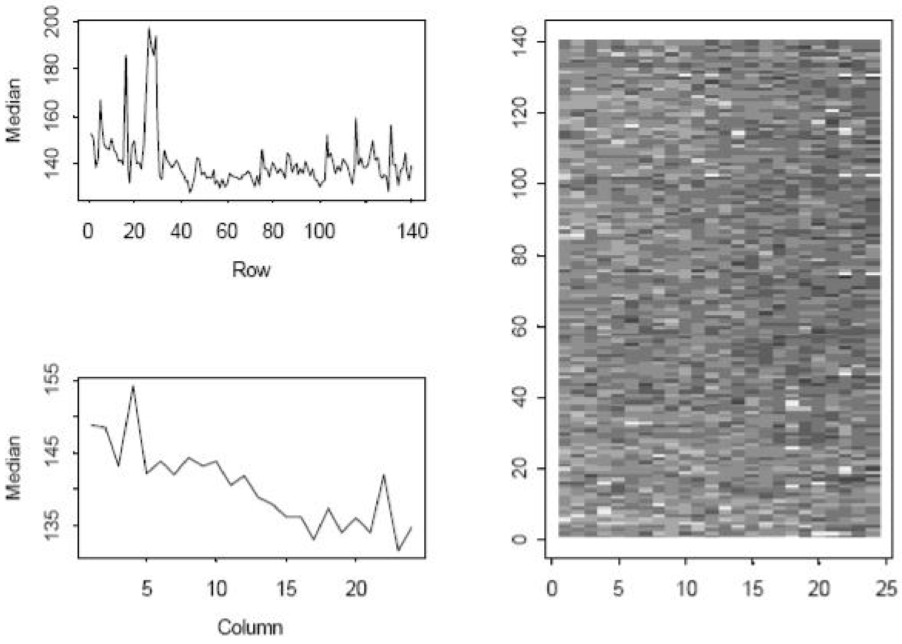


- Array experiments are often performed over time.
- There may be unnoticed changes in the experimental setup, that cause shifts in the general properties of the arrays.
- Boxplots of summary statistics ordered over time can reveal such events.
- In picture such an event has affected the last four arrays.

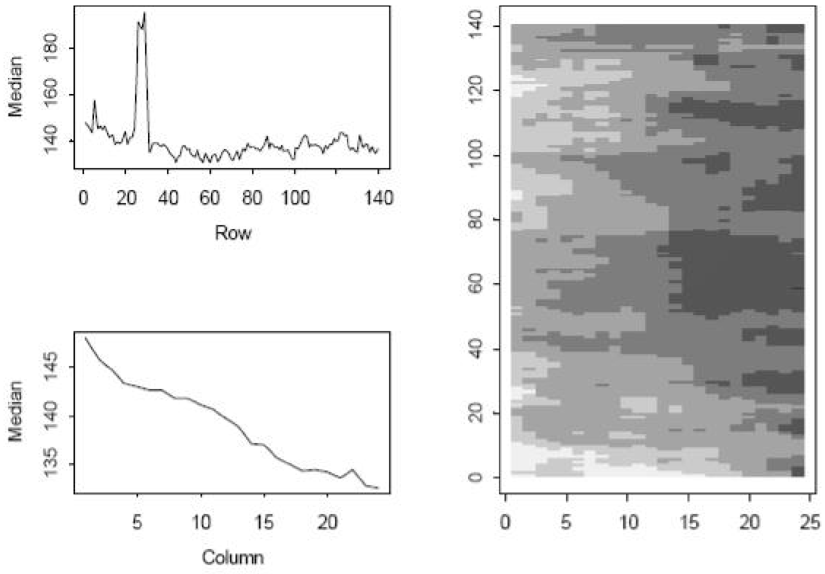
Adjusting for the background

- Pixels outside regions of interest (spots/squares) would ideally have zero intensities in the image.
- Unfortunately, this is not true because of the substrate fluorescence and non-specific binding.
- Also, spots may contain a part of the background fluorescence.
- Therefore, it is typically necessary to adjust the spot intensity values for the background effects.
- The aim is to have as correct estimate as possible for the expression signal associated with a spot.

Graphical presentations of the background behavior



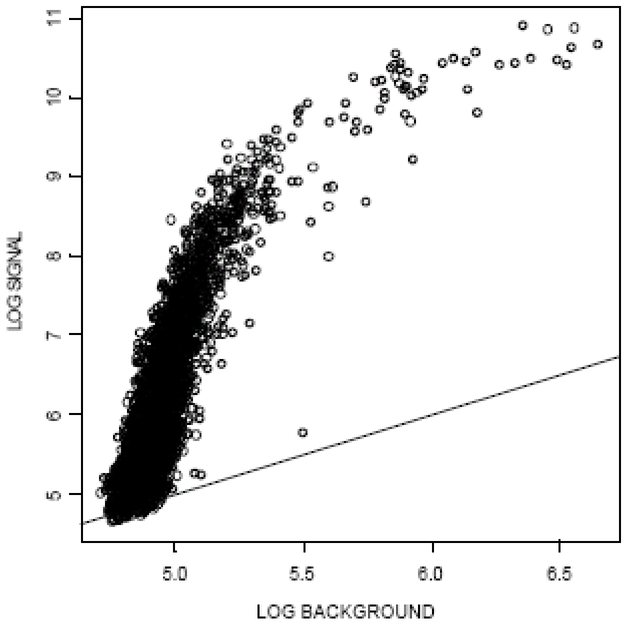
Original background intensities.



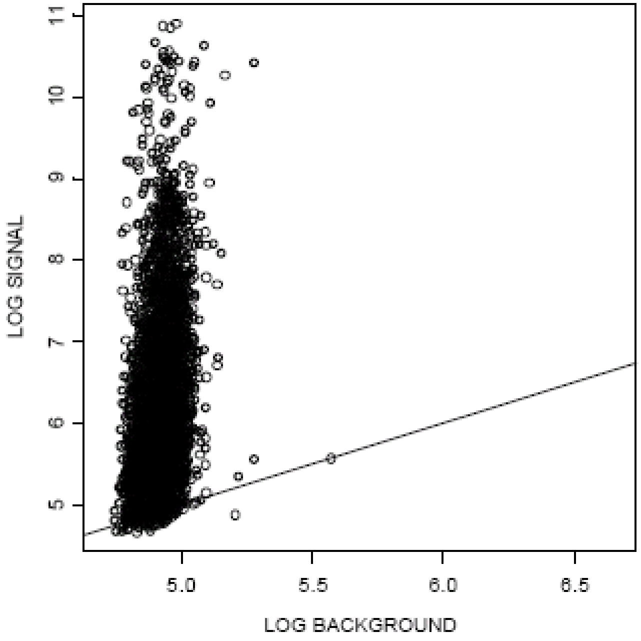
Locally smoothed background intensities.

Notice that even after the smoothing, the background intensity is not uniformly distributed over this array. This should be taken into account in the adjustment.

Graphical presentations of the signal vs. background behavior



Original background intensities.



Locally smoothed background intensities.

Notice the tendency of high background values to be associated with a higher signal (=spot intensity).

An example of background adjustment

- Let SI_g denote the estimated intensity for spot (=gene) g, and BI_g the estimated background intensity for the same location.
- Further, let T be a low percentile of the SI_g values (such as the 5th percentile).
- Then, an adjusted signal intensity may be defined as:

$$AI_g = \max(SI_g - BI_g, T)$$

However, it is still pretty much an open question how to best adjust for a spatially non-uniform background variation!

Estimating expression levels for Affy arrays

- Currently, in the Affymetrix arrays there are 11 pairs of Perfect match (PM) and Mismatch (MM) probes (earlier versions had 20 and even more).
- The PM probes represent the ‘correct’ version of a gene, whereas the MM probes contain a complementary base in the middle of the oligonucleotide sequence.
- As opposed to the early versions of the Affymetrix GeneChip, the MM and PM probes for a single gene are currently scattered over the array to avoid experimental defects masking the information for both (which could happen earlier).
- The idea with the MM probes is to provide a control measurement of non-specific binding for genes.
- Unfortunately, whilst the idea seems to be good in theory, in practice there is a lot of controversy about the use of the MM probes.
- Consequently, some people tend to ignore them in their experiments.
- Also, a portion of the probes in the Affy chip for the human genome are incorrectly annotated, as shown by some very recent research, together with corrected annotations.

An example of expression level estimation using PM and MM probes

- Let PM_{gi} and MM_{gi} denote the background-adjusted estimated intensity for the perfect match and mismatch probe i (i=1,...,n_g) of the gene g.
- Currently, n_g = 11.
- PM_{gi} - MM_{gi} acts as a measure of the hybridization level of the ith probe.
- A simple estimate of the expression signal for g is then:

$$S_g = \Sigma_\iota (PM_{gi} - MM_{gi}) / n_g$$

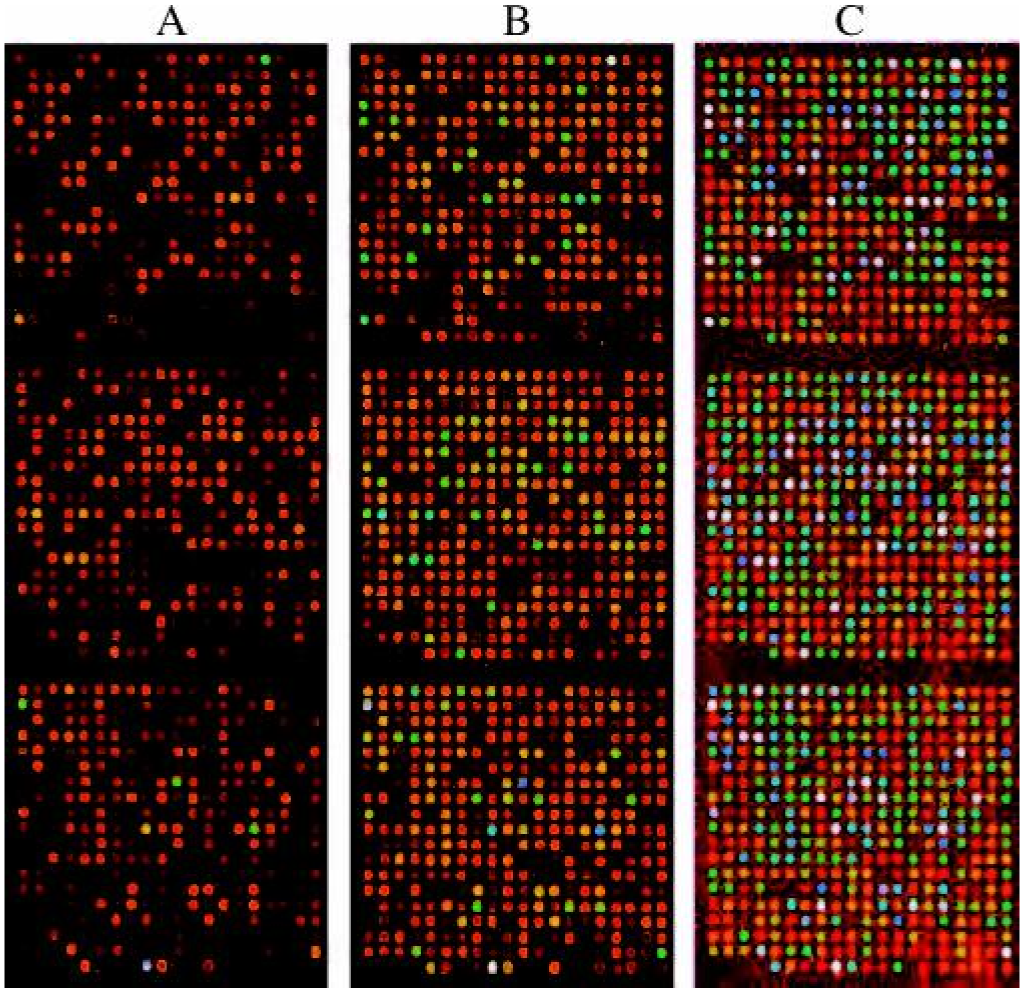
- However, the relationship between PM and MM probes has been shown to be complicated as the MM probes may contain a part of the true target signal.
- Also, it has been detected that the relationship between the intensities of the two probe types is often nonlinear, which calls for other methods.
- Lauren (IEEE Trans Nanobiosci, 2003) discusses various ways of statistical modeling to estimate the expression levels for Affymetrix arrays without using MM probes.

Multiple scans

- An issue that has so far largely been overlooked in the microarray community is that of choosing the laser intensity for scanning of the array.
- Recent research has shown that this aspect is important and should be considered when planning the experiments.
- The problem stems from the fact that a single laser intensity cannot optimally measure expression levels for all different types of genes expressed in the sample data.
- At low laser intensities, primarily highly expressed genes can be discovered.

- Conversely, with high laser intensities, the signal will be saturated for moderately or highly expressed genes, whereas weakly expressed genes will be better captured.
- For example, Skibbe et al. (Bioinformatics, 2006) showed that a multiple scan approach using varying laser intensities can better detect expressed genes than a single-intensity approach.
- Typically, a low-medium-high range setting would be sufficiently informative, however, this area would still benefit from further research.
- The picture below illustrates the observed changes in expression levels at different laser intensities.

Imaged expression levels for Maize arrays at low (A), medium (B) and high (C) laser intensities (from Skibbe et al 2006).



- Several recent papers deal with the same issue, Khondoker et al (Bioinformatics, 2006), Piepho et al (Bioinformatics, 2006), Gupta et al (Statistical Applications in Genetics and Molecular Biology, 2006).
- These papers illustrate nicely how a statistical model-based approach can combine information from several sources to improve inferences compared to the situation where such information is used in isolation.
- In summary, the better the estimation and capture of the expression levels represent the underlying biological reality in the samples, the better the downstream inferences related to the experiment.

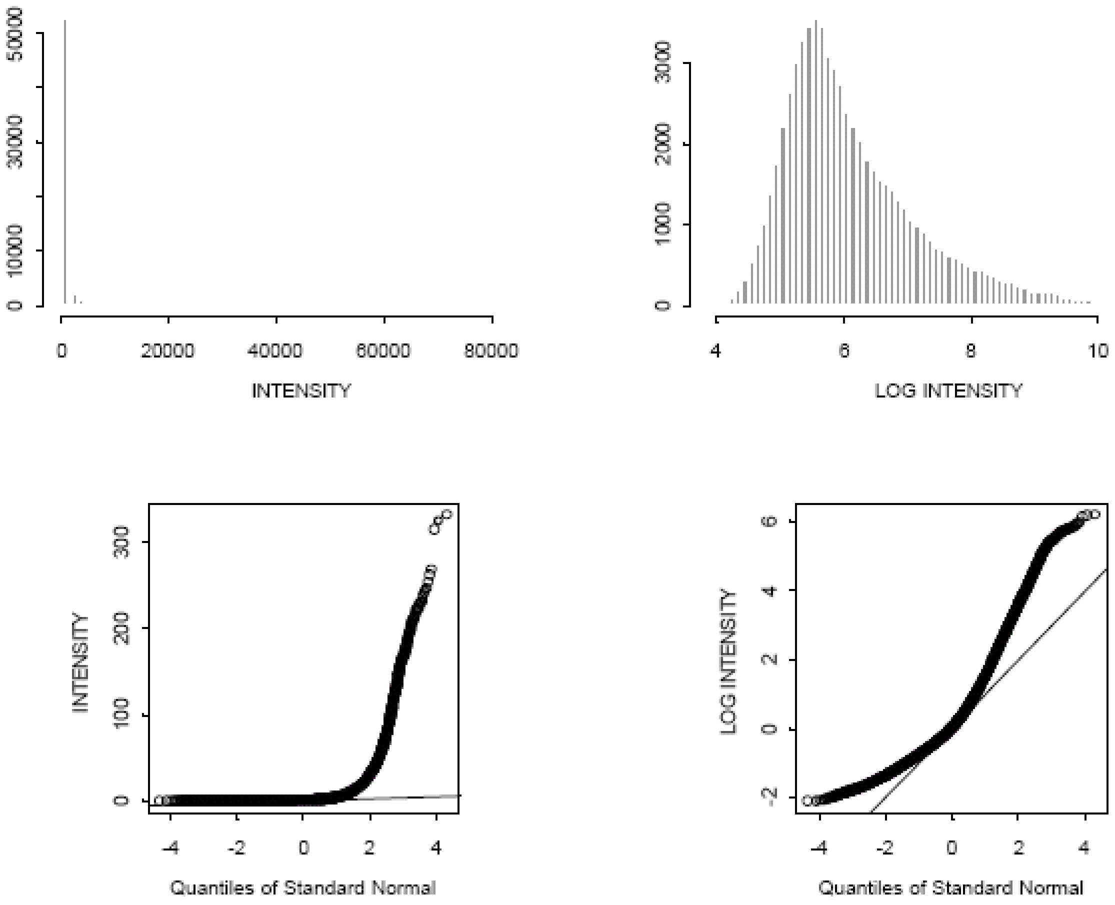
3. Preprocessing data from a microarray experiment

- The need for preprocessing the array data continues even after all the fuzz involved in the image analysis.
- Some immediate goals are as follows.
- We need to transform the data into a scale suitable for statistical analysis.
- Effects of unwanted systematic sources of variation need to be removed.
- Discrepant observations and arrays should be identified.

An example

- To illustrate the need for preprocessing, the following type of data will be considered.
- An experiment with 10 pairs of arrays C1A,C1B,...,C10A,C10B was performed.
- Each pair of arrays correspond to a single mRNA sample (labeled C1,...,C10) which was taken from a mouse and hybridized to two separate arrays (A and B).
- The two arrays in each pair are *technical replicates* as they are exposed to the same biological sample.
- The five mice from which samples C1,...,C5 were taken are controls, so they are biological replicates.
- The remaining mice were each treated with a different drug.

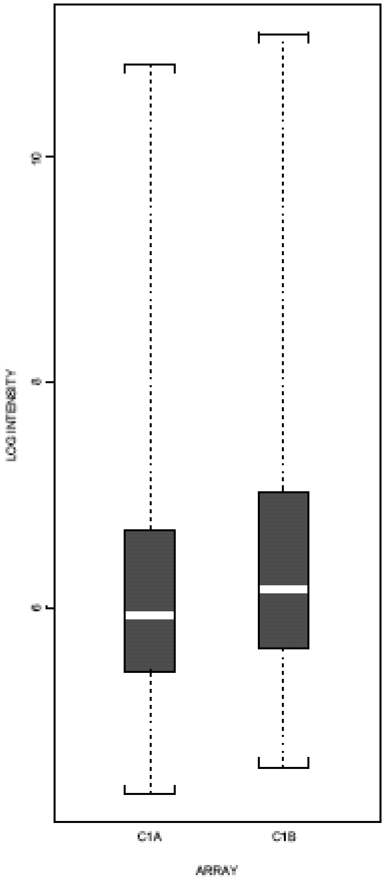
The usefulness of a logarithmic transformation illustrated by the data from array C1A.

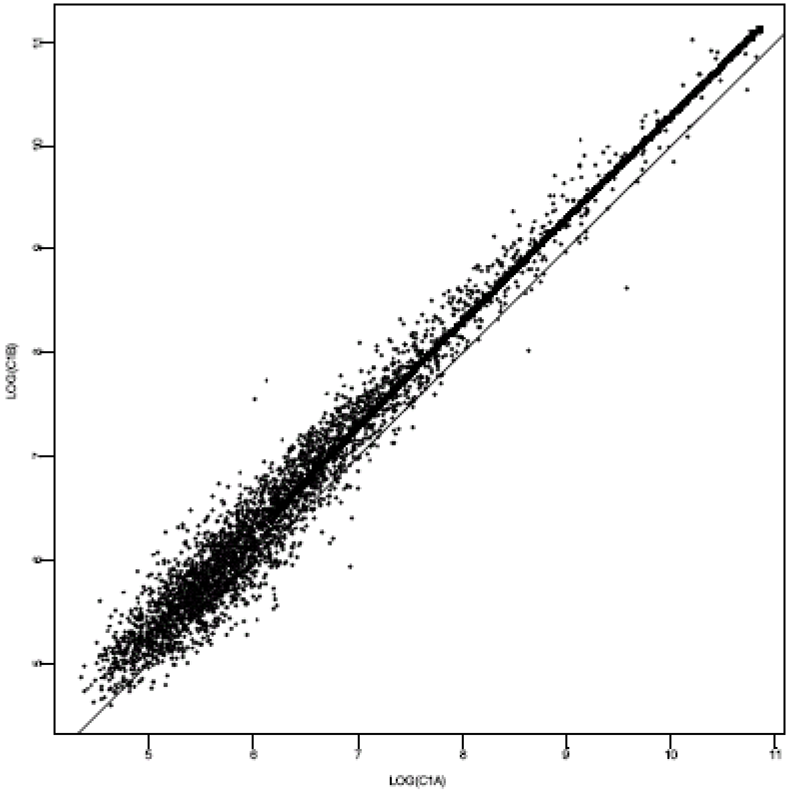


Identified sources of bias

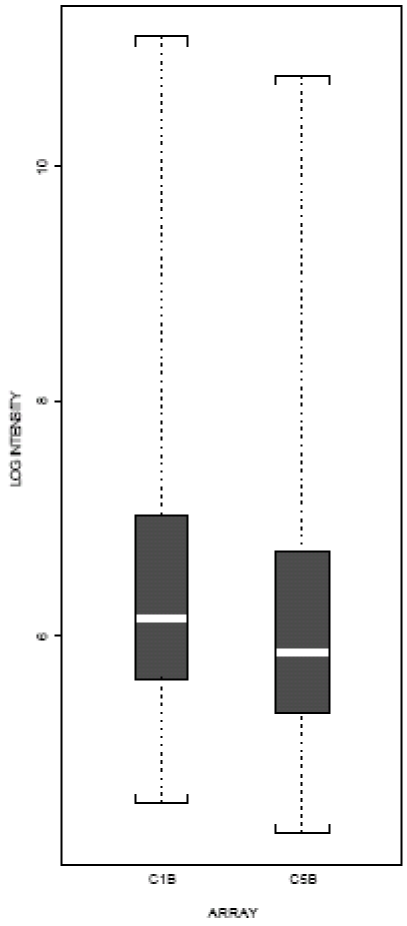
- Concentration and amount of cDNA placed on arrays.
- mRNA preparation.
- Scanner settings.
- Saturation effects.
- Equipment wearing out.
- Dye effect (red intensities higher than green)
- ...
- As such systematic effects affect different arrays in different ways, the data needs to be modified so that valid comparisons can be made, i.e. the observations must be brought onto a common scale.
- *Normalization* refers generally to the removal of systematic variation in the array data.

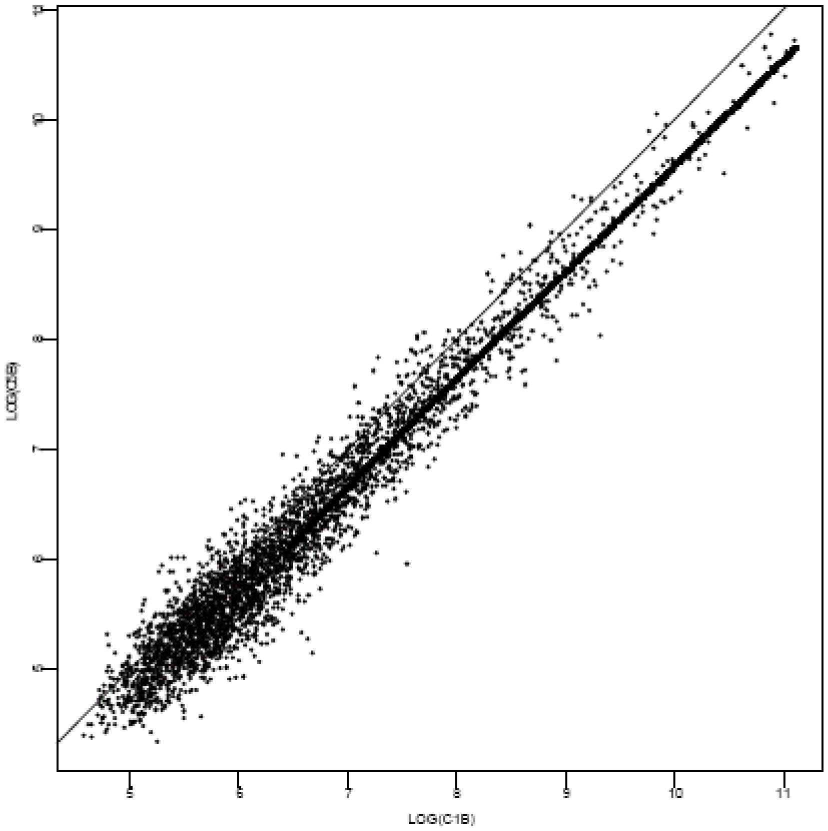
Example. Log-intensities for arrays C1A and C1B (both are from the same biological sample).





Another example. Log-intensities for arrays C1B and C5B (from two distinct control samples).

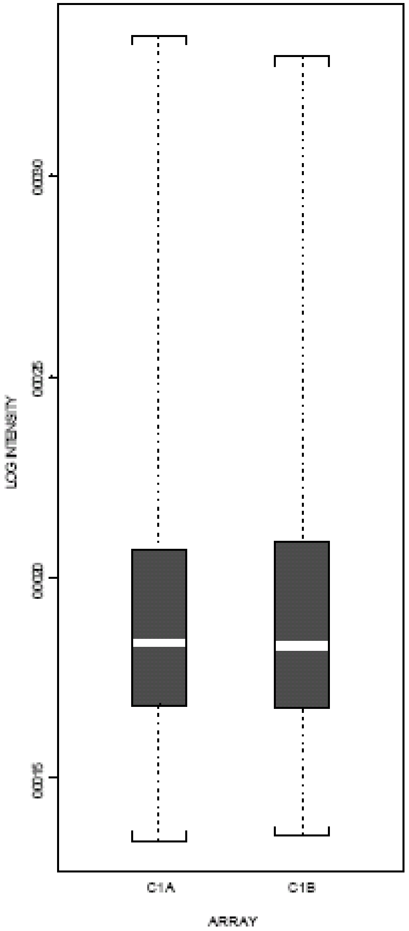


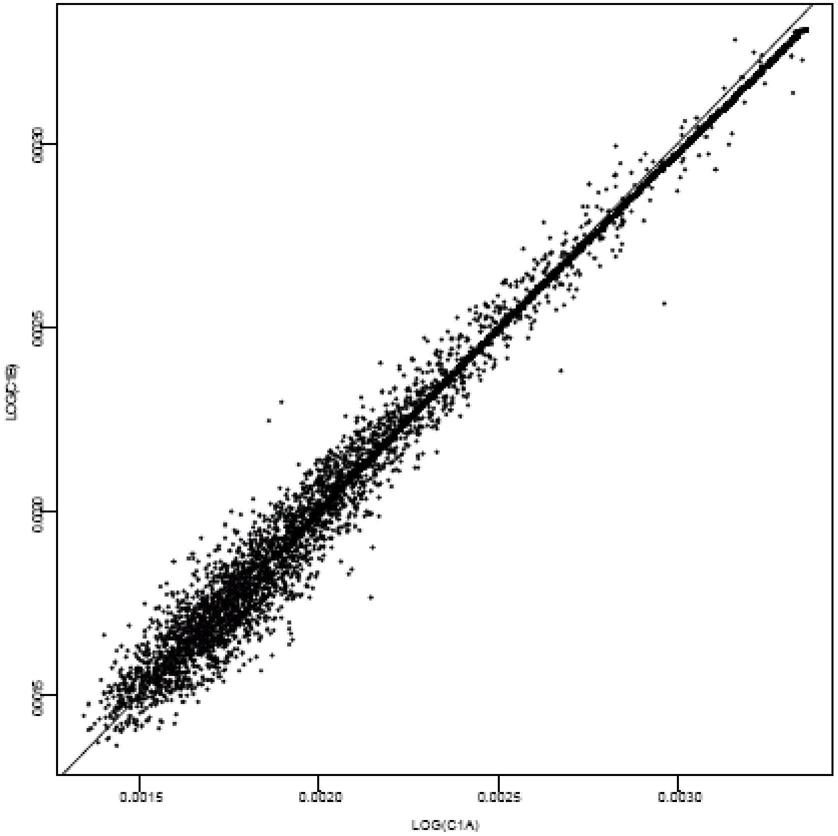


Simple normalization methods (to already log-transformed data)

- Normalization by the sum, where the sums of intensities of *m* arrays are forced to be equal (divide the values of an array *i* by the sum over *i*).
- The rationale is that the total amount of mRNA should be roughly equal across the samples.
- An equivalent method is to normalize the arrays to have equal arithmetic means (an example is given on the next page).
- Related methods are median and quartile normalizations.
- All these are examples of *global* or *linear* normalization schemes.
- These and the more advanced methods discussed later apply to cDNA as well as oligonucleotide array data.

Example. Log-intensities for arrays C1A and C1B after mean normalization.





The rationale behind linear normalization methods

- Spot intensities on every pair of the considered arrays are assumed to be linearly related without any intercept.
- This means that the lack of comparability is corrected by adjusting every spot by the same amount, regardless of its intensity.
- That amount is called normalization factor.
- As the relationships between the intensity values are most often nonlinear, more sophisticated methods are to be preferred.

Intensity-dependent normalization methods

- The nonlinear relation between the intensity values for a pair of arrays suggests the following:
- The factor necessary to adjust low-intensity measurements should be different from the corresponding factor for high-intensity measurements.
- A large number of such intensity-dependent methods have been proposed.
- These methods use a reference or baseline array against which all the considered *m* arrays are normalized.
- One possibility for constructing this is a median array, where the values are given by $median\{X_{g1},\ldots,X_{gm}\}$, which is the median out of the *m* values for gene *g*.

- However, it is usually a good strategy to apply a linear normalization to the data (such as median or Q3 normalization), *before* the reference array is constructed.
- This will bring all arrays to a common overall level so that each of them can contribute to the construction of the reference array.
- A key for the more sophisticated normalizations is the selection of an *invariant gene set*, which will be used to estimate the normalization functions.

Invariant genes should have the following characteristics:

1. Constant expression levels across the considered arrays.
2. Their expression levels should span the entire range of expression values to avoid the need of extrapolation.
3. The normalization relationship for these genes across the arrays should be representative for all arrayed genes.

Invariant genes could be:

- Control genes present on the array solely for the normalization purposes (possible problems with condition 3).
- Housekeeping genes (possible problems with conditions 1 & 2).
- Unchanging genes. These could be chosen from the raw data by looking for genes that are least differentially expressed over the arrays (possible problems with conditions 1 & 2).
- All the genes on the array. As a very small number of genes are differentially expressed in a typical study, the ones doing that will only induce a small bias when the array contains many genes.

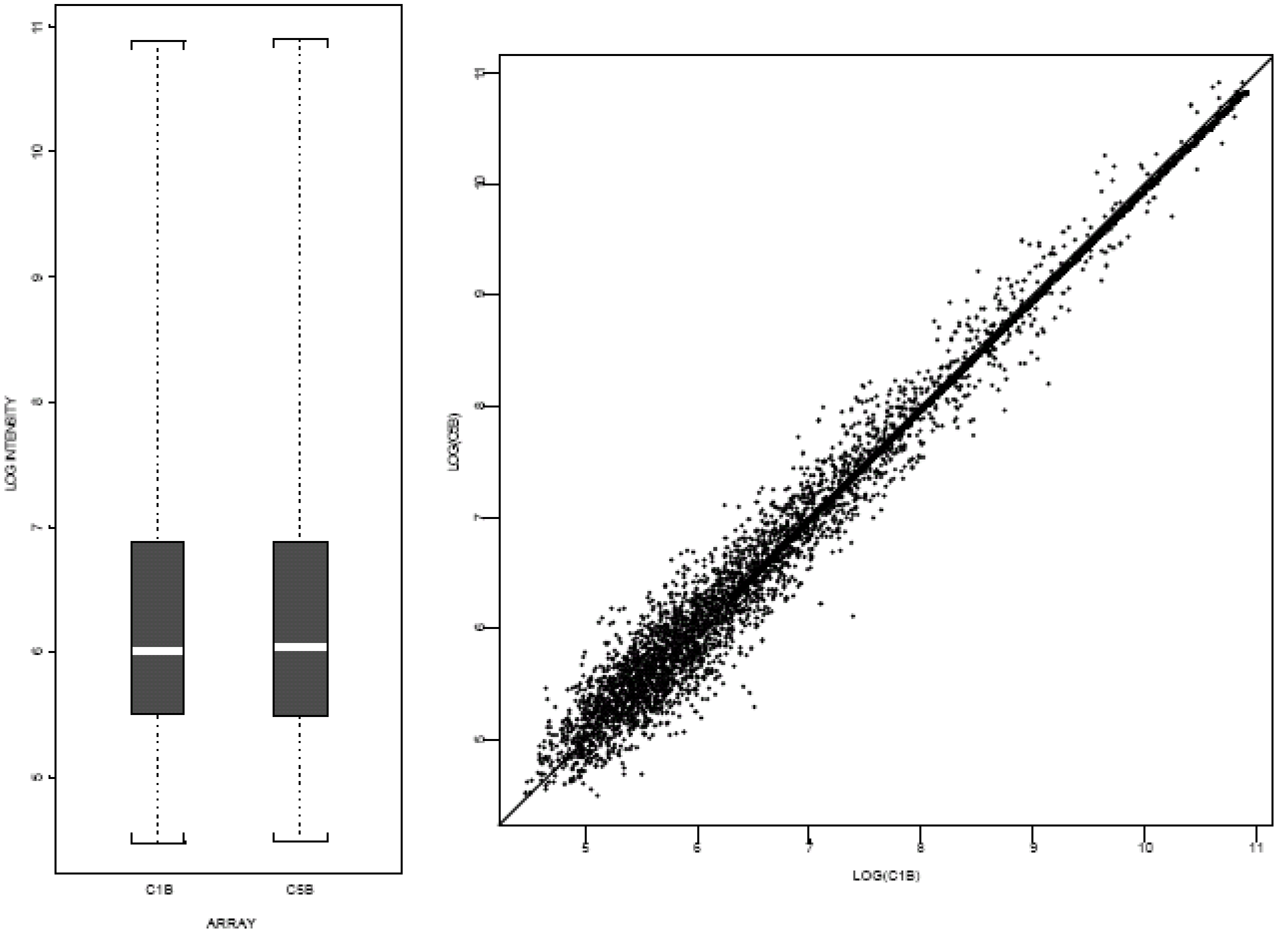
Smooth function normalization

- In *smooth function* normalization each array is normalized by fitting a locally adaptive function to the invariant gene set.
- Let *i* index an array.
- Let M_g be the median intensity for the gene *g* in the invariant set.
- With n genes in the invariant set and m arrays, we can fit the following model:

$$X_{gi} = f_i(M_g) + \varepsilon_{gi}$$

- The transformed values, say X_{gi}^{*}, for all genes in the *i*th array are then obtained by using the inverse of the smoother function $X_{gi}^{*} = f^{-1}(X_{gi})$.
- Typically, splines or kernel-based methods are used as smoothers.
- The rationale is to locally capture the non-uniform variation of the intensities with respect to the median intensity across the range of different intensity values among the invariant genes.

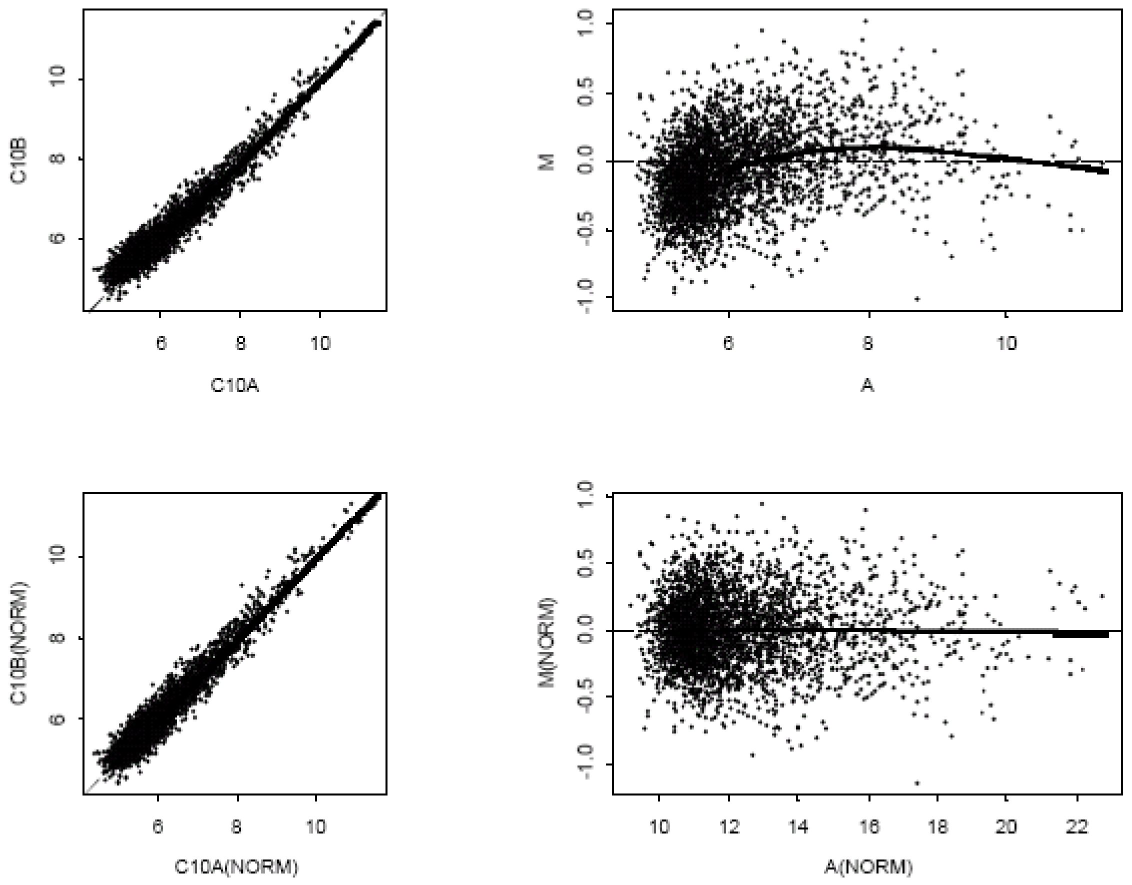
Example. Log-intensities for arrays C1A and C5B after spline normalization.



Normalization of the dye effect

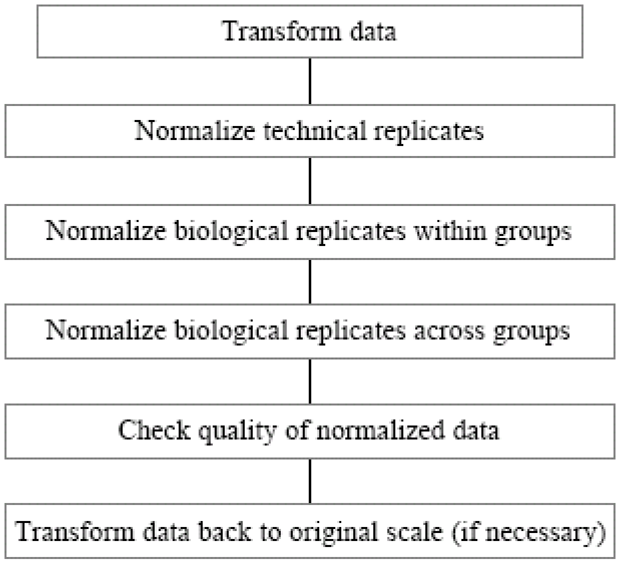
- Consider log-intensities X_{gR} and X_{gG} for the red (R) and green (G) channels of a cDNA array experiment.
- If there is no systematic dye bias, the values of X_{gR} versus X_{gG} should fall on the diagonal when plotted.
- If there is a systematic deviation, a normalization is necessary.
- However, it is easier to detect deviations using an MVA plot according to the following.
- Let $M_g = X_{gR} - X_{gG}$ and $A_g = (X_{gR} – X_{gG})/2$.
- In case of no dye bias, the values of M_g versus A_g should be nicely scattered around the zero line.
- If they are not, then the previously discussed normalization methods can be applied to remove the dye bias.

Example. Scatterplots of the log-intensities for arrays C10A and C10B before and after normalization.



Stagewise normalization

- When the data includes both technical and biological replicates, it is more practical to do the normalization in stages.



Transform data

Normalize technical replicates

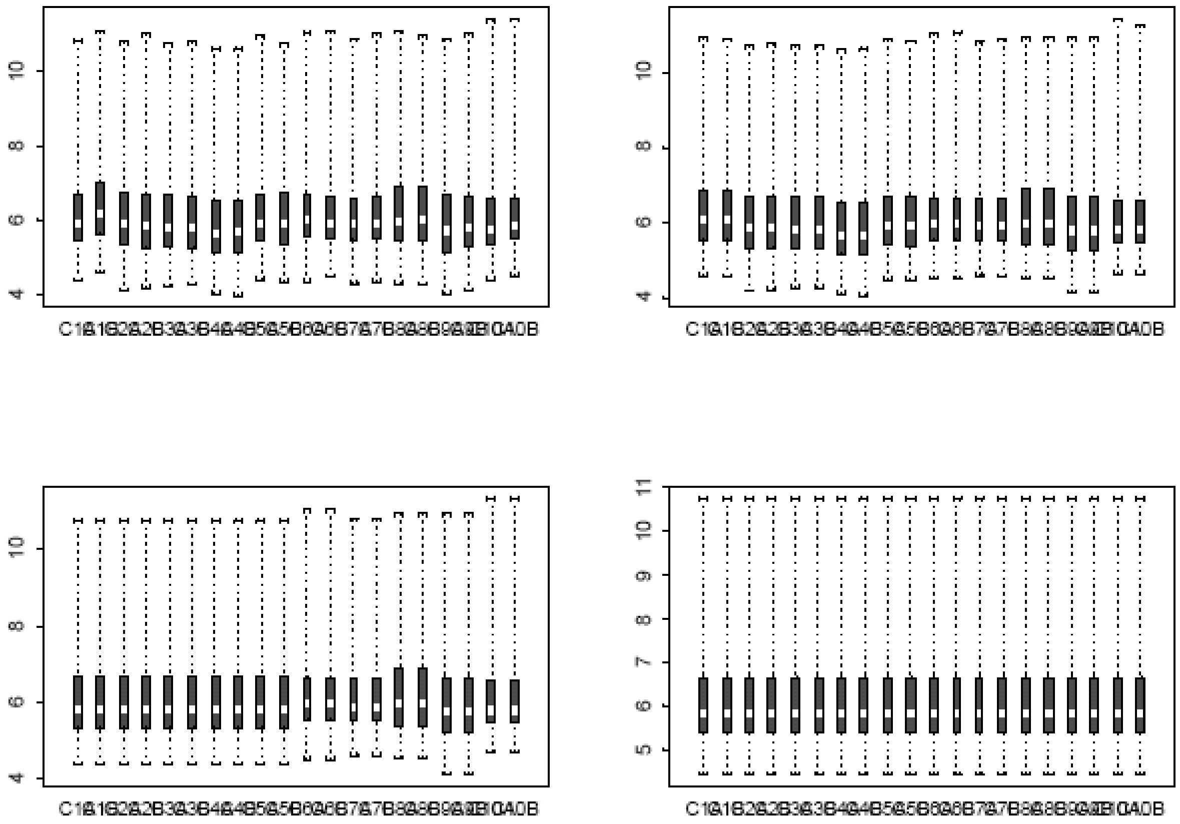
Normalize biological replicates within groups

Normalize biological replicates across groups

Check quality of normalized data

Transform data back to original scale (if necessary)

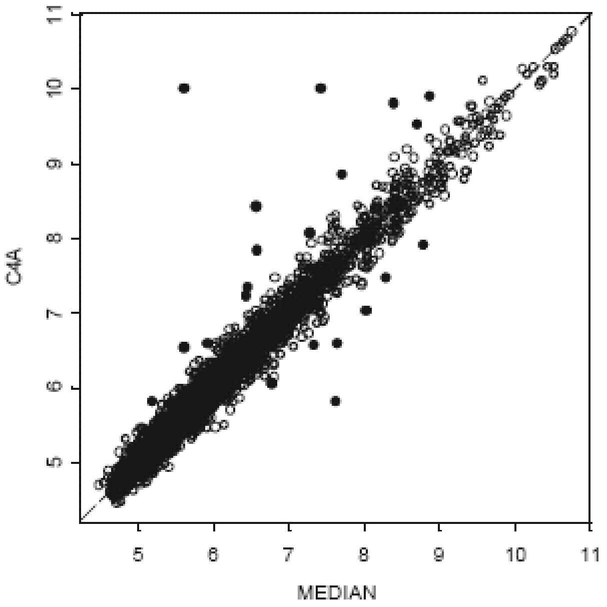
Figure 5.10: Side-by-side boxplot displays of arrays C1A to C10B at various stages of stagewise normalization (S0) before normalization, (S1) after normalization of technical replicates, (S2) after normalization of biological replicates in control group, (S3) after quantile normalization

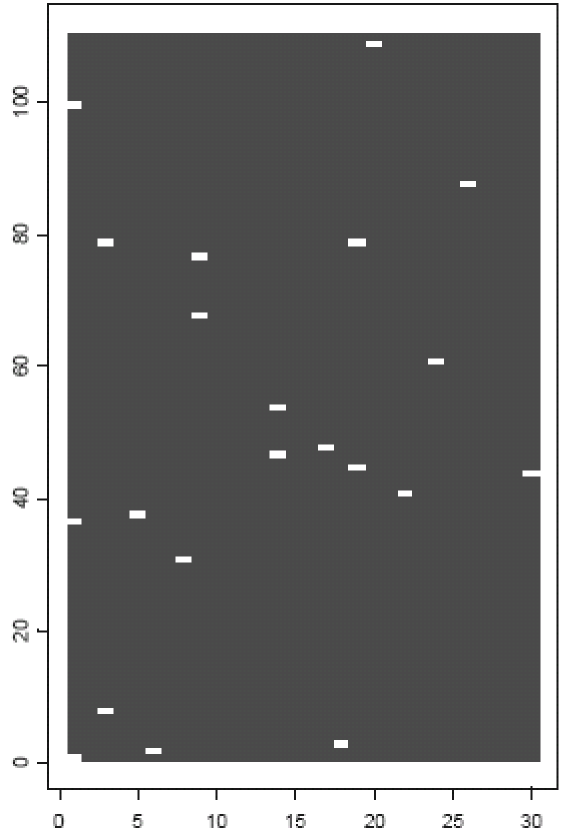


Outlier detection

- We end the preprocessing phase discussion by considering outliers.
- Outliers can be interpreted as observations that appear to be inconsistent with the majority of the data.
- Note that in certain settings outliers are ’true’ observations, in the sense that they are not produced by measurement errors.
- Here graphics are again valuable.
- Many general statistical methods exist for this purpose.

Example. Scatterplot for the normalized array C4A and the corresponding image plot (putative outliers indicated by the filled circles and the corresponding white squares at the locations of the genes in the array).


C4A
MEDIAN



0
20
40
60
80
100
0
5
10
15
20
25
30