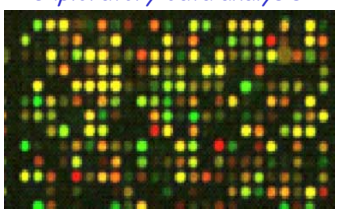
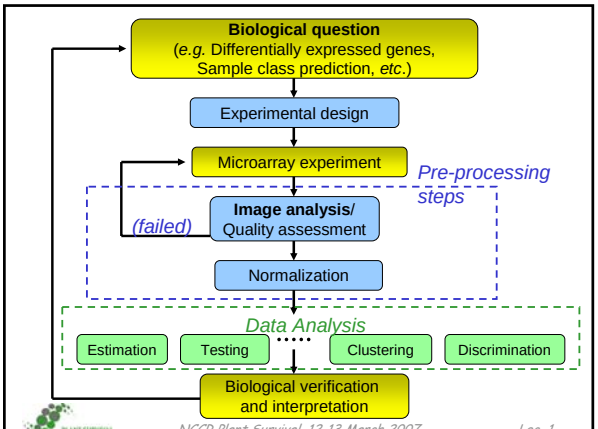
Statistics for cDNA Microarrays

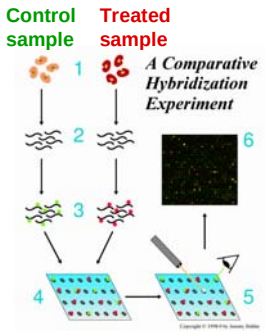
Image Analysis; Quality assessment and exploratory data analysis



http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/



Microarray Experiment



Measures of center: Mean

- The *mean* value of a variable is obtained by computing the total of the values divided by the number of values
- Appropriate for data patterns that are fairly *symmetrical*
- It is sensitive to presence of outliers, since all values contribute equally

Measures of center: Median

- The *median* value of a variable is the number having 50% (half) of the values smaller than it (and the other half bigger)
- It is NOT sensitive to presence of outliers, since it 'ignores' almost all of the data values
- The median is thus usually a more appropriate summary for *skewed* (asymmetrical) data patterns

Quantification of Expression

For each spot on the slide, calculate

Red intensity = Rfg - Rbg

(fg = foreground, bg = background) and

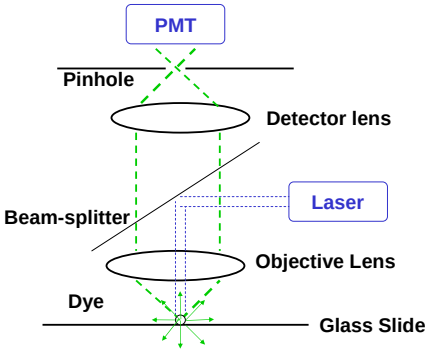
Green intensity = Gfg - Gbg

and combine them in the log (base 2) ratio

$$\mathbf{Log_2(Red/Green)}$$

Often, fg = mean and bg = median of relevant pixel intensities

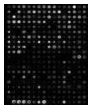
Scanner



Scanner Process

Laser → Dye → Photons (excitation) → PMT → Electrons (amplification) → A/D Convertor → Signal (Filtering, Time-space averaging)

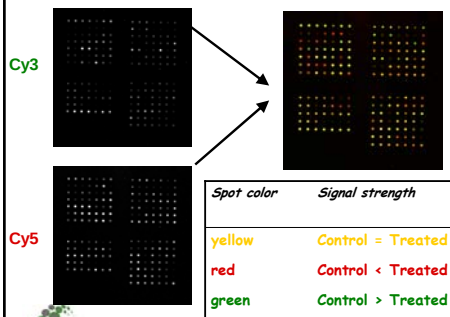
Images from Scanner



- Resolution
  - standard 10μm [currently, best ~ 5μm]
  - 100μm spot on chip = 10 pixels in diameter
- Image format
  - TIFF (tagged image file format) 16 bit (65,536 levels of gray) – also other formats
  - 1cm x 1cm image at 16 bit = 2Mb
- Separate image for each fluorescent sample
  - channel 1, channel 2, *etc.*

Images : examples

Pseudo-color overlay

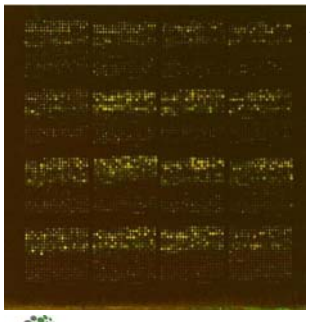


Spot color	*Signal strength*	*Gene expression*
yellow	Control = Treated	unchanged
red	Control < Treated	induced
green	Control > Treated	repressed

Steps in Images Processing

- *Addressing* (or *Gridding*)
  - Assigning coordinates to each spot
- *Segmentation*
  - *Classification of pixels* as either foreground (signal) or background
- *Information Extraction*
  - Foreground fluorescence intensity pairs (*R*,*G*)
  - Background intensities
  - Quality measures

Addressing

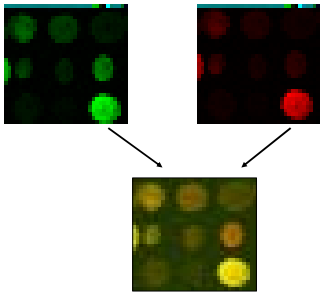


This is the process of assigning coordinates to each of the spots.

Automating this part of the procedure permits high throughput analysis.

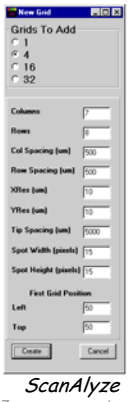
4 by 4 grids
19 by 21 spots per grid

Addressing — Registration



Addressing (I)

- Basic structure of images known (determined by the arrayer)
- Parameters to address spot positions
  - Separation between rows and columns of grids
  - Individual translation of grids
  - Separation between rows and columns of spots within each grid
  - Small individual translation of spots
  - Overall position of the array in the image



ScanAlyze

Steps in Images Processing

- *Addressing* (or *Gridding*)
  - Assigning coordinates to each spot
- *Segmentation*
  - *Classification of pixels* as either foreground (signal) or background
- *Information Extraction*
  - Foreground fluorescence intensity pairs (*R*,*G*)
  - Background intensities
  - Quality measures

Segmentation

- Classification of pixels as foreground or background
  → fluorescence intensities are calculated for each spot as measure of transcript abundance
- Production of a *spot mask:* set of foreground pixels for each spot

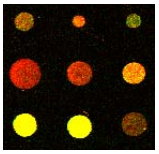
Segmentation methods in some programs

Fixed circle	ScanAlyze, GenePix, QuantArray
Adaptive circle	GenePix, Dapple, SignalViewer (uses ellipse)
Adaptive shape	Spot, region growing and watershed
Histogram	ImaGene, QuantArray, DeArray and adaptive thresholding

Fixed circle segmentation

- Fits a circle with a *constant diameter* to all spots in the image
- Easy to implement
- The spots should be of the same shape and size

May not be good for this example

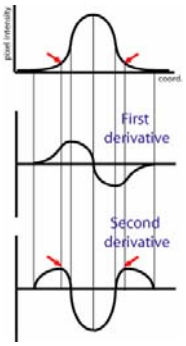


Adaptive circle segmentation

- The circle diameter is estimated *separately* for each spot

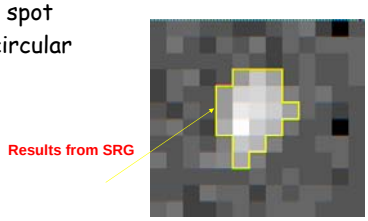
Dapple finds spots by detecting edges of spots (second derivative)

- *Problematic* if spot exhibits oval shapes

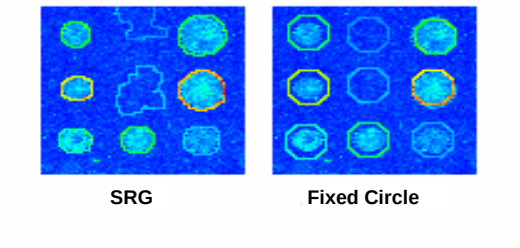


Limitation of circular segmentation

- Small spot
- Not circular



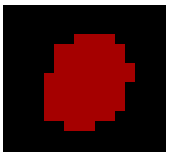
Limitation of fixed circles



SRG Fixed Circle

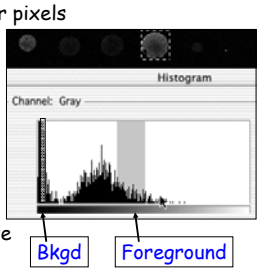
Adaptive shape segmentation

- Specification of *starting points* or *seeds*
  - Bonus: already know geometry of array
- Regions grow outwards from the seed points preferentially according to the difference between a pixel's value and the running mean of values in an adjoining region



Histogram segmentation

- Choose target mask larger than any spot
- Fg and bg intensities determined from the histogram of pixel values for pixels within the masked area
- Example : QuantArray
  - Background : mean between 5th and 20th percentile
  - Foreground : mean between 80th and 95th percentile
- May not work well when a large target mask is set to compensate for variation in spot size

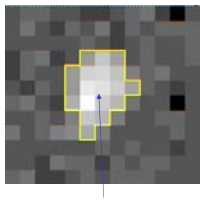


Steps in Images Processing

- *Addressing* (or *Gridding*)
  - Assigning coordinates to each spot
- *Segmentation*
  - *Classification of pixels* as either foreground (signal) or background
- *Information Extraction*
  - Foreground fluorescence intensity pairs (*R*,*G*)
  - Background intensities
  - Quality measures

Information Extraction

- *Spot Intensities*
  - mean of pixel intensities
  - median of pixel intensities
  - Pixel variation (e.g. IQR)
- *Background values*
  - None
  - Local
  - Constant (global)
  - Morphological opening
- *Quality Information*



Spot 'foreground' intensity

- The total amount of hybridization for a spot is proportional to the *total fluorescence* generated by the spot
- Spot intensity = sum of pixel intensities within the spot mask
- Since later calculations are based on *ratios* between Cy5 and Cy3, we compute the average* pixel value over the spot mask

**alternative* : ratios of medians may be better than means if bright specks present

Background intensity

- The measured fluorescence intensity includes a contribution of *non-specific hybridization* and *other chemicals* on the glass
- Fluorescence from regions not occupied by DNA should be *different* from regions occupied by DNA
  → one solution is to use *local negative controls* (spotted DNA that should not hybridize)

BG: None

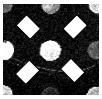
- Do not consider the background
  - Probably not accurate in many cases, but may be better than some forms of local background determination

BG: Local

- Focus on small regions surrounding the spot mask
- Median of pixel values in this region
- Most software implements such an approach


ScanAlyze


ImaGene

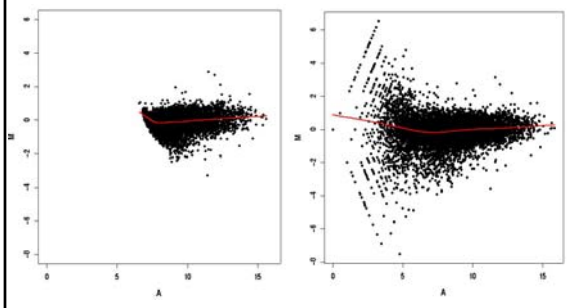

Spot, GenePix

- By ignoring pixels immediately surrounding the spots, bg estimate is *less sensitive* to the performance of the segmentation procedure

BG: Constant

- *Global method* which subtracts a constant background for all spots
- Some evidence that the binding of fluorescent dyes to 'negative control spots' is lower than the binding to the glass slide
- → More meaningful to estimate background based on a *set of negative control spots*
  - If no negative control spots : approximation of the average background = third percentile of all the spot foreground values

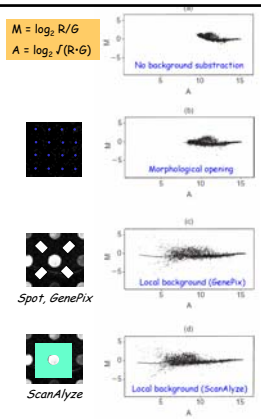
Background matters

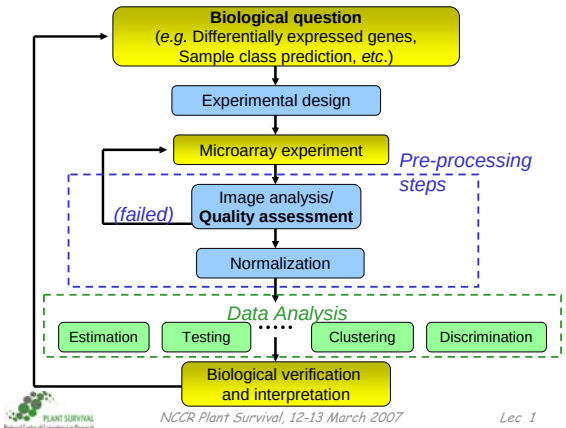


From Spot From GenePix

Summary

- Choice of bg correction method often has larger impact on the log-intensity ratios than segmentation method
- Bg adjustment has larger impact on low intensity spots





Preprocessing: Data Visualization, Exploratory Data Analysis (EDA)

- Was the experiment a success?
- Are there any specific problems?
- What analysis tools should be used?

Microarray data preprocessing

- Primary (dual channel) microarray data are pixel-level values from the *two tiff files*
- Prior to analysis, usually microarray data are *preprocessed* so that there is a *single value* for each spot
- The main preprocessing steps are
  - Image analysis (we just looked at this)
  - Normalization (to come after the break)
- We assume now that image analysis has been carried out, so that pixel values are summarized as Rf, Rb, Gf, Gb

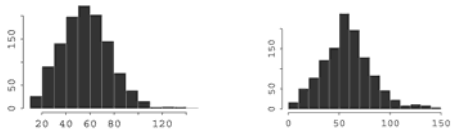
Quality Measurements

- Array
  - Correlation between spot intensities
  - Percentage of spots with no signals
  - Distribution of spot signal area
- Spot
  - Signal / Noise ratio
  - Variation in pixel intensities
  - Identification of "bad spot" (spots with no signal)
- Ratio (2 spots combined)
  - Circularity
- *Flag* or *weight* spots based on these (or other) criteria

Quality of Array

Distribution of areas

- Judge by eye
- Look at variation. (e.g. SD)



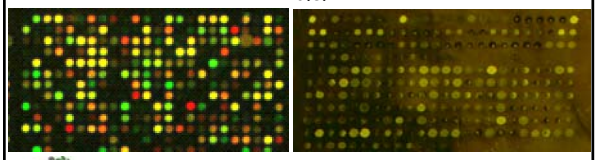
Cy3 area
- mean 57
- median 56
- SD 20.67

Cy5 area
- mean 59
- median 57
- SD 24.34

Red/Green overlay images

Co-registration and *overlay* offers a quick visualization, revealing information on color balance, uniformity of hybridization, spot uniformity, background, and artifacts such as dust or scratches

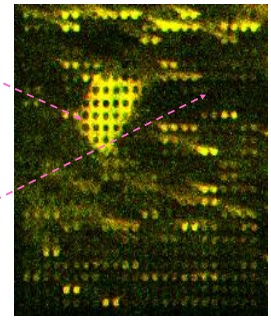
Good: low bg, detectable d.e. Bad: high bg, ghost spots, little d.e.



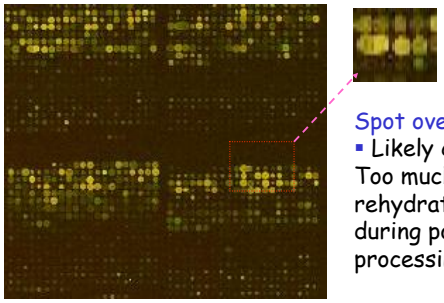
Practical Problems 1

High Background
- 2 likely causes:
  - Insufficient blocking
  - Precipitation of the labeled probe

Weak Signals



Practical Problems 2



Spot overlap
- Likely cause: Too much rehydration during post - processing

Artifacts in microarrays

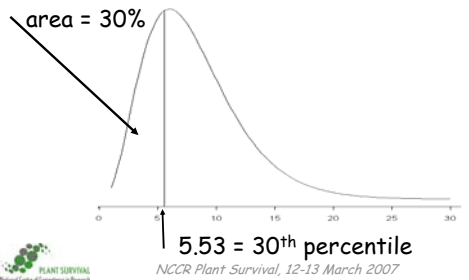
- We are interested in finding true *biologically meaningful differences* between sample types
- Due to other sources of systematic variation, there are also usually *artifactual differences*
- Sources of artifacts include:
  - print tips - differences in subarrays
  - plate effects – differences in rows within subarray
  - batch effects
  - hybridization artifacts

Looking for artifacts

- Exploratory data analysis (EDA) is an important component of microarray data preprocessing
- EDA involves identifying data artifacts
- We will use several types of plots for data visualization, primarily
  - *image plots*
  - *scatterplots*
  - *boxplots*
  - *spatial plots*

Quantiles

- The p^{th} *quantile* is the number that has the proportion p of the data values smaller than it

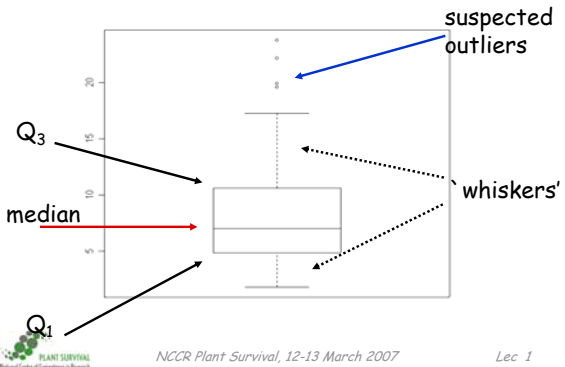


Five-number summary and boxplot

- The 25th (Q_1), 50th (median), and 75th (Q_3) percentiles divide the data into 4 equal parts; these special percentiles are called *quartiles*
- An overall summary of the distribution of variable values is given by the five values:

  Min, Q_1, Median, Q_3, and Max
- A *boxplot* provides a visual summary of this five-number summary

Sample boxplot

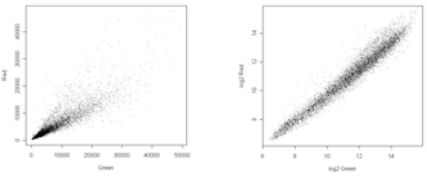


Scatterplots

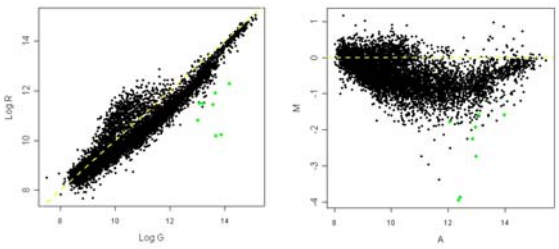
- MA plot (also sometimes called MVA plot, RI plot, SI plot):

  A = *average* of $\log_2$(Red) and $\log_2$(Green)

  M = difference of $\log_2$(Red) and $\log_2$(Green)
- Why we take logs:



Scatterplots: always log*, always rotate



$\log_2 R$ vs $\log_2 G$ $M=\log_2 R/G$ vs $A=\log_2 \sqrt{RG}$

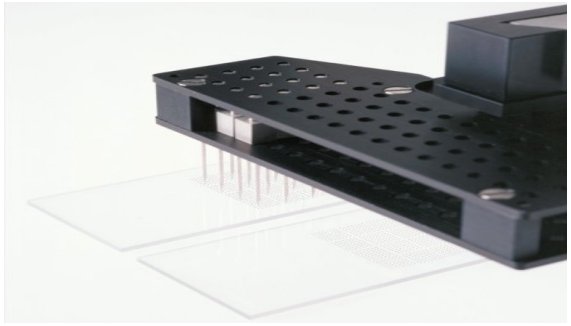
* Other transformations may provide improvement

Building the array

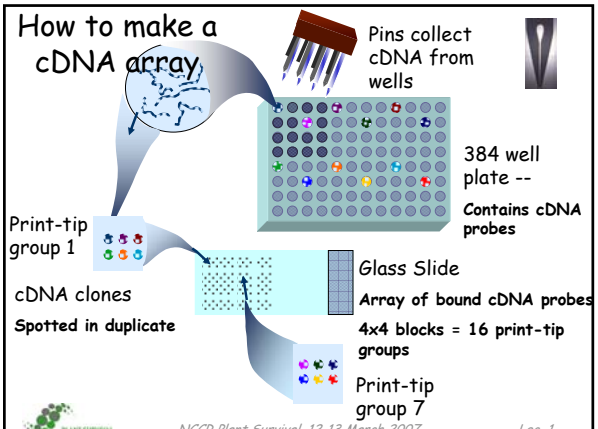
Ngai Lab arrayer , UC Berkeley

Print-tip head

Print-tip head *(lego)*



How to make a cDNA array



(*BREAK*)