cDNA Microarray

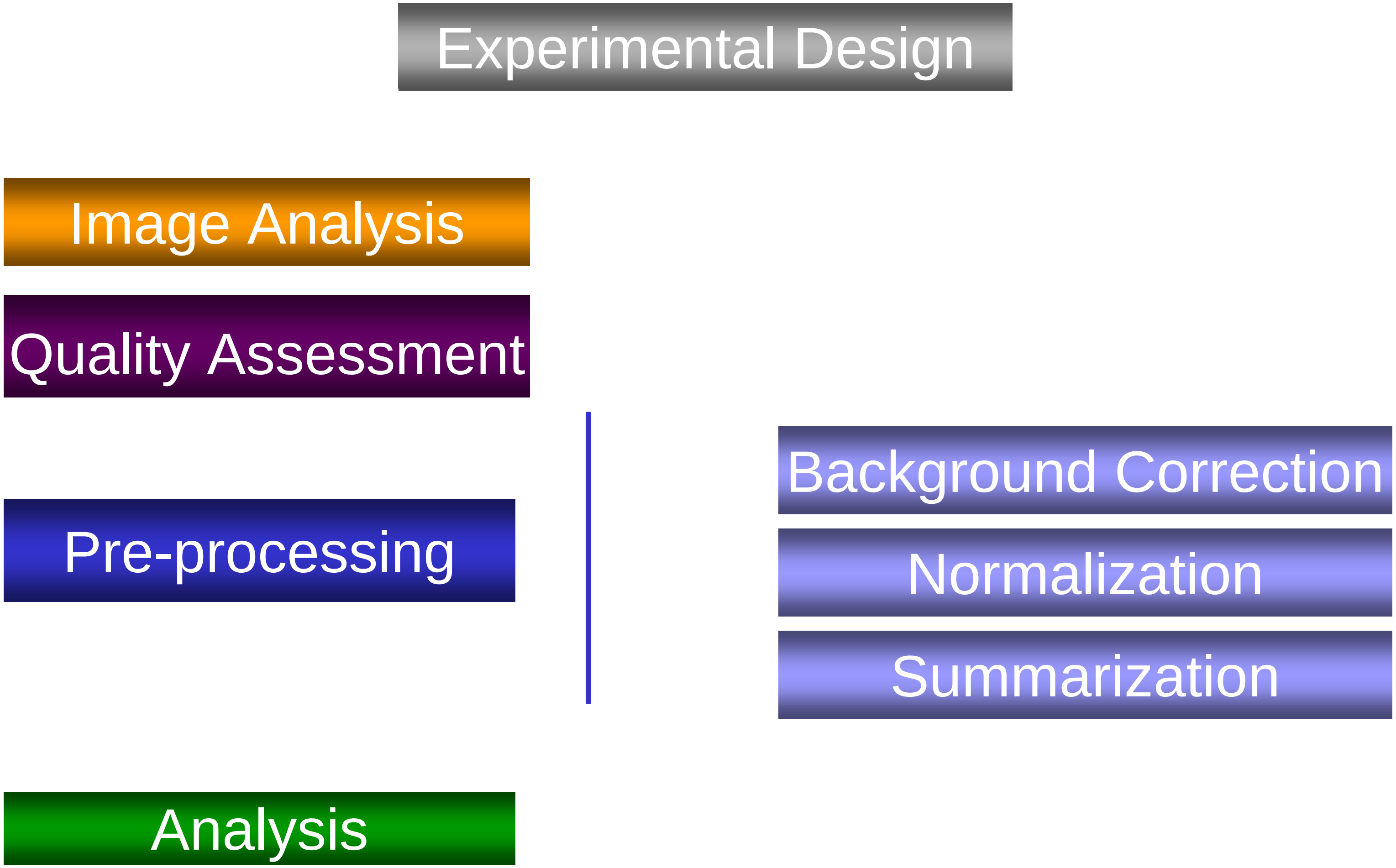
Data Analysis

with BioConductor packages

Nolwenn Le Meur

October MiniCourse

Microarrays Experiment



Experimental Design

Image Analysis

Quality Assessment

Pre-processing

Background Correction

Normalization

Summarization

Analysis

Outline

- **Data acquisition & Pre-processing (chap. 4)**
  - Image analysis
  - Quality assessment
  - Pre-processing

- **Lab : case studies (chap 4)**
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)

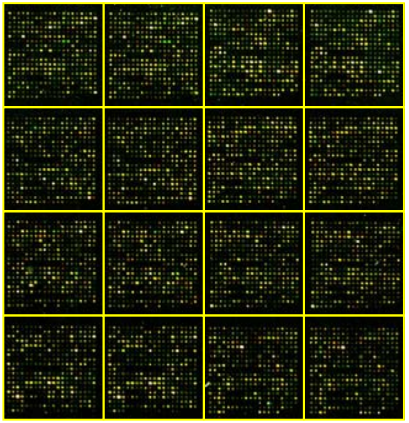
Terminology

- **Target:** DNA hybridized to the array, mobile substrate.
- **Probe:** DNA spotted on the array (spot).
- **print-tip-group :** collection of spots printed using the same print-tip (or pin), aka. grid.

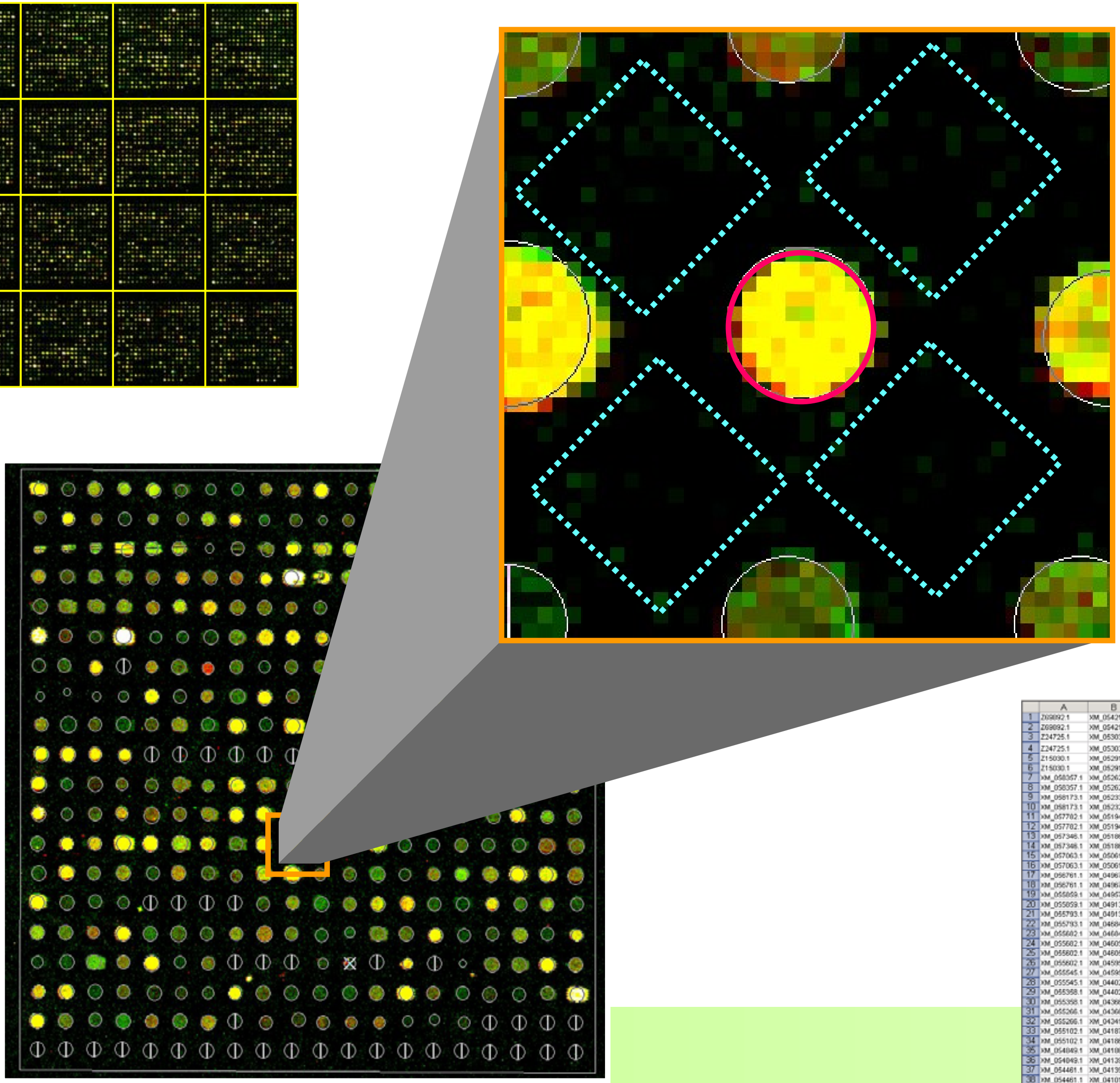
- **G, Gb:** Cy 3 signal and background intensities
- **R, Rb:** Cy5 signal and background intensities
- **M** = log2(R) - log2(G)
- **A** = 1/2(log2(R) + log2(G))

Image Analysis

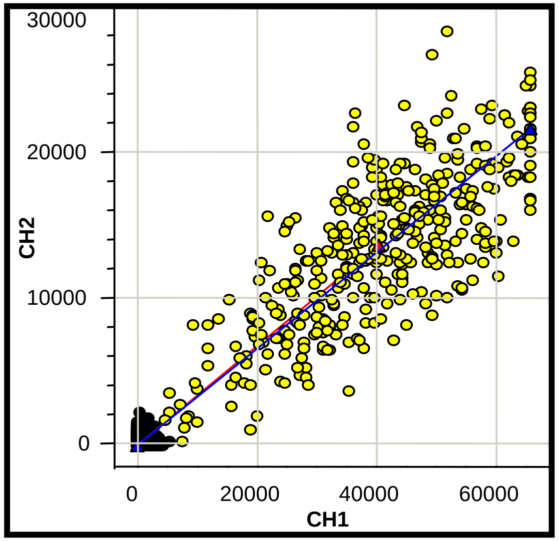
1. Location

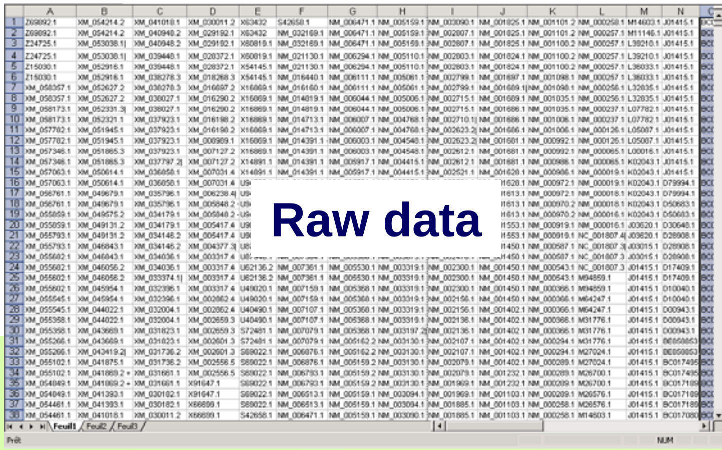


2. Segmentation

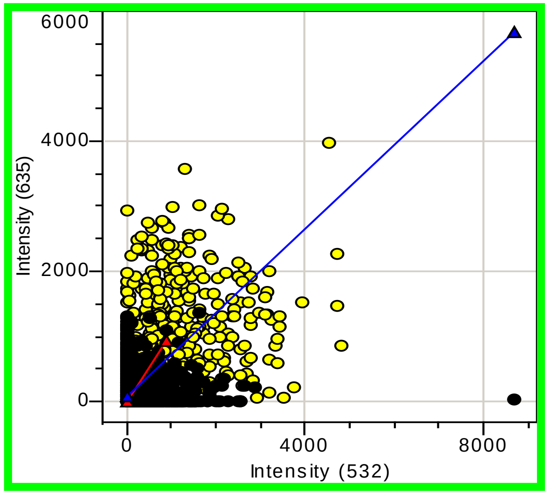


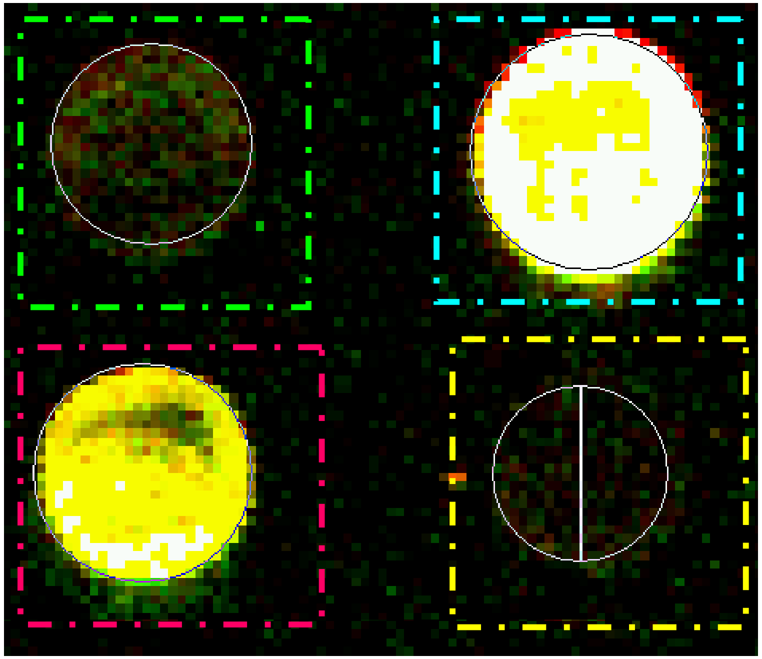
3. Quantification

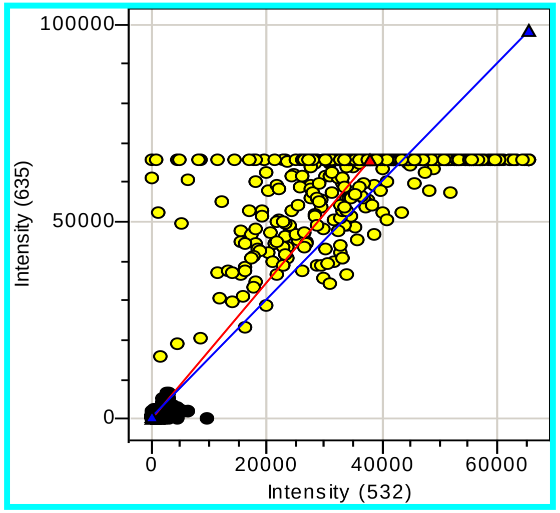


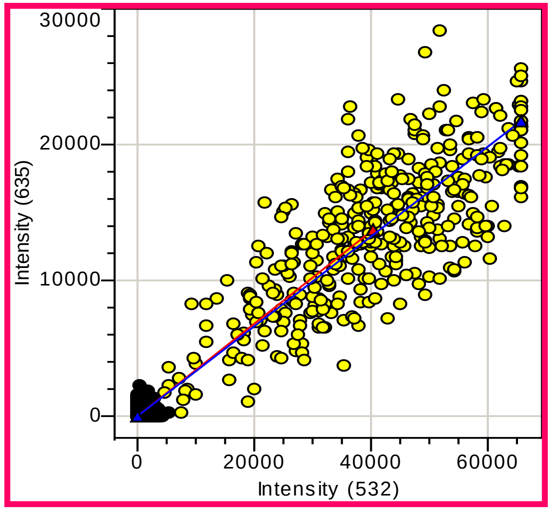


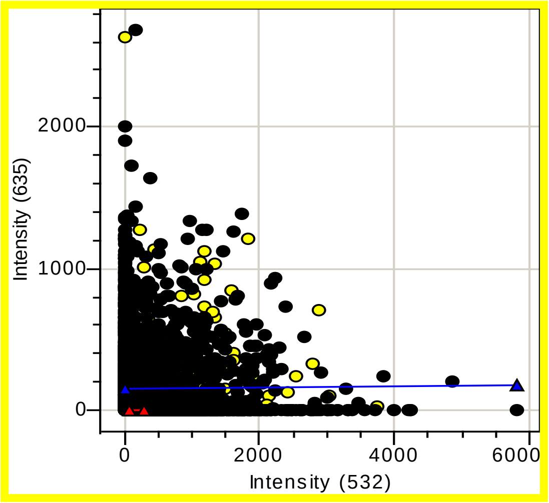
Quality Filtering











● Background

● Foreground

Quality Assessment

For at the probe-level:

- **Sources**
  - faulty printing, uneven distribution, contamination with debris, magnitude of signal relative to noise, poorly measured spots
- **Spot quality**
  - *Brightness:* foreground/background ratio
  - *Uniformity:* variation in pixel intensities and ratios of intensities within a spot
  - *Morphology:* area, perimeter, circularity
  - *Spot Size:* number of foreground pixels
- **Action**
  - use weights for measurements to indicate reliability in later analysis.
  - set measurements to NA (missing values)

Quality Assessment

For each array

- **Problems**
  - array fabrication defect
  - problem with RNA extraction
  - failed labeling reaction
  - poor hybridization conditions
  - faulty scanner
- **Quality measures**
  - Percentage of spots with no signal (~30% exlcuded spots)
  - Range of intensities
  - (Av. Foreground)/(Av. Background) > 3 in both channels
  - Distribution of spot signal area

Quality Assessment

For each array:

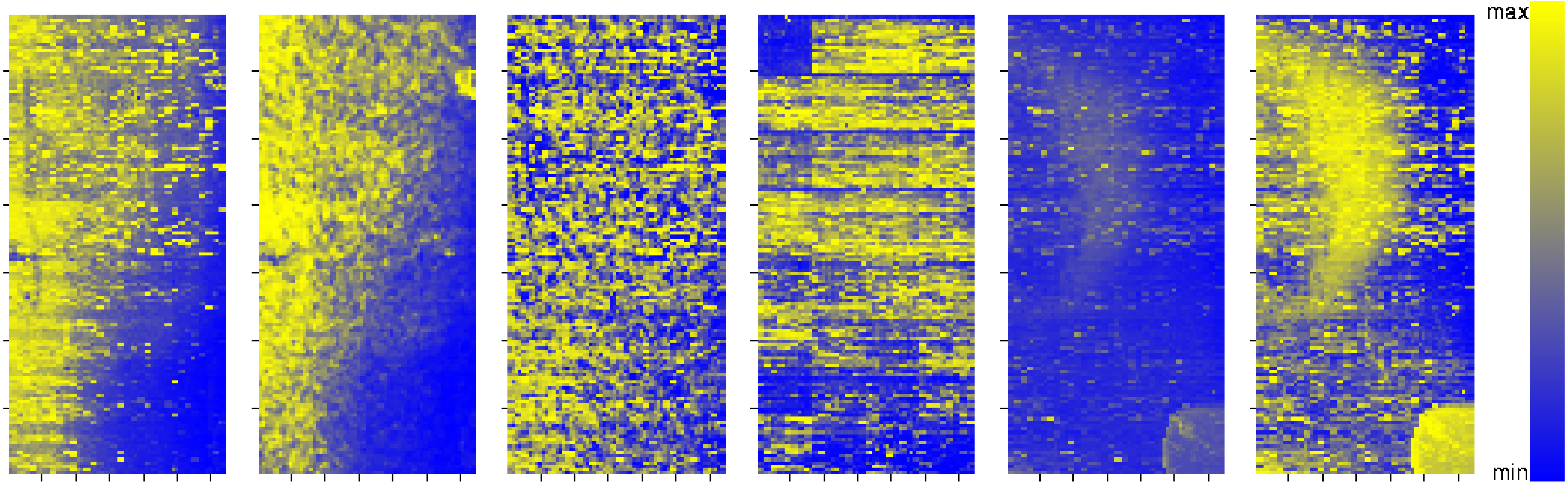
- **Visual inspection**
  - hairs, dust, scratches, air bubbles, dark regions, regions with haze
- **Diagnostics plots** of spot statistics

e.g. R and G log-intensities, M, A, spot area.

- – 2D spatial images;
- – ECDF plots;
- – Boxplots;
- – Scatter-plots;
- – Density plots.

- **Stratify** plots according to layout parameters, *e.g.* print-tip-group, plate.

Spatial Effects – Image Plots



R Rb R-Rb

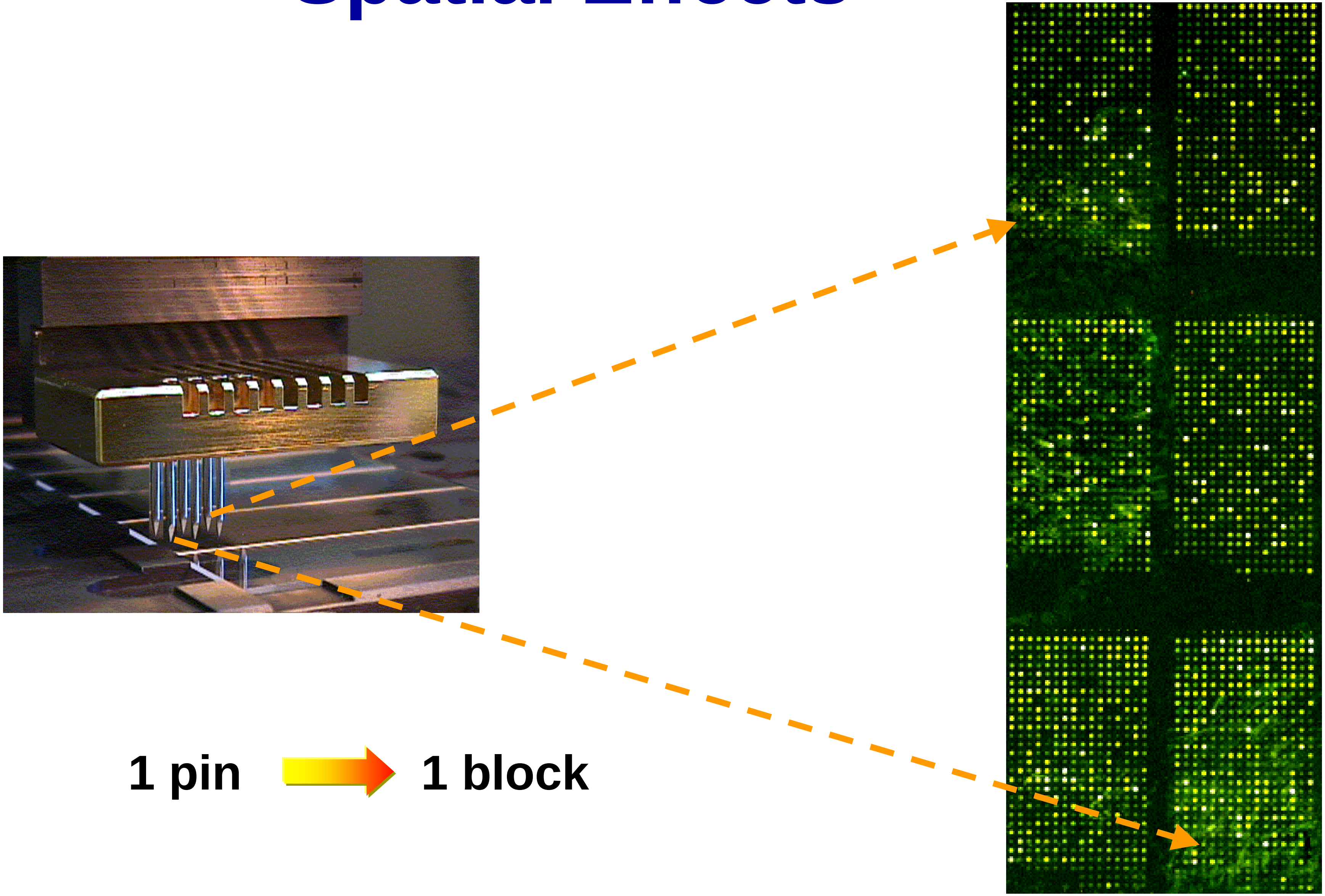
color scale by rank

Print-tip

Air buble

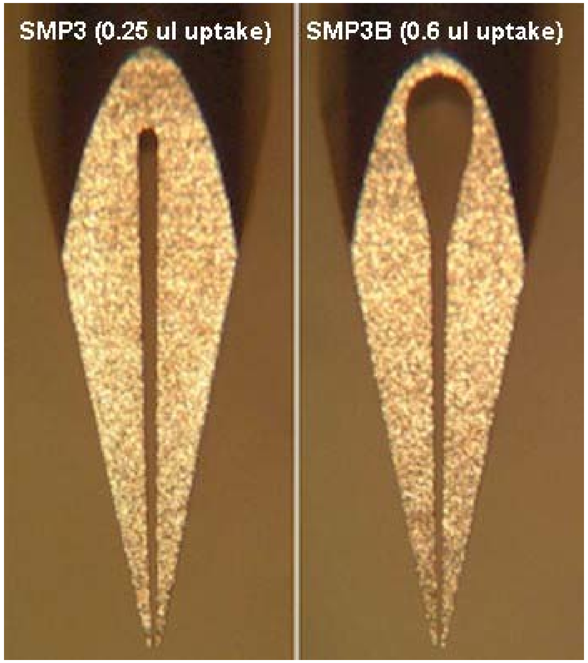
washing

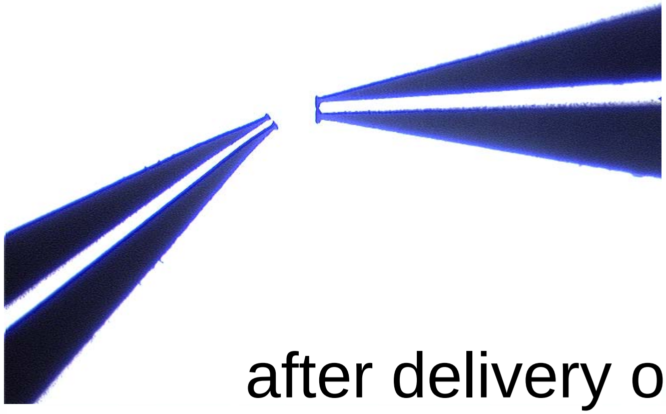
Spatial Effects



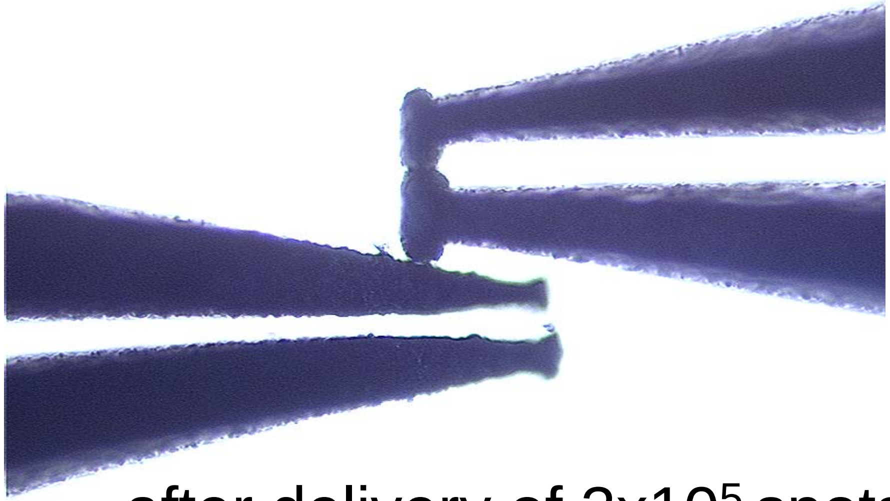
1 pin → 1 block

Spotting Pin Quality Decline





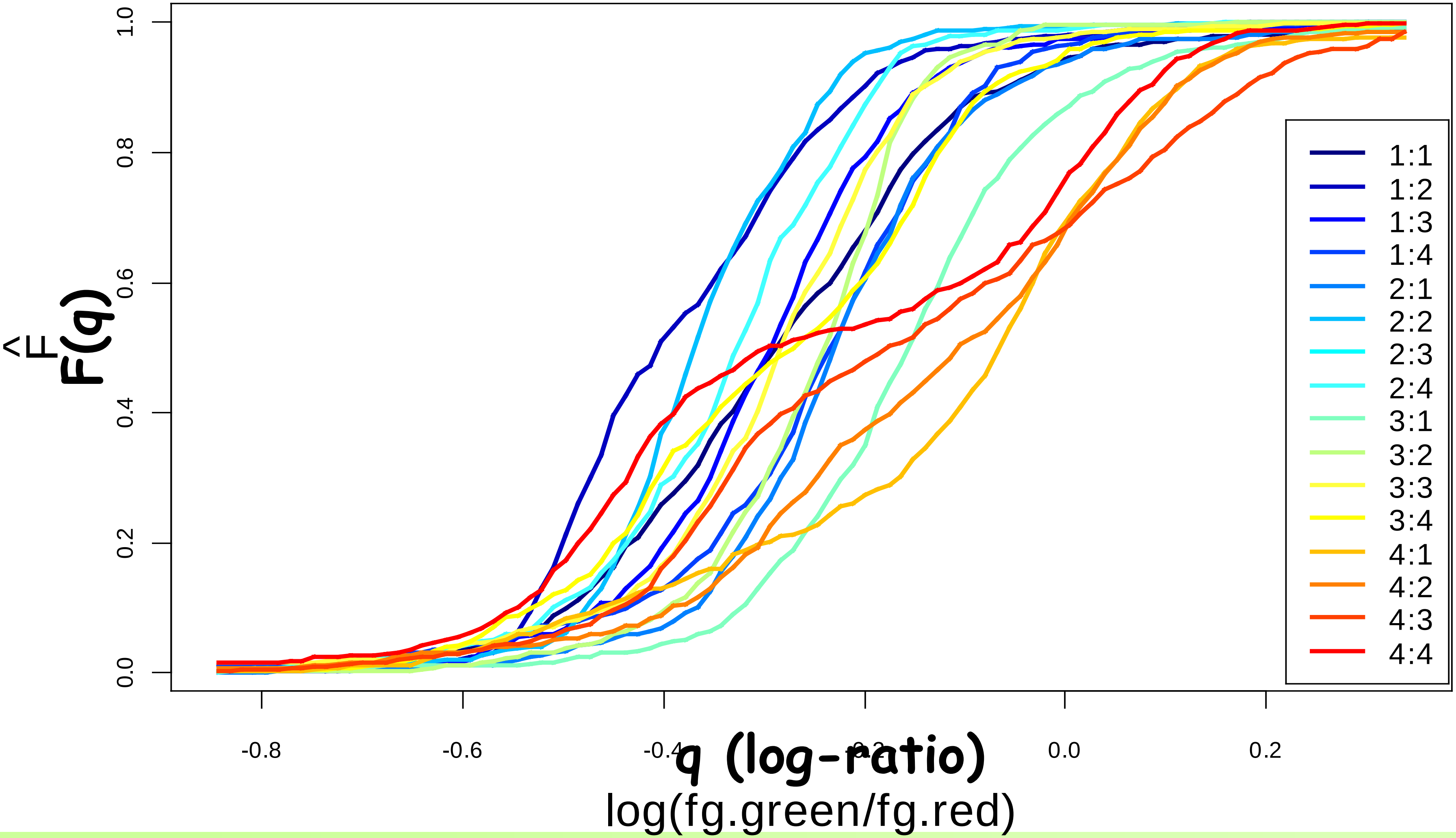
after delivery of $5\text{x}10^5$ spots



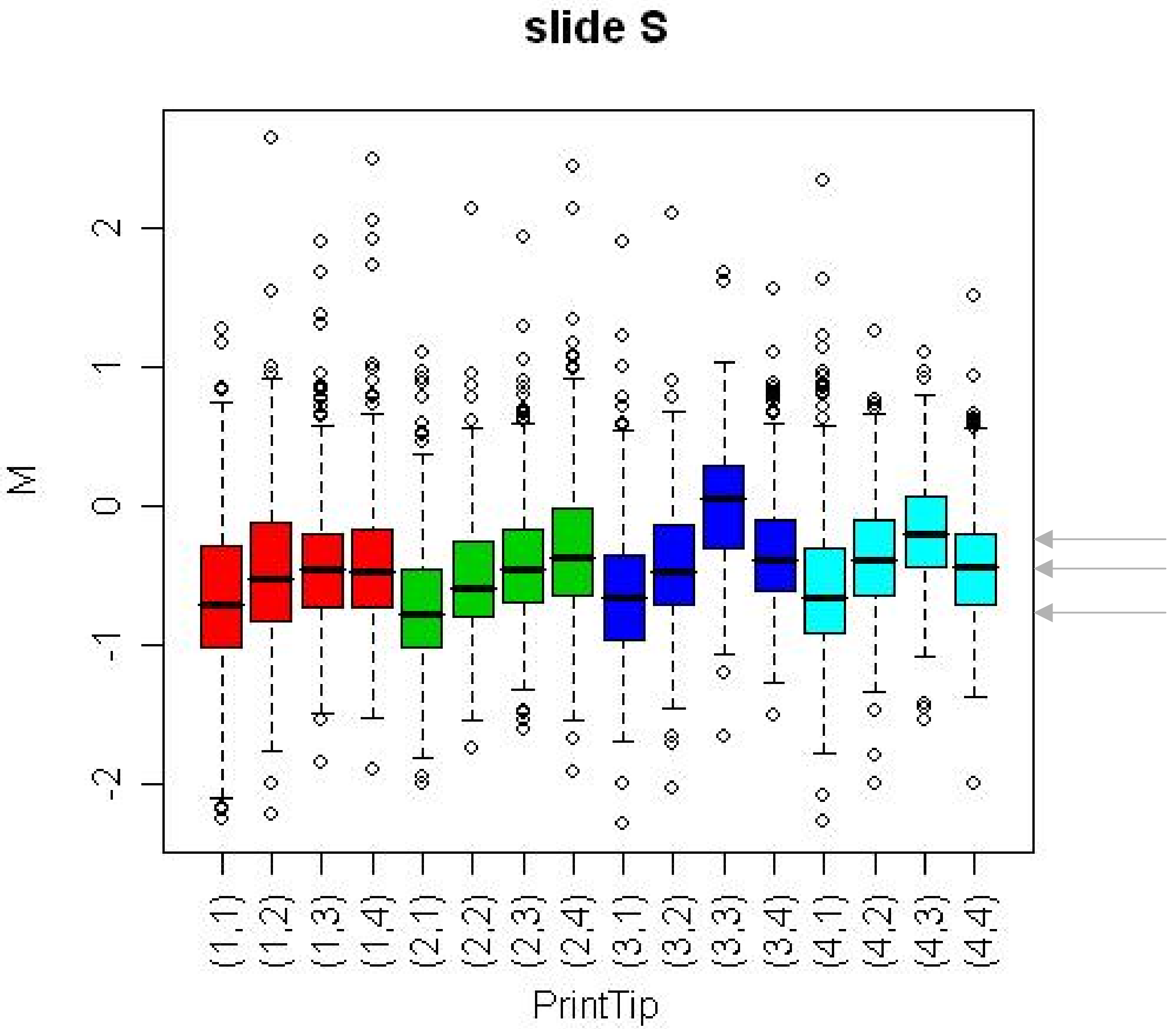
after delivery of $3\text{x}10^5$ spots

H. Sueltmann DKFZ/MGA

Print-tip Effects – ECDF plot



Print-tip Effects - Boxplots

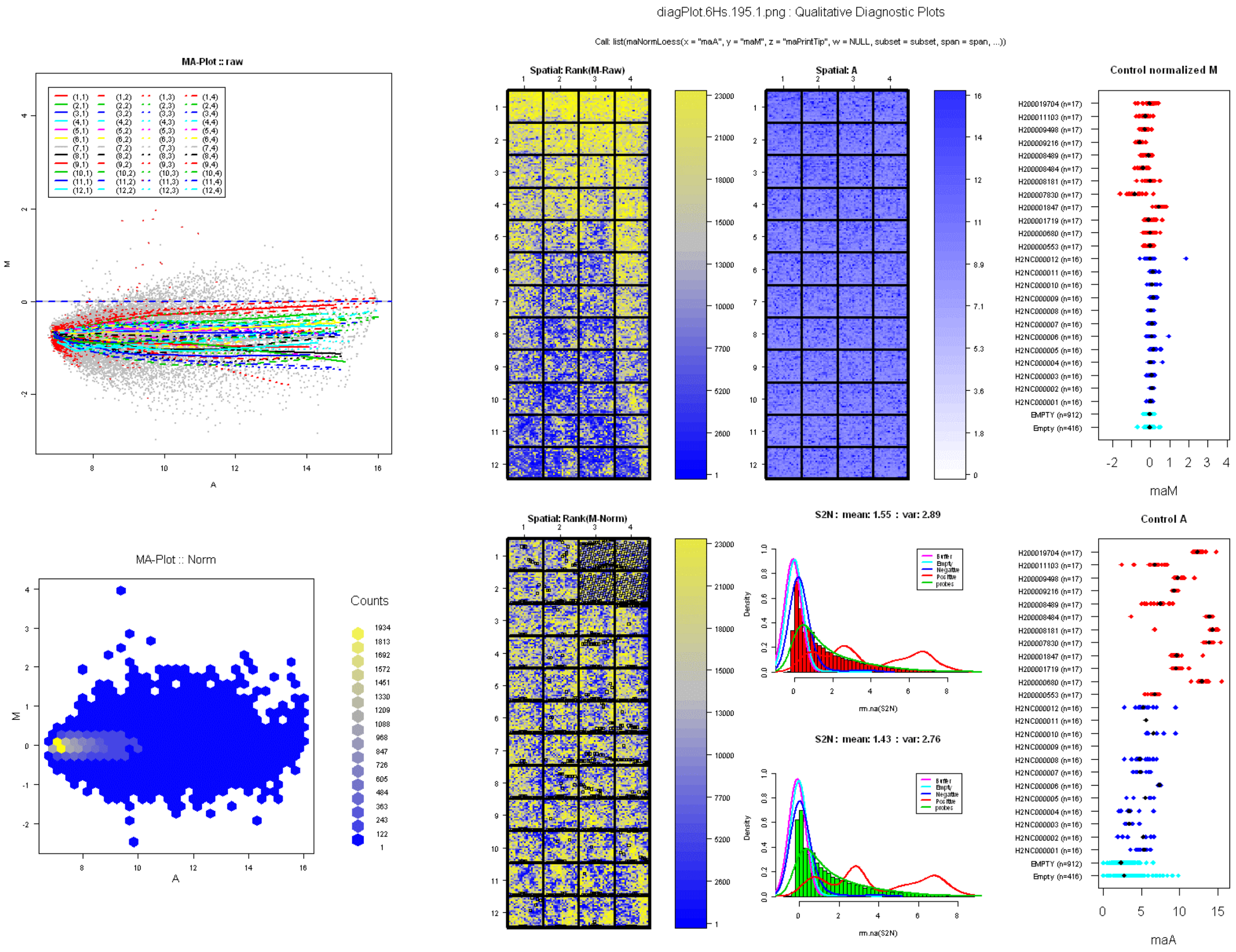


Q3 75% quantile

median

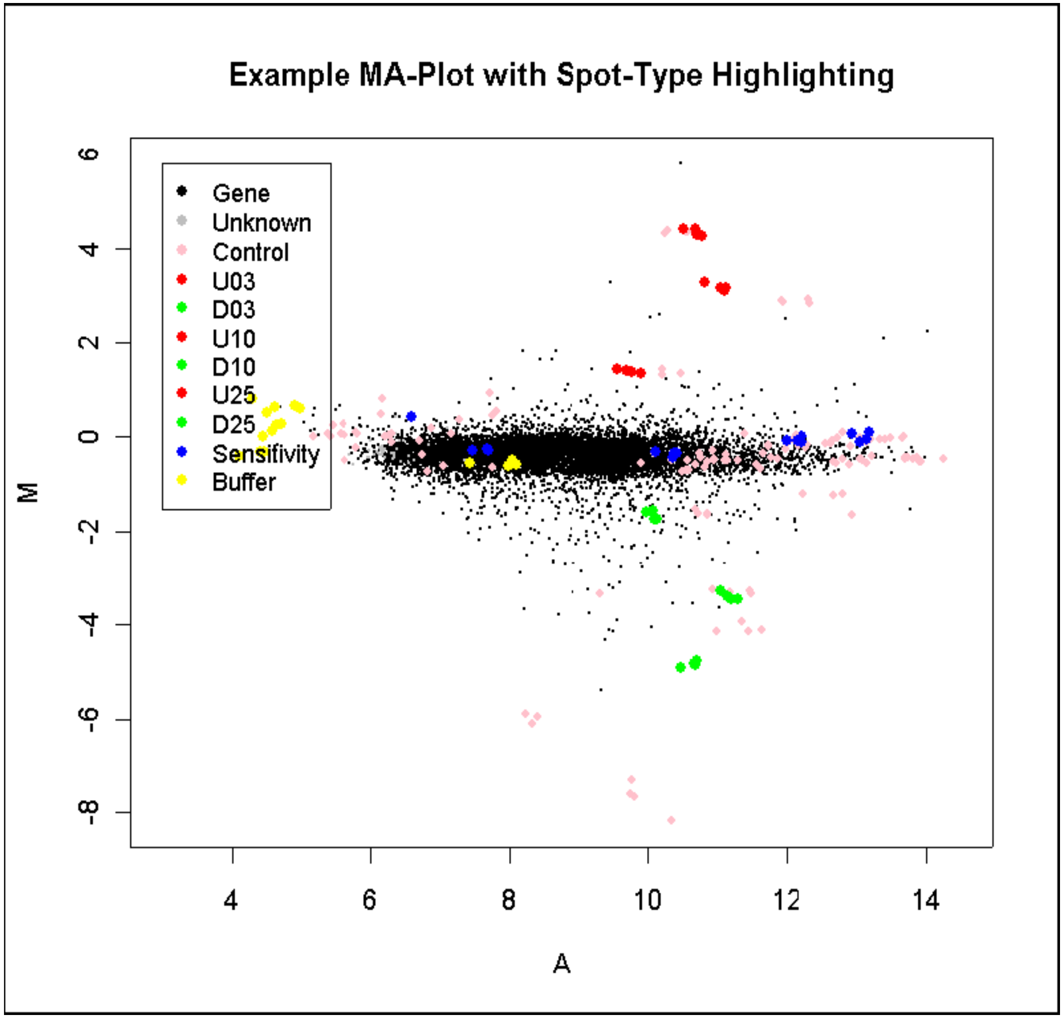
Q1 25% quantile

Diagnostic plot with *arrayQuality*


diagPlot.6Hs.195.1.png : Qualitative Diagnostic Plots

Call: list(maNormLoess(x = "maA", y = "maM", z = "maPrintTip", w = NULL, subset = subset, span = span, ...))

Data Exploration with *limma*



(Limma user Guide)

Quality Assessment: Summary

For each spot:

- weight

For each array:

- Diagnostics plots
- Stratify
- Controls

BioC packages:

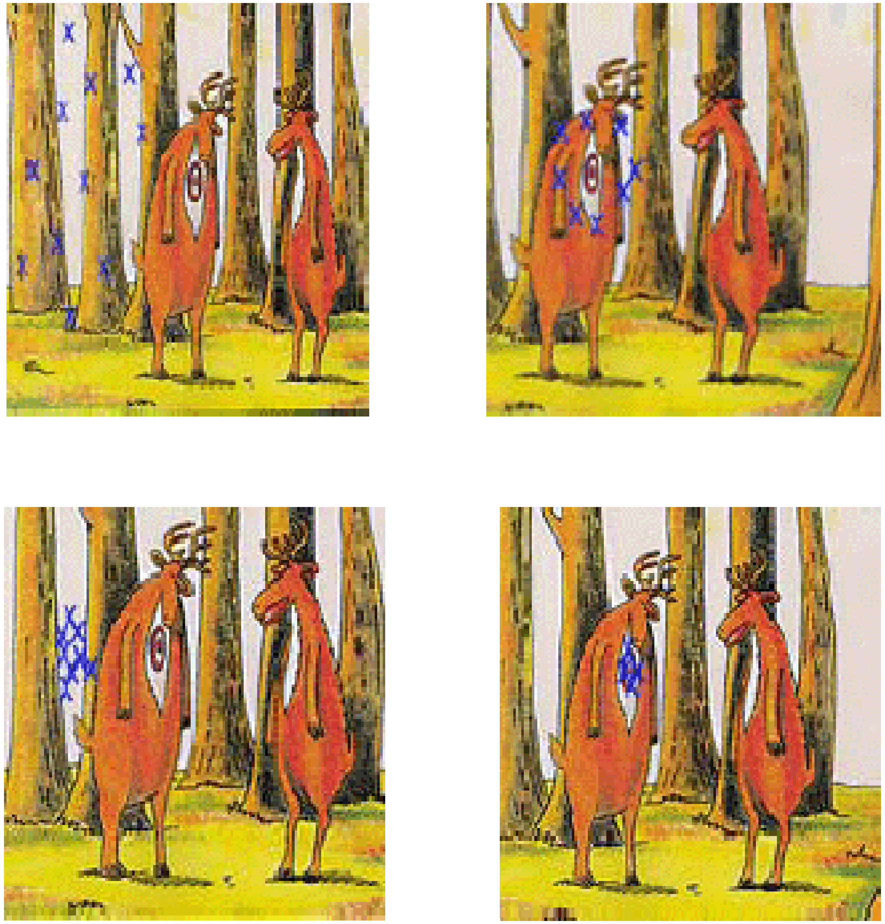
- *arrayQuality*
- *arrayMagic*
- *...*

Outline

- Data acquisition & Pre-processing (chap. 4)
  - Image analysis
  - Quality assessment
  - **Pre-processing**

- Lab : case studies (chap 4)
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)

Variance-Bias trade off



<- Bias

Accuray ->

Variance ->

<- Precision

Sources of Variation

- RNA extraction
- reverse transcription
- labeling efficiencies
- Scanner settings

Systematic

•similar effect on many measurements

•corrections can be estimated from data

Calibration

- PCR
- DNA concentration
- Printing or pin
- cross-hybridization
- …

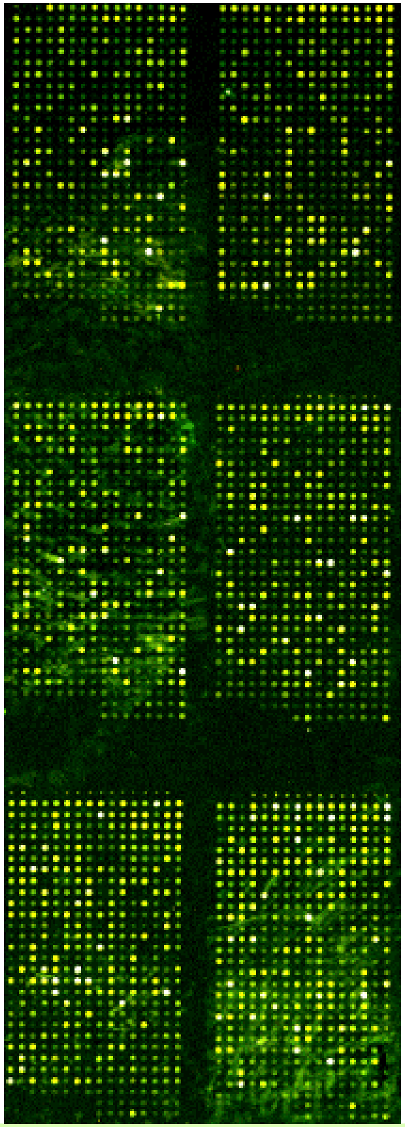
Stochastic

• too random to be explicitely accounted for

• “noise”

Error Model

Background Correction

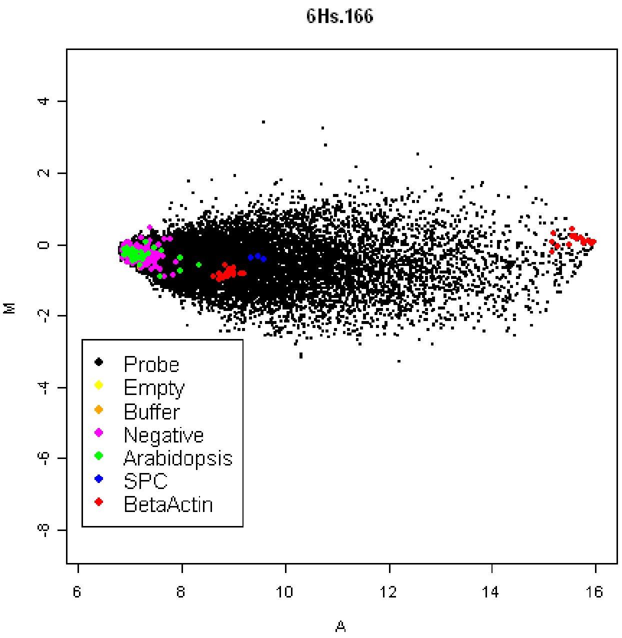


- none
- subtraction, movingmin
- *Minimun,edwards, normexp,…*

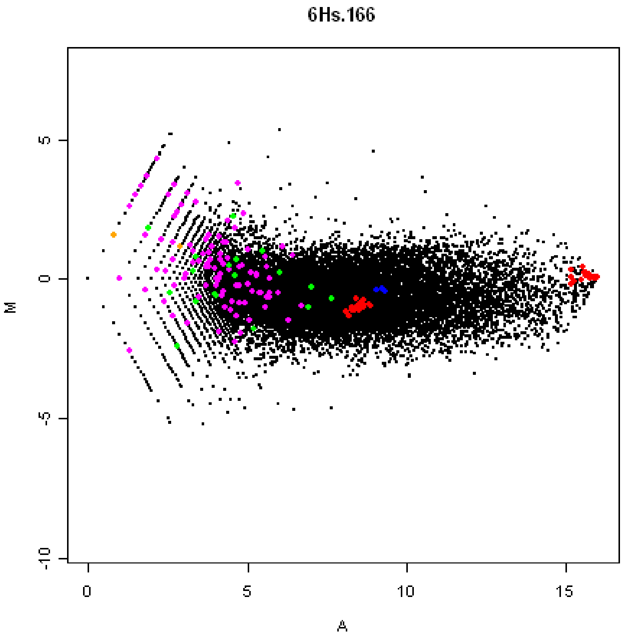
- More details … *limma*

  >?backgroundCorrect

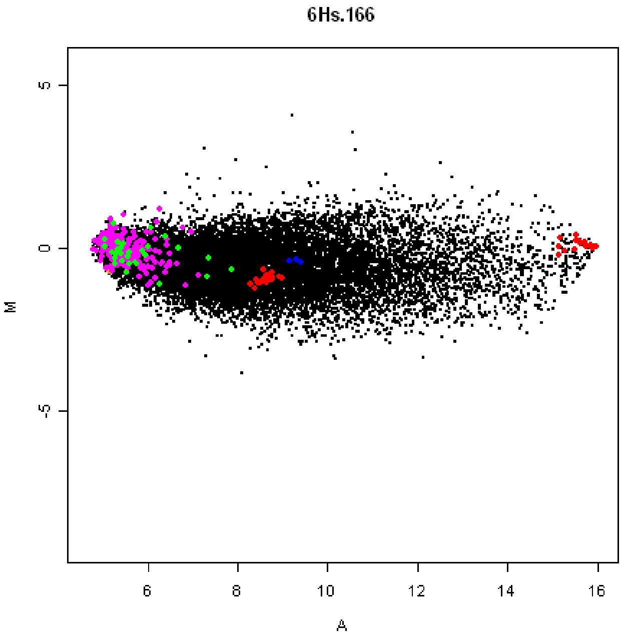
Background Correction



none

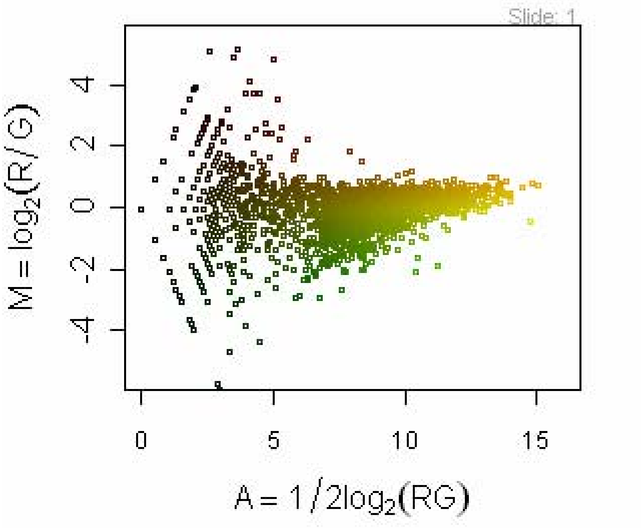


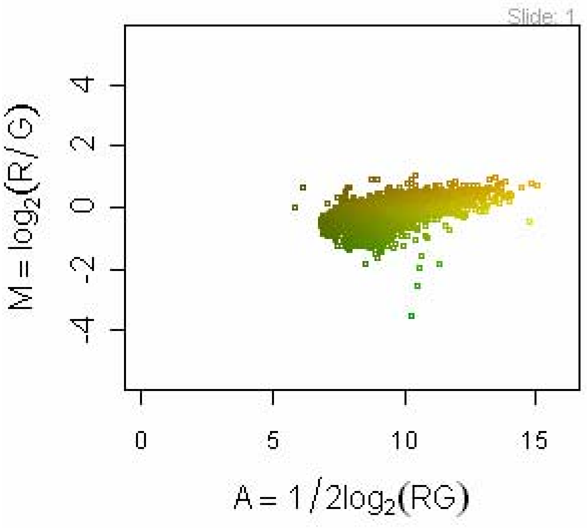
substraction

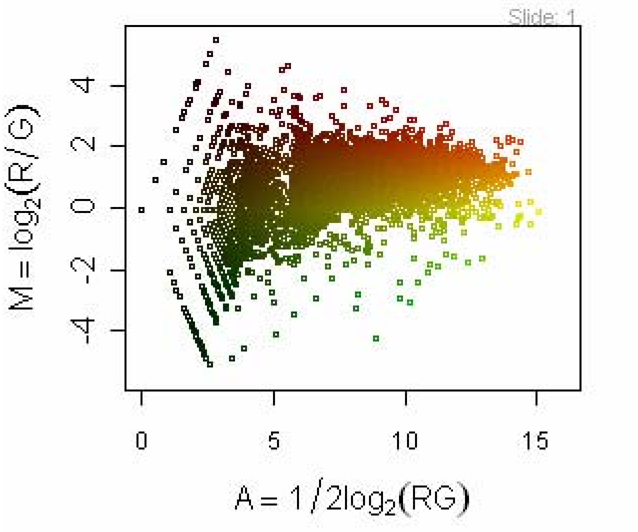


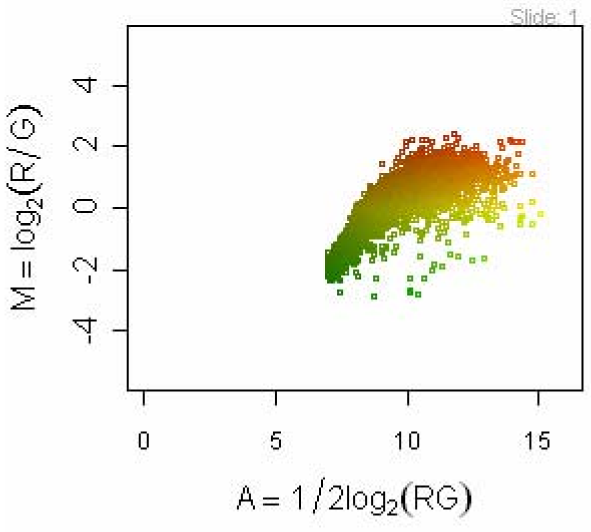
normexp

Background Correction



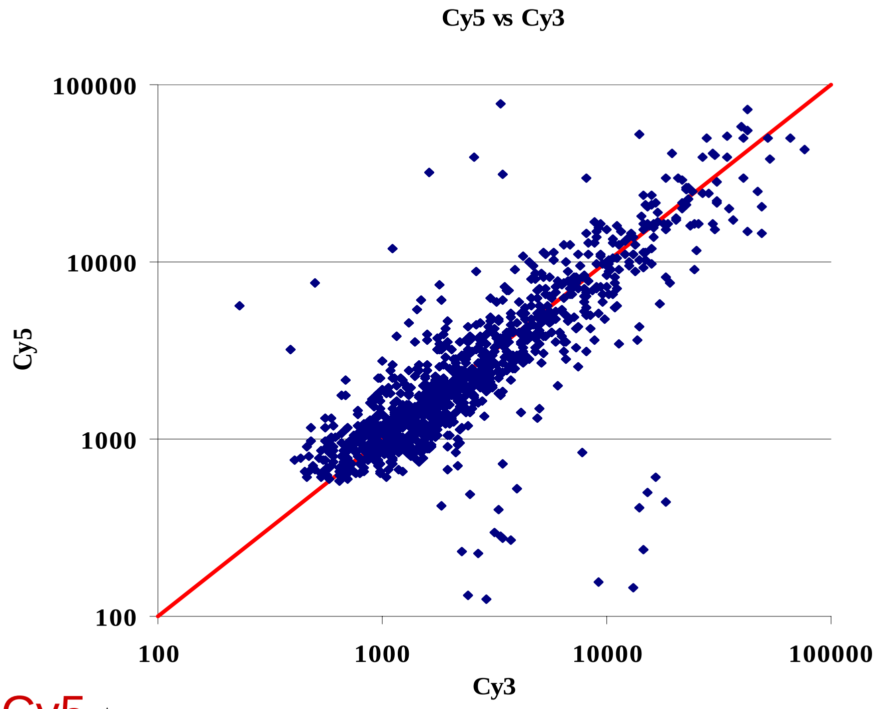


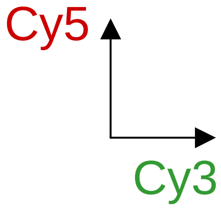




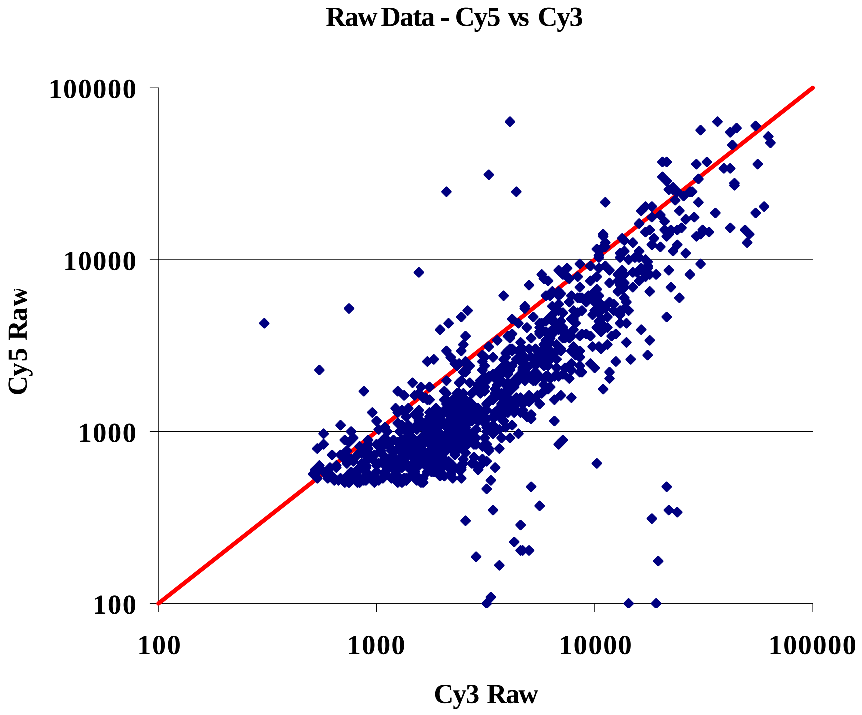
Why Normalize?

Theory





Reality



Normalization

Identify and remove the effects of systematic variation

- Normalization is closely related to quality assessment. In a ideal experiment, no normalization would be necessary, as the technical variations would have been avoided.

- Normalization is needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact.

- Normalization is necessary before any analysis which involves within or between slide comparisons of intensities, e.g., clustering, testing.

Normalization methods

- median
- loess
- 2D loess
- print-tip loess
- variance stabilisation
- …..

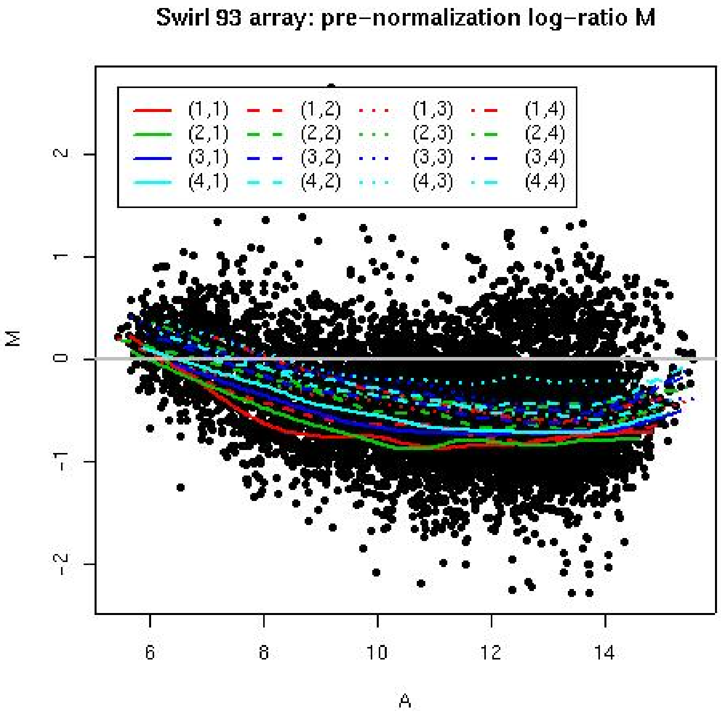
Two-channel

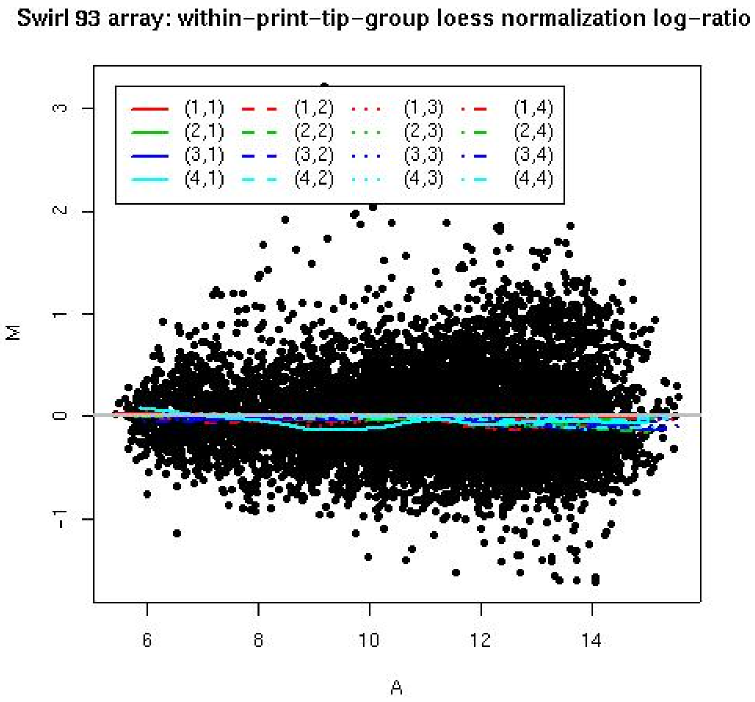
Separate-channel

Smyth, G. K., and Speed, T. P. (2003). In: *METHODS: Selecting Candidate Genes from DNA Array Screens: Application to Neuroscience*

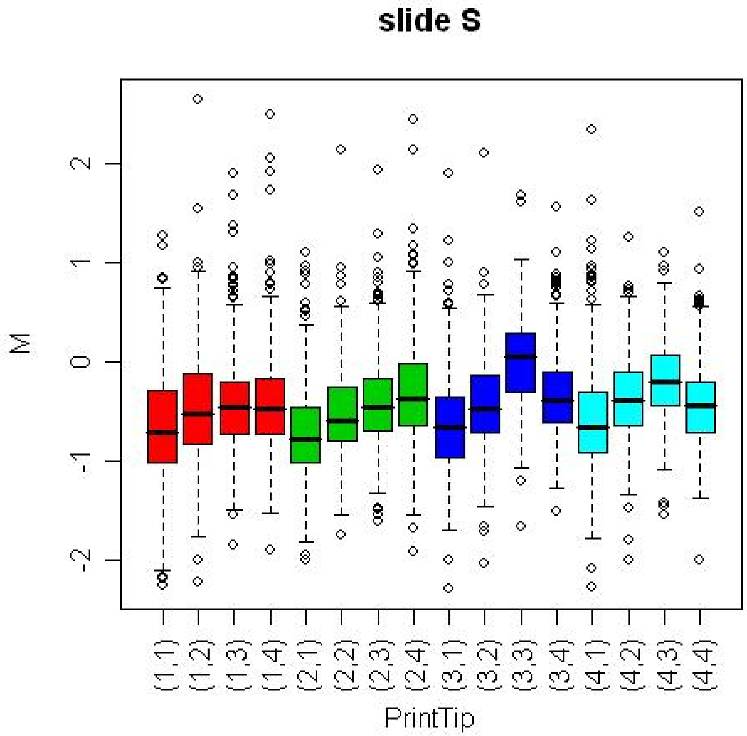
Two channel normalization

- Location: centers log-ratios around zero using A and spatial dependent bias

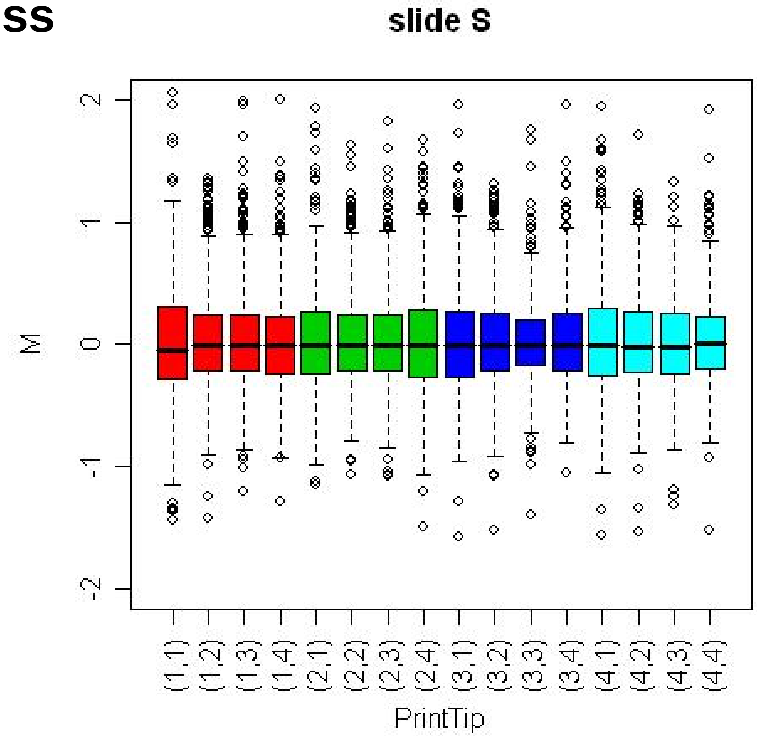




Two channels normalization



Print-tip lowess

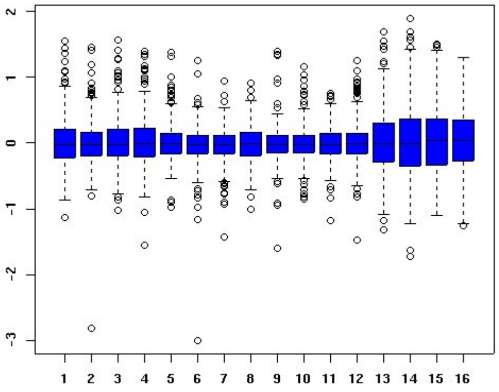


Two channels normalization

- Location: centers log-ratios around zero using A and spatial dependent bias

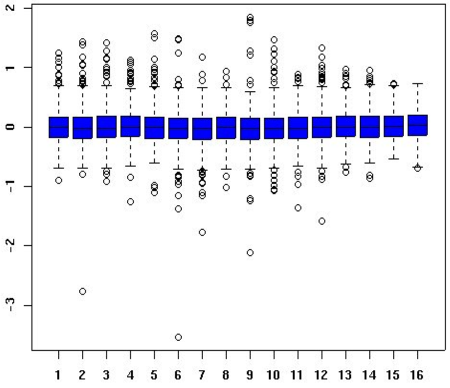
- Scale: adjust for different in scale between multiple arrays

median centered



Scaling

median centered & MAD scaled



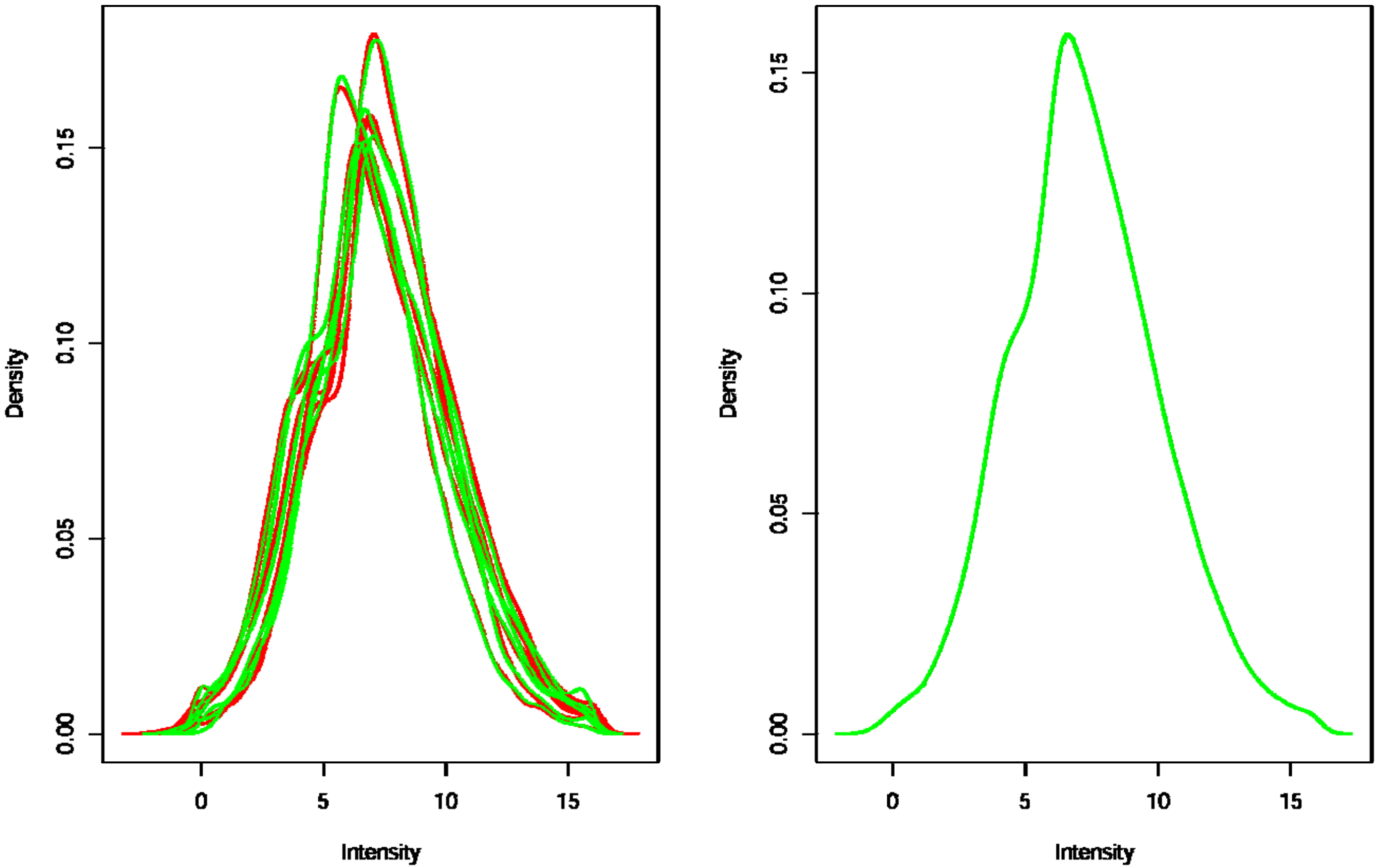
One channel normalization

- As technology improves the spot-to-spot varation is reduced
- Development of normalization techniques that work on the absolute intensities

Ex: quantile normalization (*limma*)

variance stabilization (*vsn*)

Quantile Normalization



Before

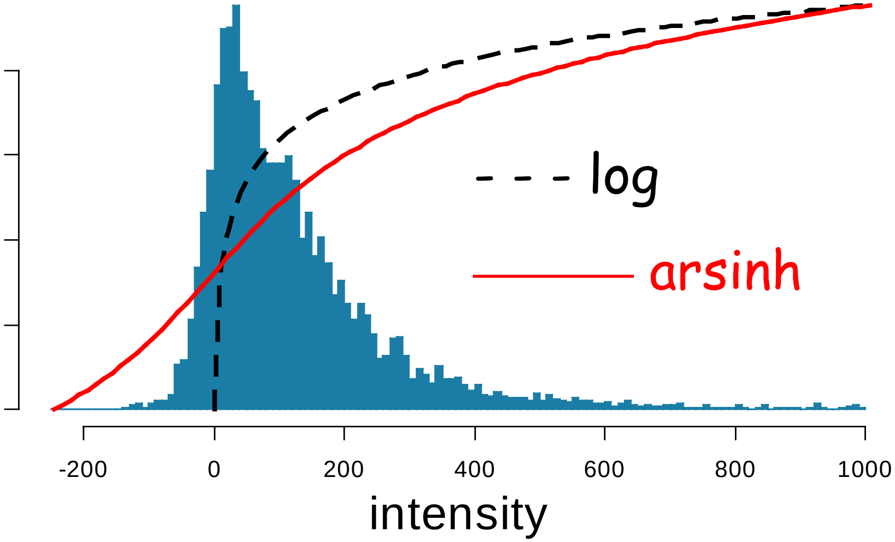
After

Bolstand *et al*.(2003)

Variance Statibilizing Transformation

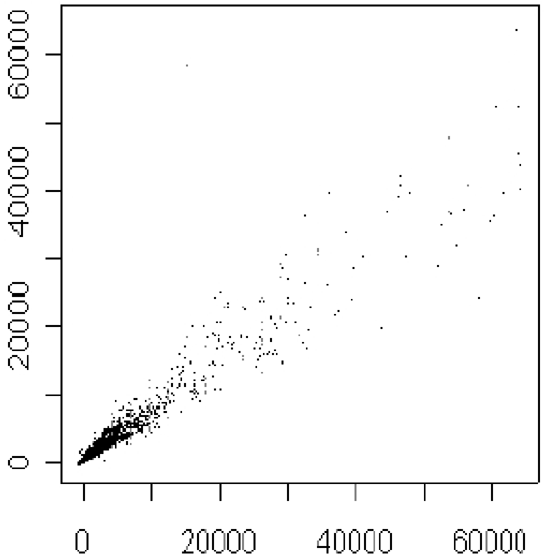
- log-transformation is replaced by a arcsinh transformation
  - Meaningful around 0
  - Original intensities may be negatives

- Estimation of transformation parameters (location, scale) based on Maximun Likelihood paradigm
- vsn–normalized data behaves close to the normal distribution

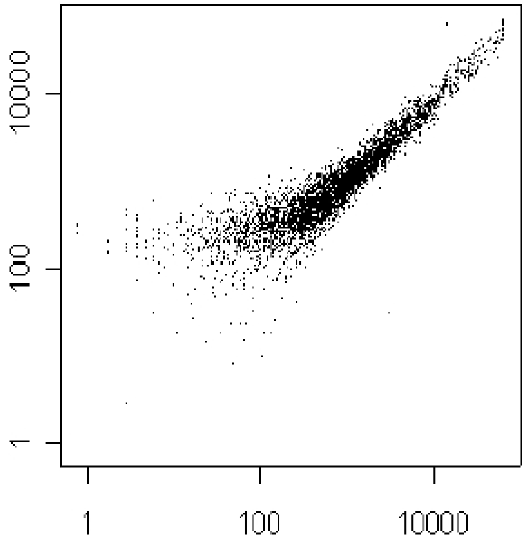


(Huber *et al.* 2004)

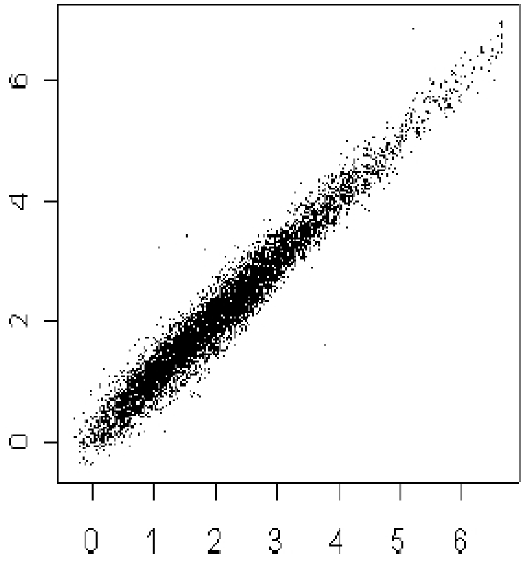
Variance stabilization (*vsn*)



linear



log



arsinh

Preprocessing : Summary

For each array:

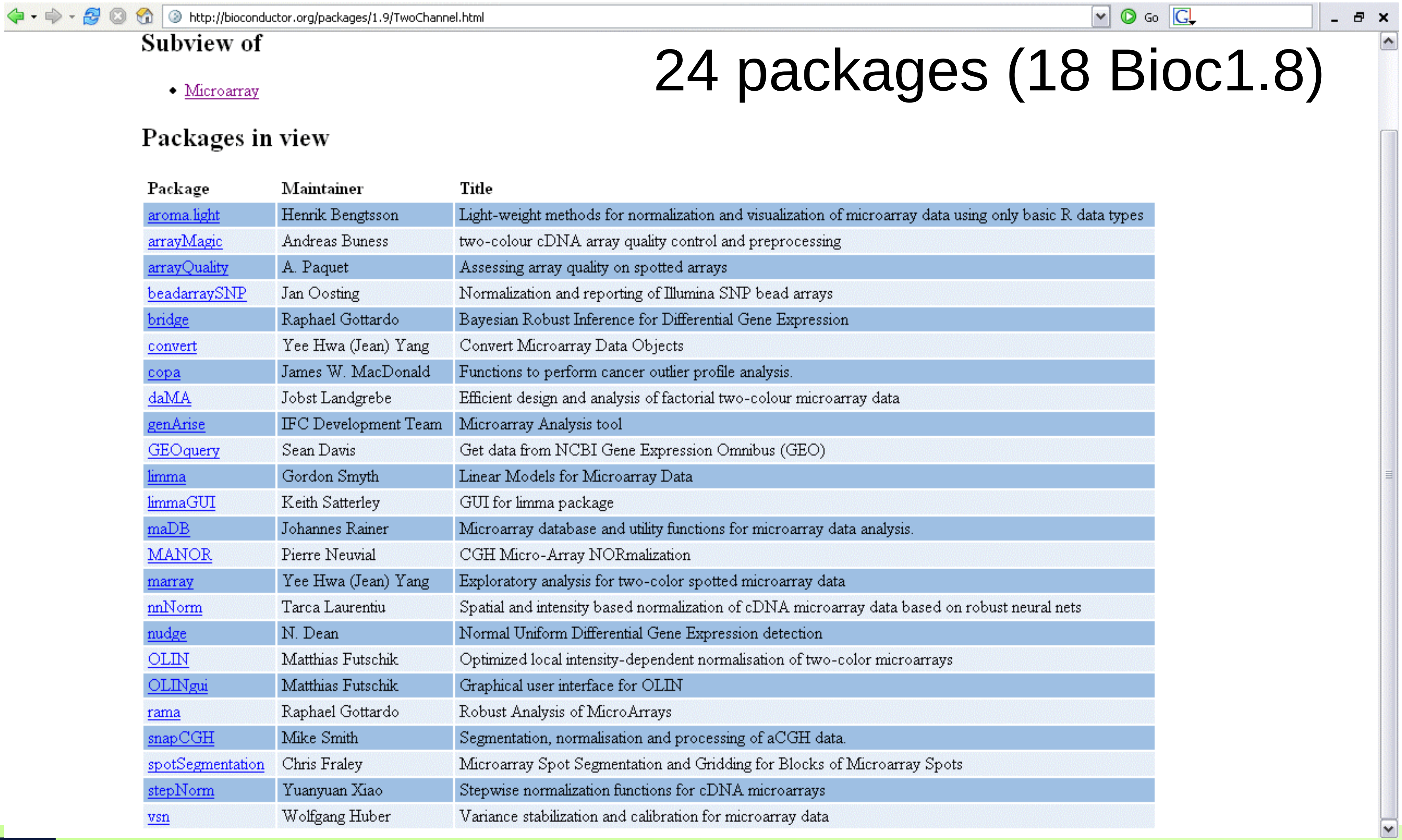
- Background correction or not
- Normalization: bias-variance trade-off
- Diagnostic plots

BioC pacakges:

- *marray*
- *limma*
- *…*

BioC Task View: TwoChannel

24 packages (18 Bioc1.8)



Subview of

- Microarray

Packages in view

Package	Maintainer	Title
aroma.light	Henrik Bengtsson	Light-weight methods for normalization and visualization of microarray data using only basic R data types
arrayMagic	Andreas Buness	two-colour cDNA array quality control and preprocessing
arrayQuality	A. Paquet	Assessing array quality on spotted arrays
beadarraySNP	Jan Oosting	Normalization and reporting of Illumina SNP bead arrays
bridge	Raphael Gottardo	Bayesian Robust Inference for Differential Gene Expression
convert	Yee Hwa (Jean) Yang	Convert Microarray Data Objects
copa	James W. MacDonald	Functions to perform cancer outlier profile analysis.
daMA	Jobst Landgrebe	Efficient design and analysis of factorial two-colour microarray data
genArise	IFC Development Team	Microarray Analysis tool
GEOquery	Sean Davis	Get data from NCBI Gene Expression Omnibus (GEO)
limma	Gordon Smyth	Linear Models for Microarray Data
limmaGUI	Keith Satterley	GUI for limma package
maDB	Johannes Rainer	Microarray database and utility functions for microarray data analysis.
MANOR	Pierre Neuvial	CGH Micro-Array NORmalization
marray	Yee Hwa (Jean) Yang	Exploratory analysis for two-color spotted microarray data
nnNorm	Tarca Laurentiu	Spatial and intensity based normalization of cDNA microarray data based on robust neural nets
nudge	N. Dean	Normal Uniform Differential Gene Expression detection
OLIN	Matthias Futschik	Optimized local intensity-dependent normalisation of two-color microarrays
OLINgui	Matthias Futschik	Graphical user interface for OLIN
rama	Raphael Gottardo	Robust Analysis of MicroArrays
snapCGH	Mike Smith	Segmentation, normalisation and processing of aCGH data.
spotSegmentation	Chris Fraley	Microarray Spot Segmentation and Gridding for Blocks of Microarray Spots
stepNorm	Yuanyuan Xiao	Stepwise normalization functions for cDNA microarrays
vsn	Wolfgang Huber	Variance stabilization and calibration for microarray data