cDNA Microarray

Quality Assessment and Quality Control with BioConductor packages

Nolwenn Le Meur

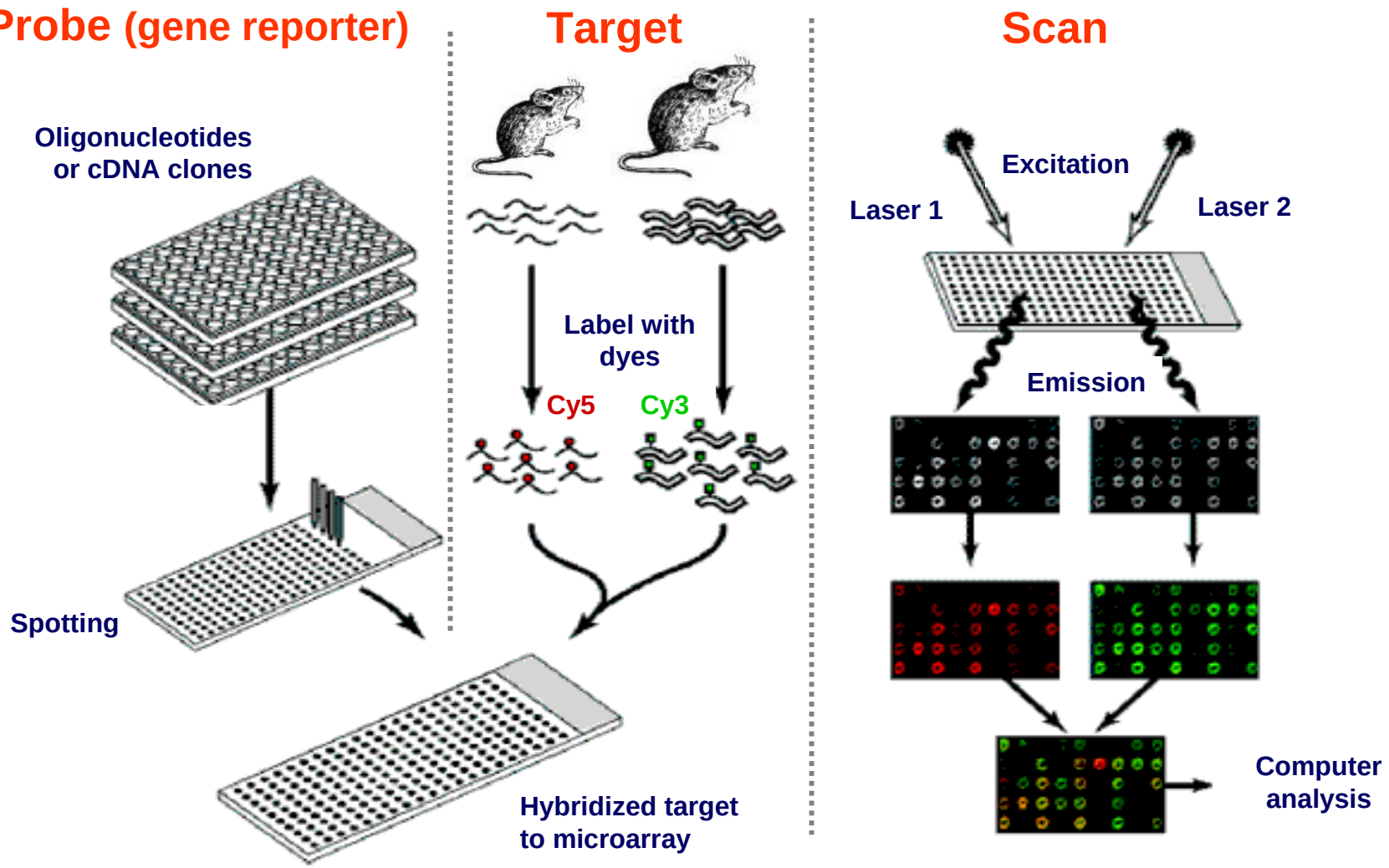
May 2007

FRED HUTCHINSON CANCER RESEARCH CENTER

Outline

- Image analysis
- Quality Assessment
- Pre-processing
  - Background correction
  - Normalization
  - Outliers detection

Two-color Microarray



(adapted from Duggan et al., Nat. Gen. , 1999)

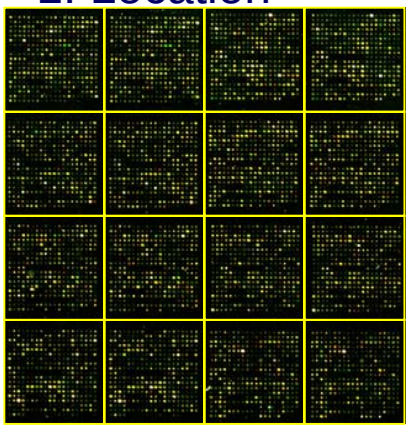
Terminology

- **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.

- **G, Gb:** Cy3 signal and background intensities
- **R, Rb:** Cy5 signal and background intensities
- **M** = log2(R) - log2(G)
- **A** = 1/2(log2(R) + log2(G))

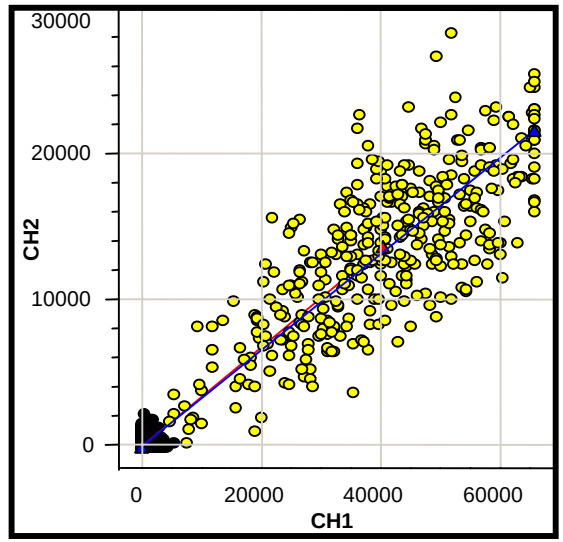
Image Analysis

1. Location

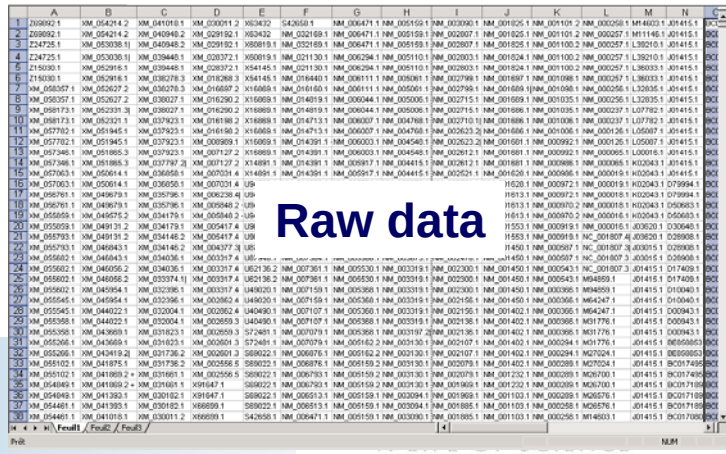


2. Segmentation

3. Quantification



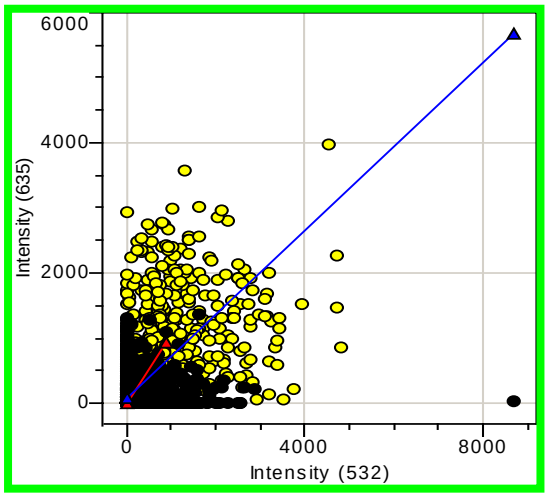
Raw data

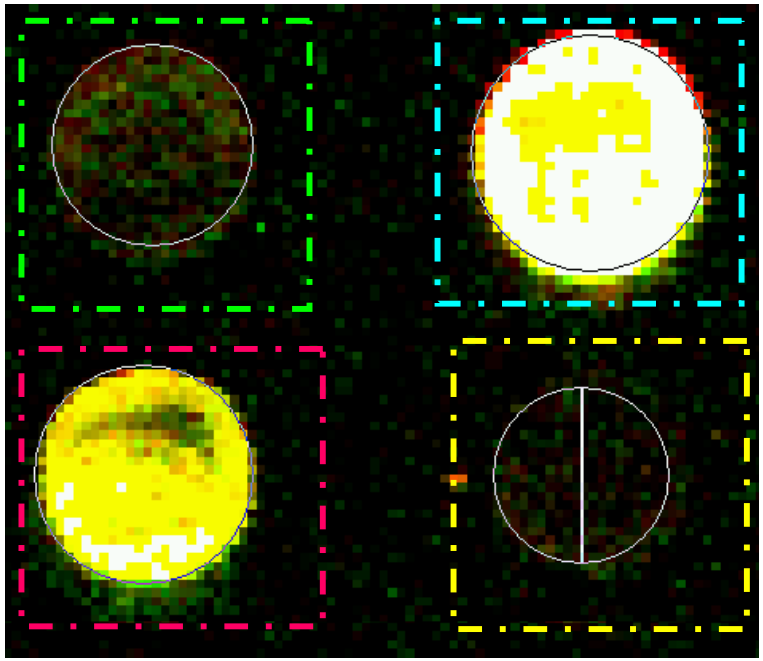


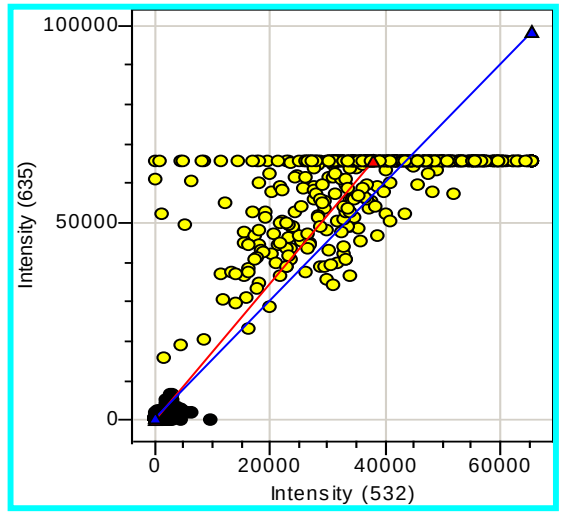
BIOCONDUCTOR

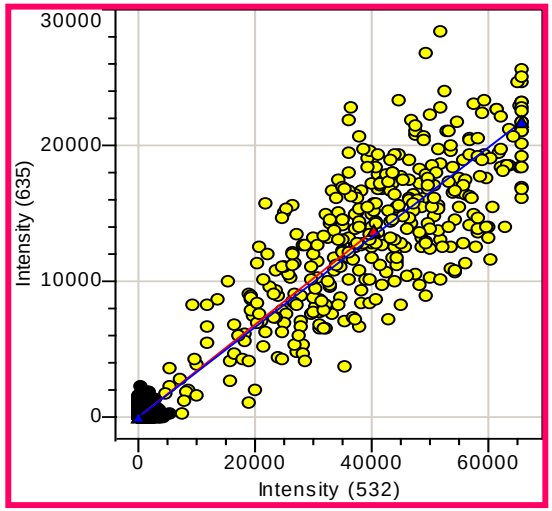
Quality Assessment

Probe level

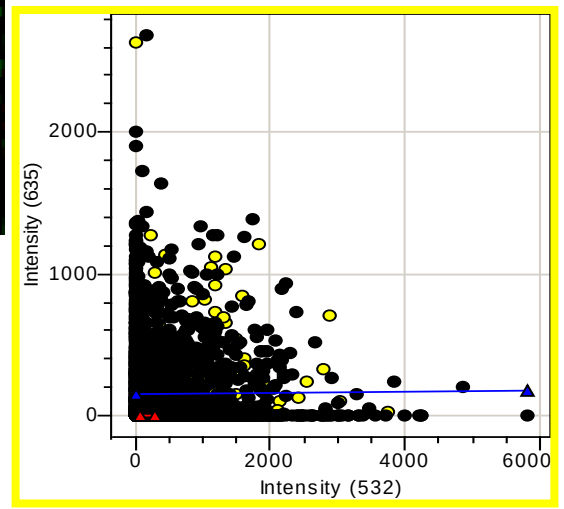








- Background
- Foreground



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% excluded spots)
  - Range of intensities
  - (Av. Foreground)/(Av. Background) > 3 in both channels
  - Distribution of spot signal area

Quality Assessment

For each array:

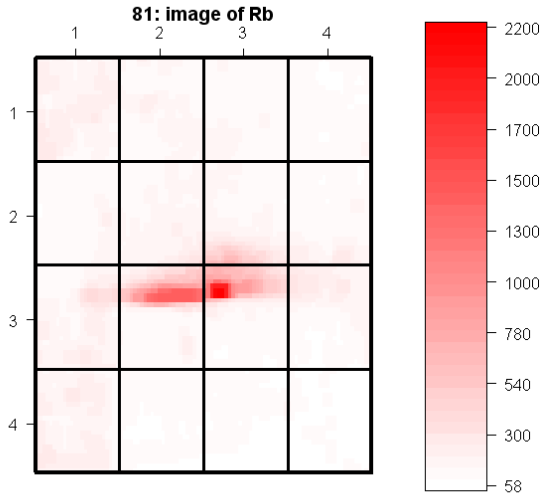
- **Visual inspection**
  - hairs, dust, scratches, air bubbles, dark regions, regions with haze
- **Diagnostics plots** of spot statistics

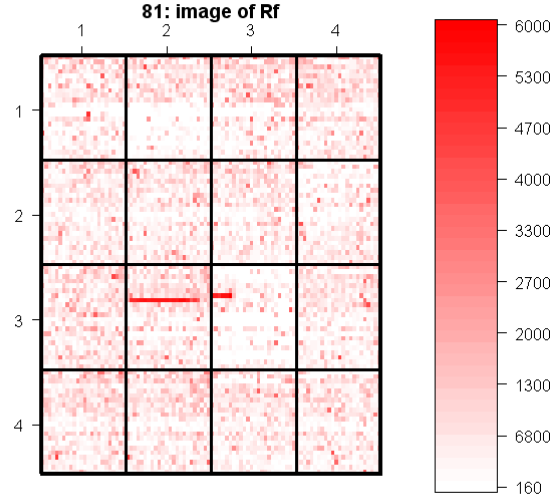
  e.g. R and G log-intensities, M, A, spot area.
  - 2D spatial images;
  - ECDF plots;
  - Boxplots;
  - Scatter-plots;
  - Density plots.
- **Stratify** plots according to layout parameters, *e.g.* print-tip-group, plate.

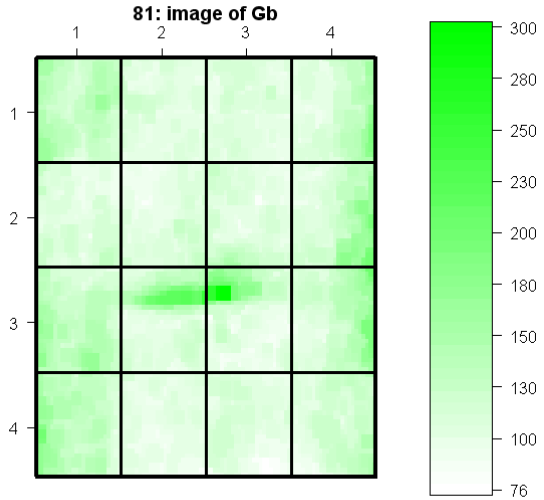
Image Plots

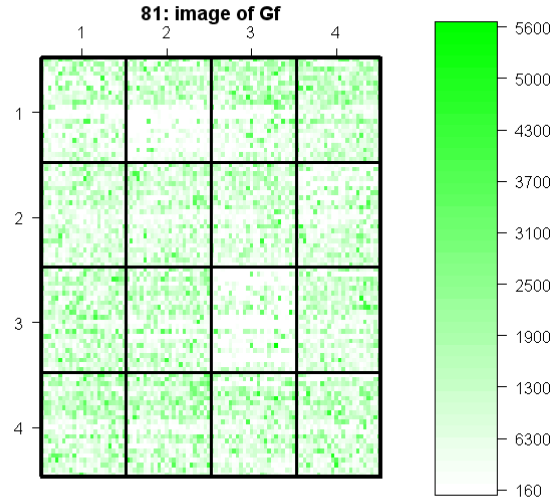
R Console

```
> library(«marray»)
> data(swirl)
> Gcol <- maPalette(low="white", high="green",k = 50)
> Rcol <- maPalette(low ="white",high="red",k = 50)
>image(swirl[,1],xvar= "maRb",col=Rcol)
>image(swirl[,1],xvar= "maRf",col=Rcol)
>image(swirl[,1],xvar= "maGb",col=Gcol)
>image(swirl[,1],xvar= "maRf",col=Gcol)
```

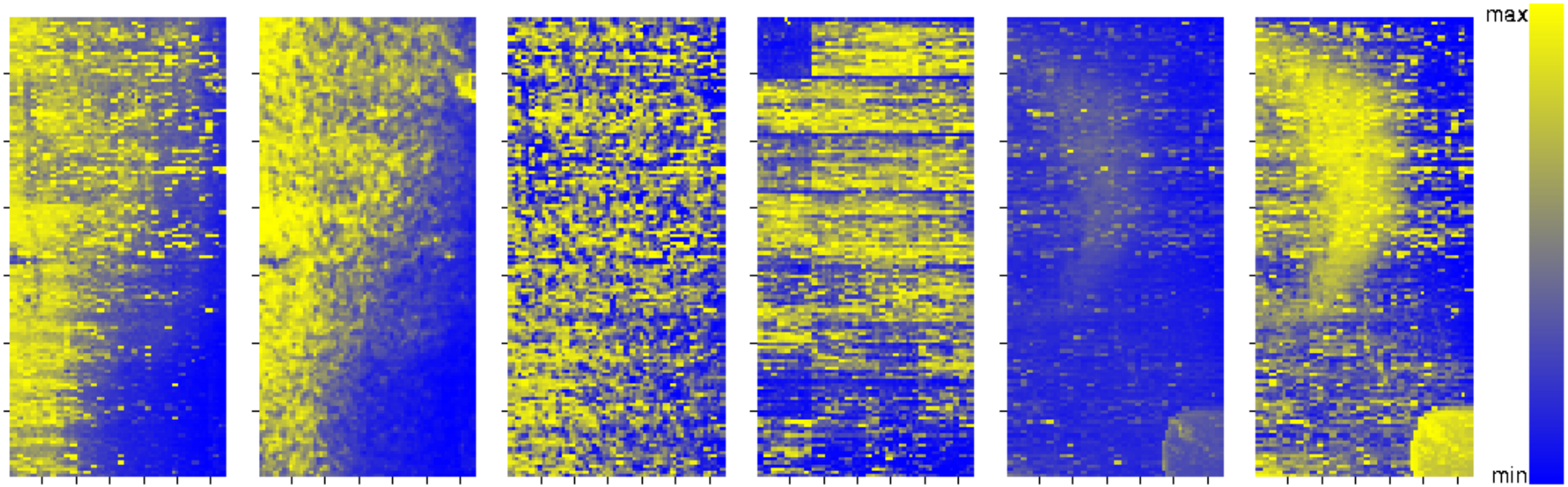








Spatial Effects – Image Plots

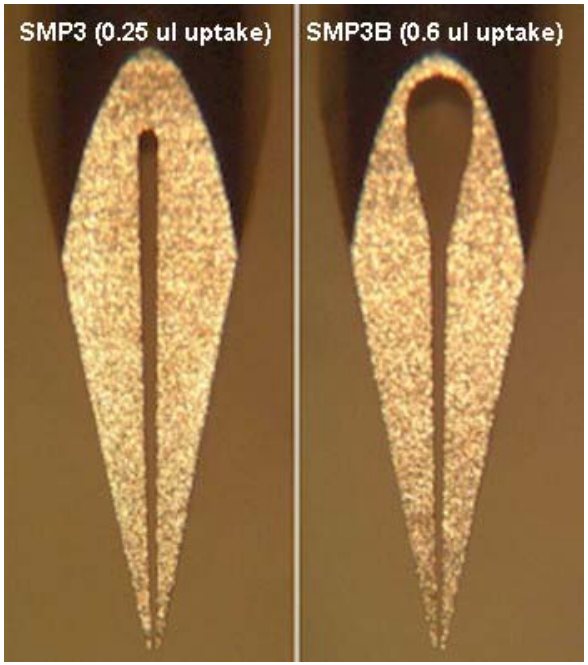


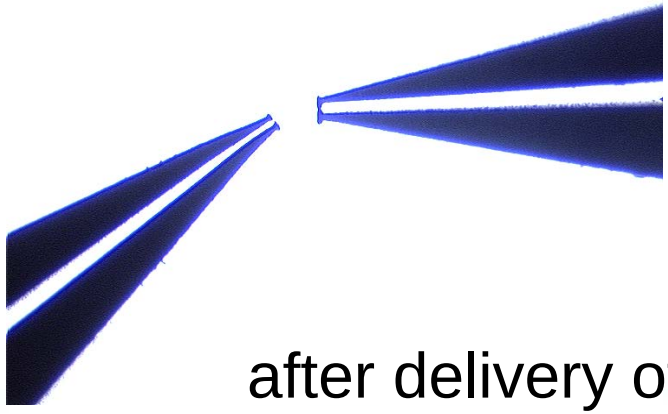
R **Rb** **R-Rb**

color scale by rank

Print-tip **Washing**

Spotting Pin Quality Decline

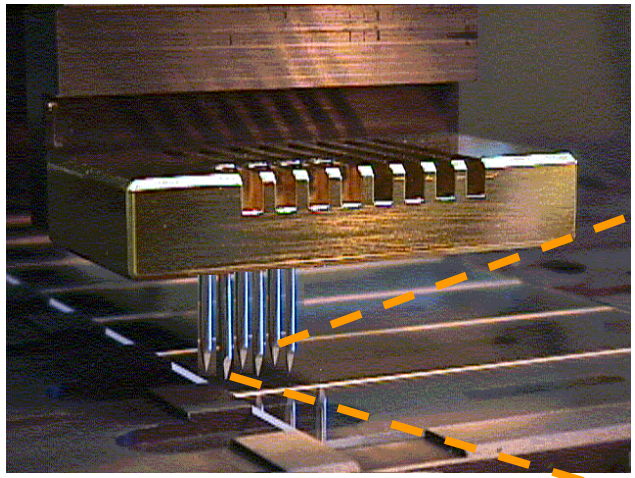


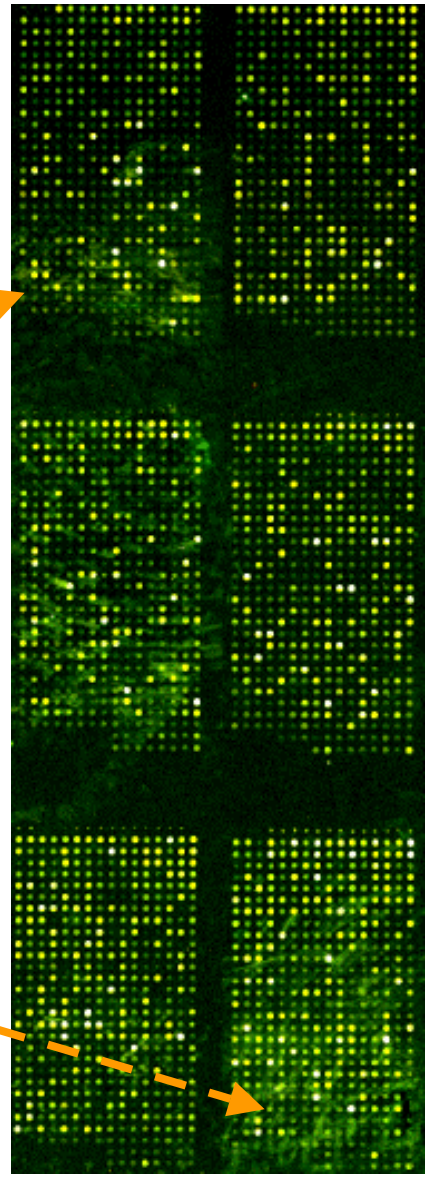


after delivery of $5x10^5$ spots

after delivery of $3x10^5$ spots

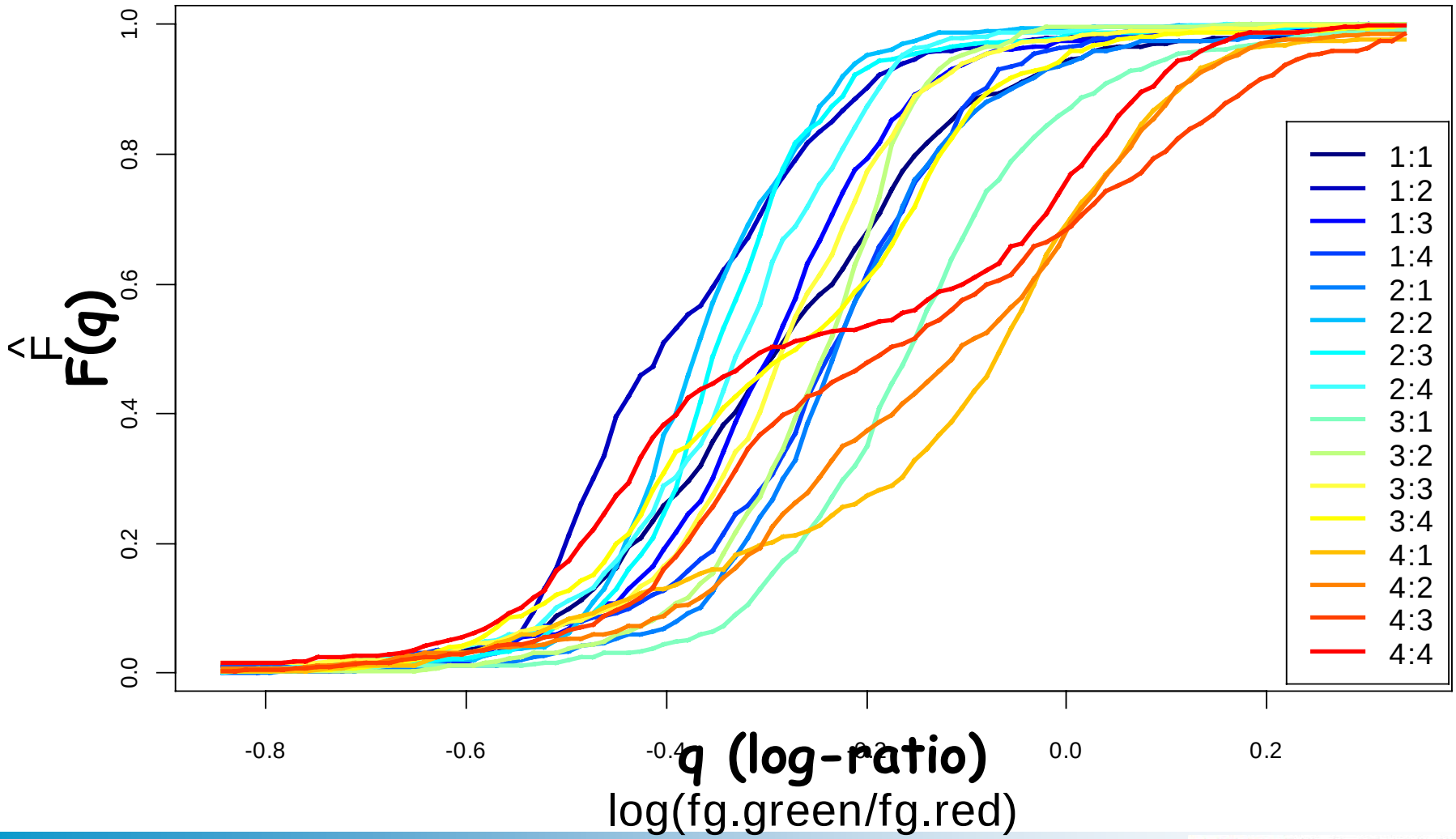
Spatial Effects



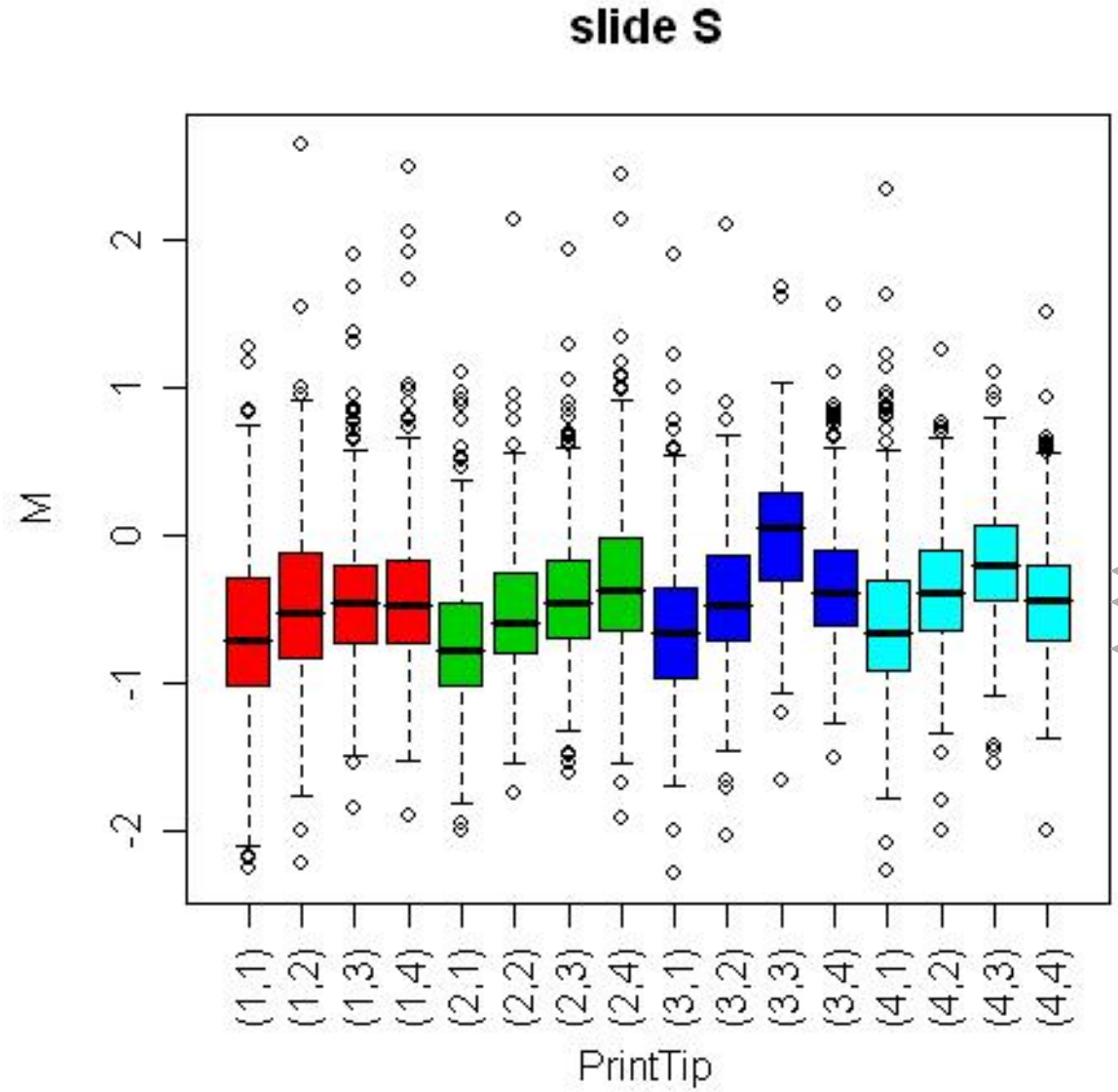


1 pin → **1 block**

Print-tip Effects – ECDF plot

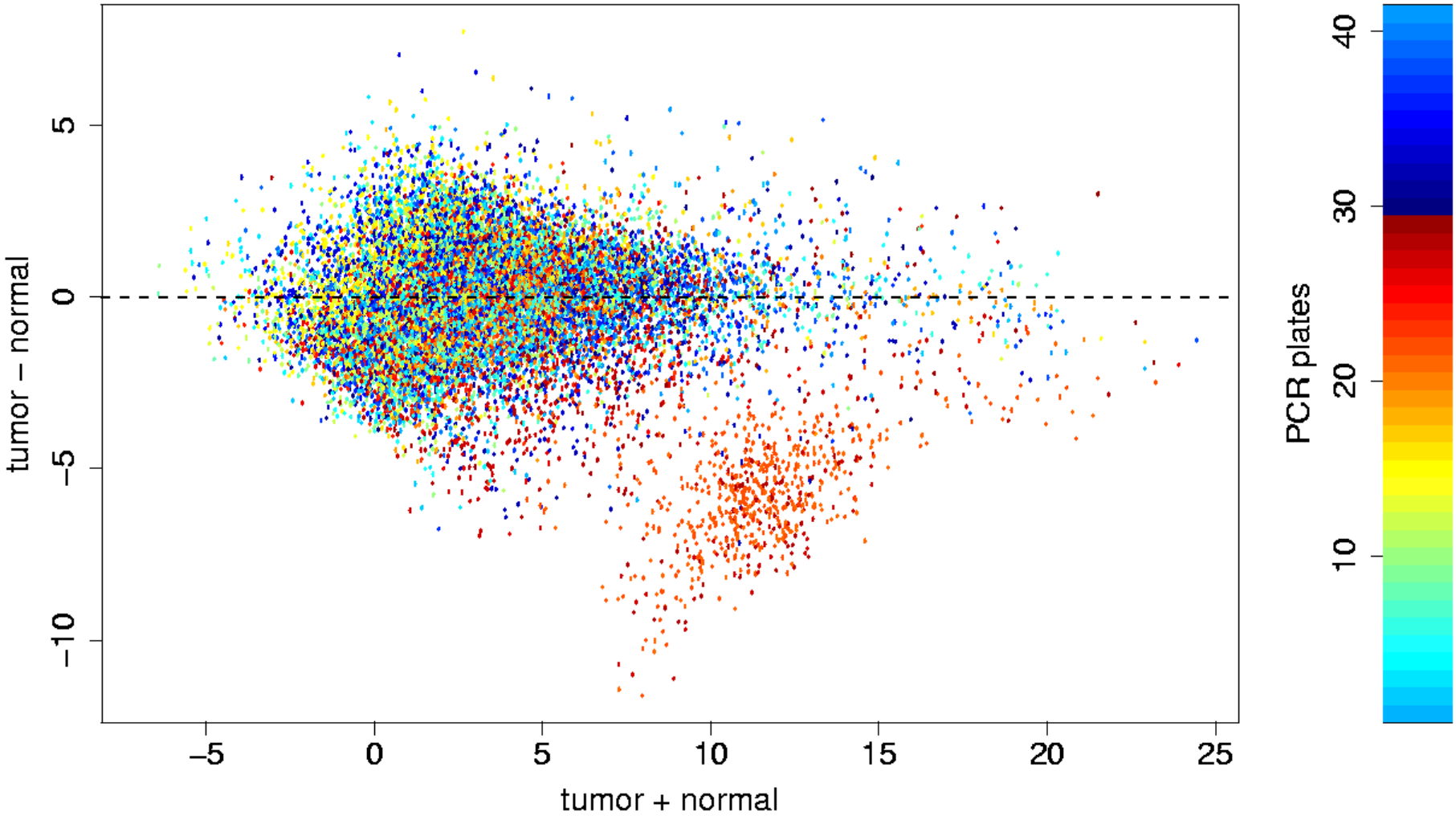


Print-tip Effects - Boxplots

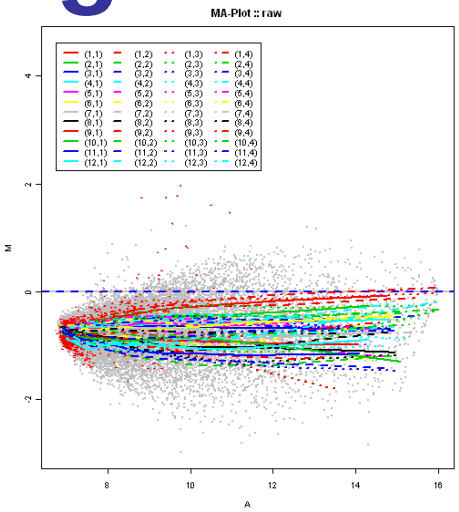


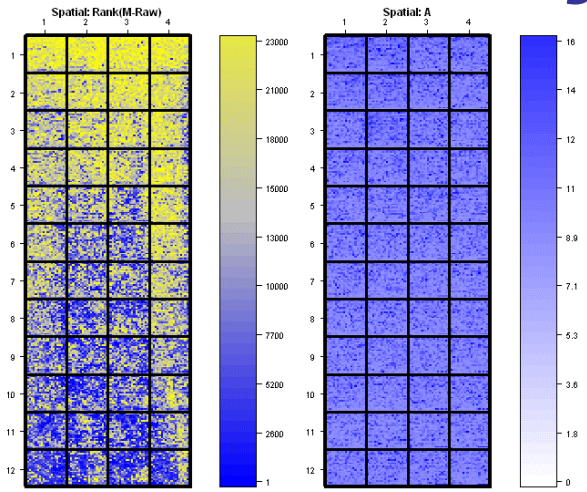
Q3 75% quantile
median
Q1 25% quantile

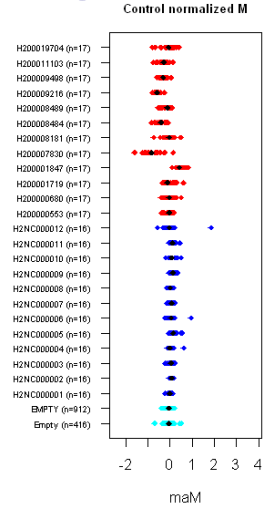
PCR plates

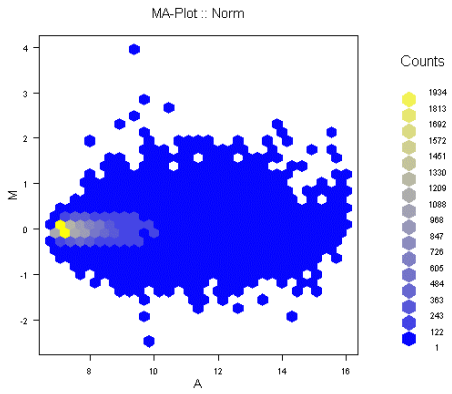

tumor – normal
5
0
−5
−10
−5
0
5
10
15
20
25
tumor + normal
PCR plates
40
30
20
10

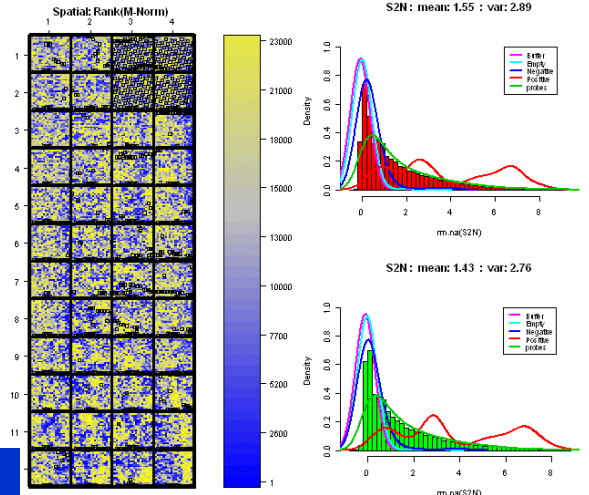
Diagnostic plot with *arrayQuality*

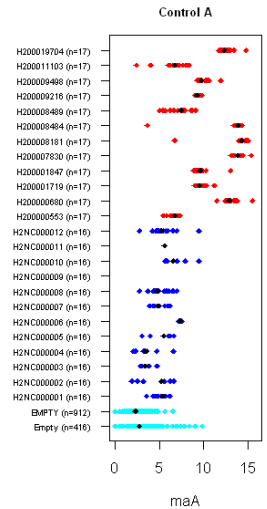












R Console

```
library(«arrayQuality»)
?maQualityPlots
```

BIOCONDUCTOR

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Diagnostic with *arrayMagic*

FileName	DL32	DL31	DL30	HuPr_4005_D
Cy3	PEC34_cntrl	PEC34_HGF	PEC34_SFN-	PEC34_SFN
Cy5	Control	Control	Control	Control
hybridisation	HuPr_4008_DL3	HuPr_4007_DL3	HuPr_4006_D	HuPr_4005_D
width	0.39	0.36	0.43	0.43
medianDistance	0.22	0.21	0.25	0.28
correlation	0.94	0.95	0.94	0.94
correlationLogRaw	0.82	0.86	0.82	0.79
meanSignalGreen	2701.65	2664.29	1558.81	2928.64
meanSignalRed	2797.52	2683.97	2037.56	3194.82
meanSignal	2749.59	2674.13	1798.19	3061.73
signalRange.Green	11001.25	11188.90	6289.65	11982.30
signalRange.Red	11446.00	11207.00	8602.65	13497.80
backgroundRange.Green	21.00	28.00	14.00	23.00
backgroundRange.Red	28.00	19.00	28.00	37.00
signalToBackground.Green	9.14	8.30	5.67	10.88
signalToBackground.Red	14.01	14.00	10.53	16.08

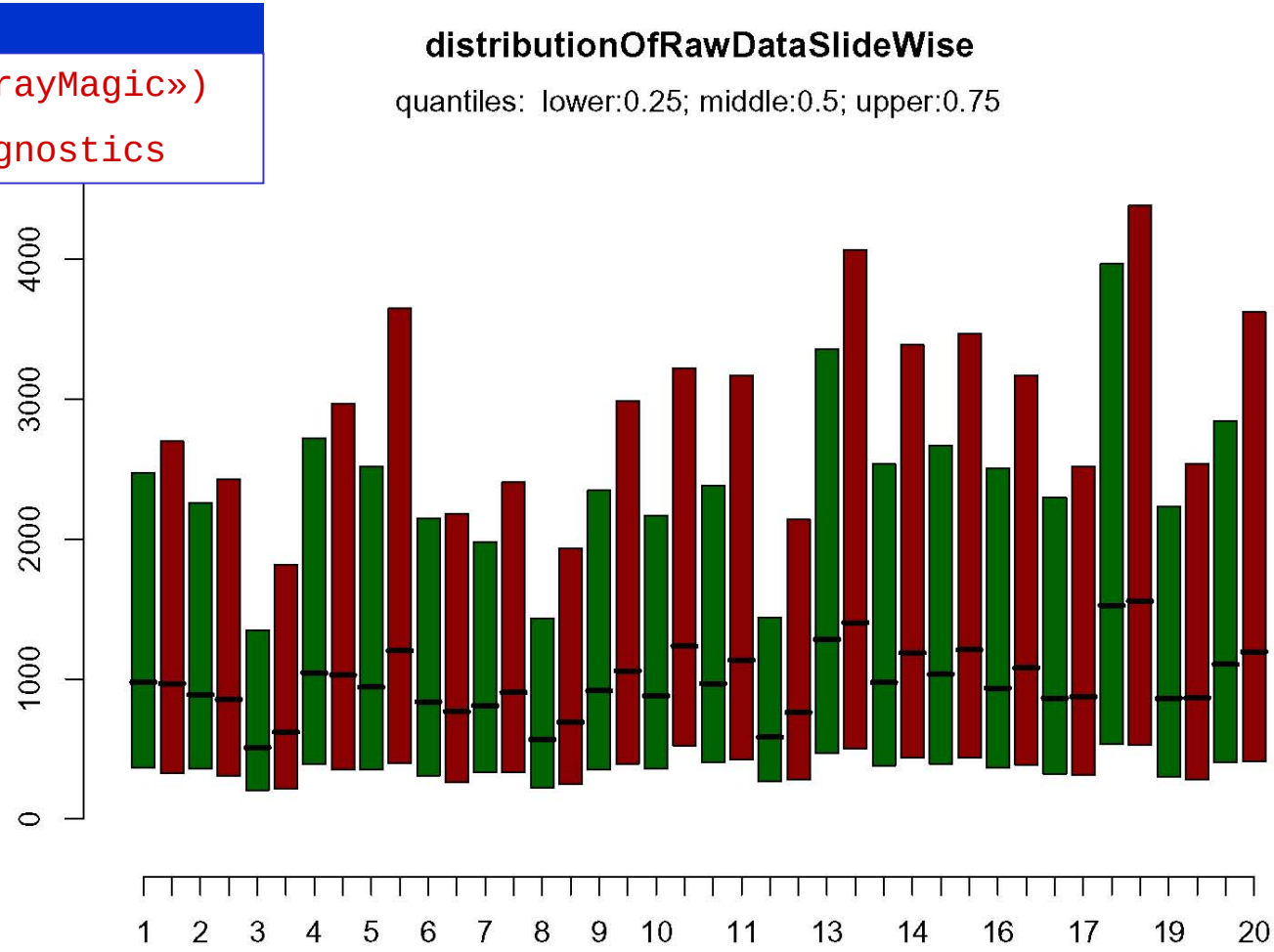
R Console

```
library(«arrayMagic»)
?qualityParameters
```

Diagnostic with *arrayMagic*

R Console

```
library(«arrayMagic»)
?qualityDiagnostics
```



Quality Assessment: Summary

For each spot:

- weight

For each array:

- Diagnostics plots
- Stratify
- Controls

BioC packages:

- *arrayQuality*
- *arrayMagic*
- *…*

Outline

- Image analysis
- Quality Assessment
- Pre-processing
  - Background correction
  - Normalization
  - Outliers detection

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 quality
- Spotting efficiency
- cross-hybridization
- …

Stochastic

- too random to be explicitely accounted for
- “noise”

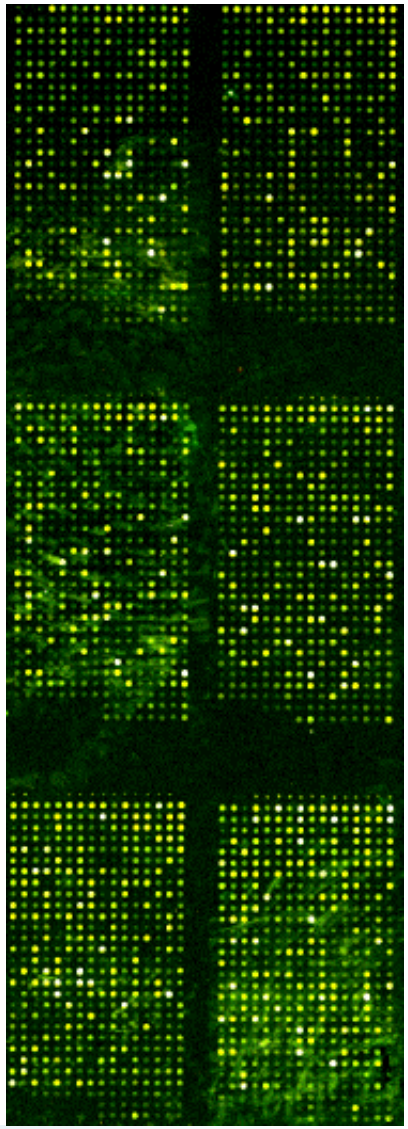
Error Model

Background Correction

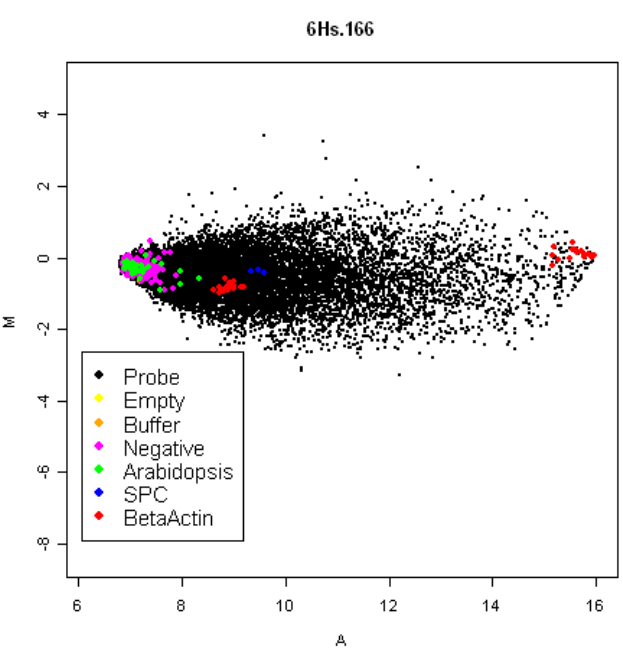
- none
- subtraction, movingmin
- *Minimun,edwards, normexp,…*

- More details … *limma*

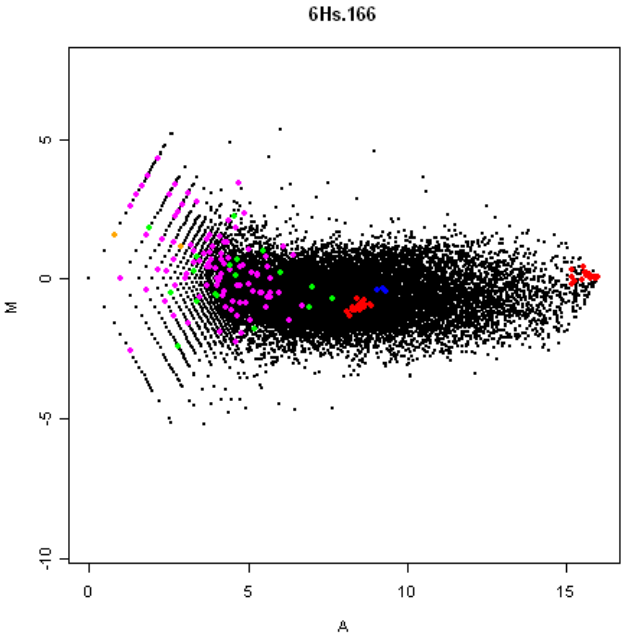
  >?backgroundCorrect



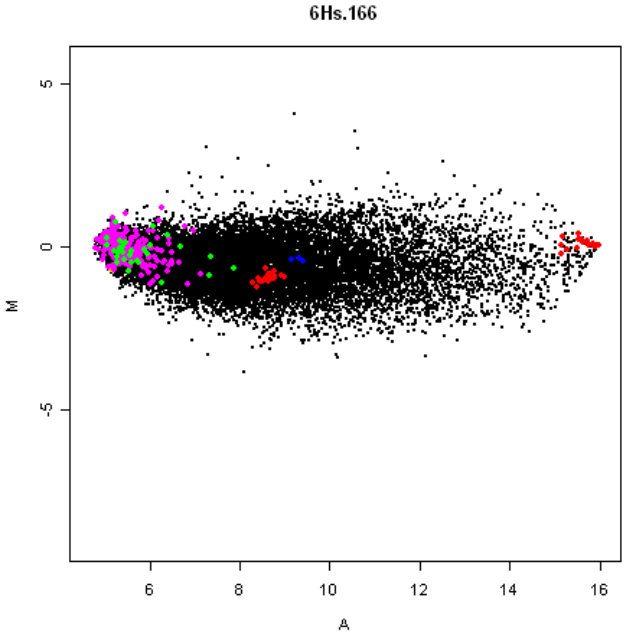
Background Correction



none

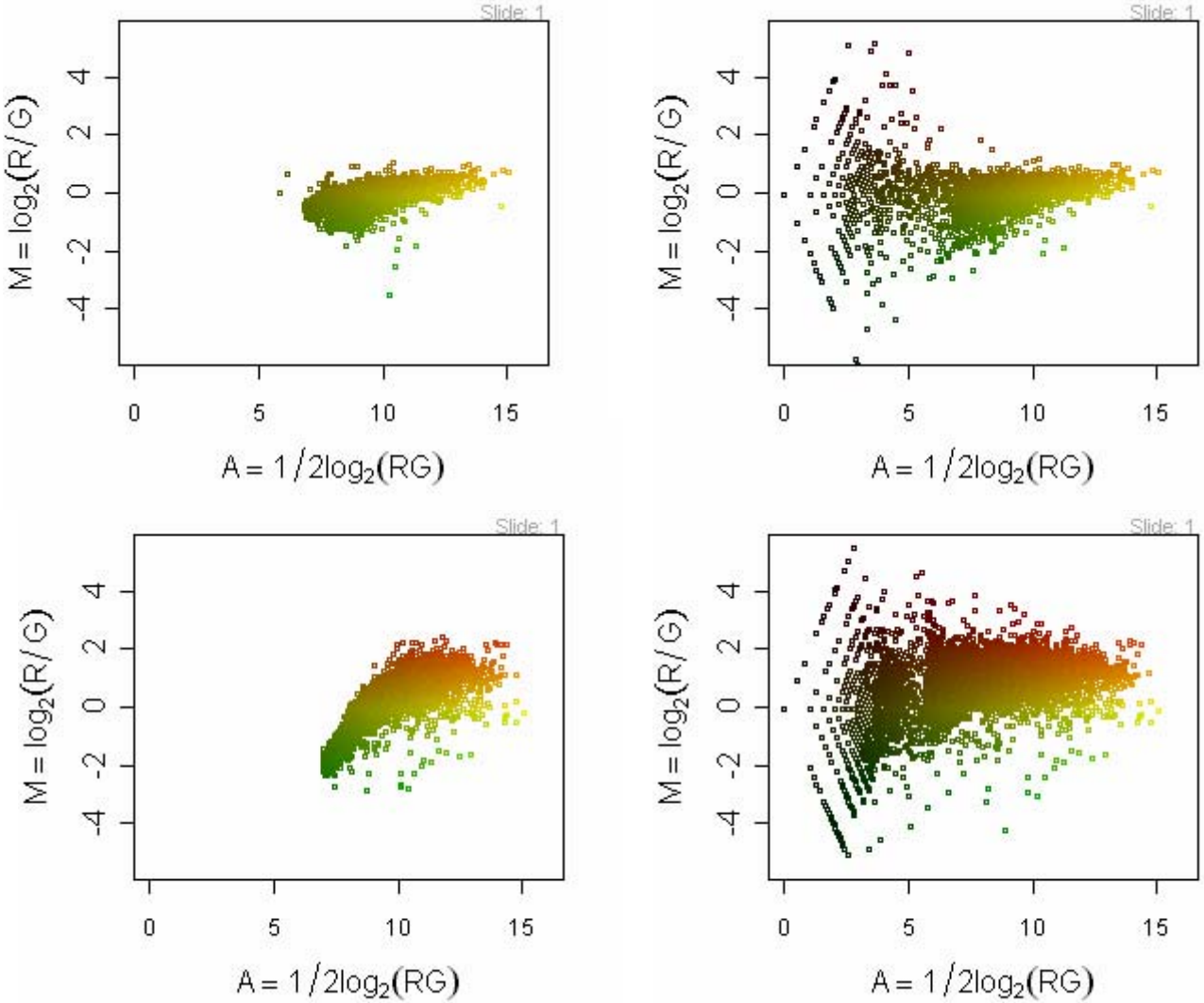


substraction



normexp

Background Correction



Normalization

Identify and remove the effects of systematic variation

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

- 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 between slide comparisons of intensities, e.g., clustering, testing.

Normalization methods

- median
- loess
- print-tip loess
- …
- variance stabilisation
- Quantile
- …..

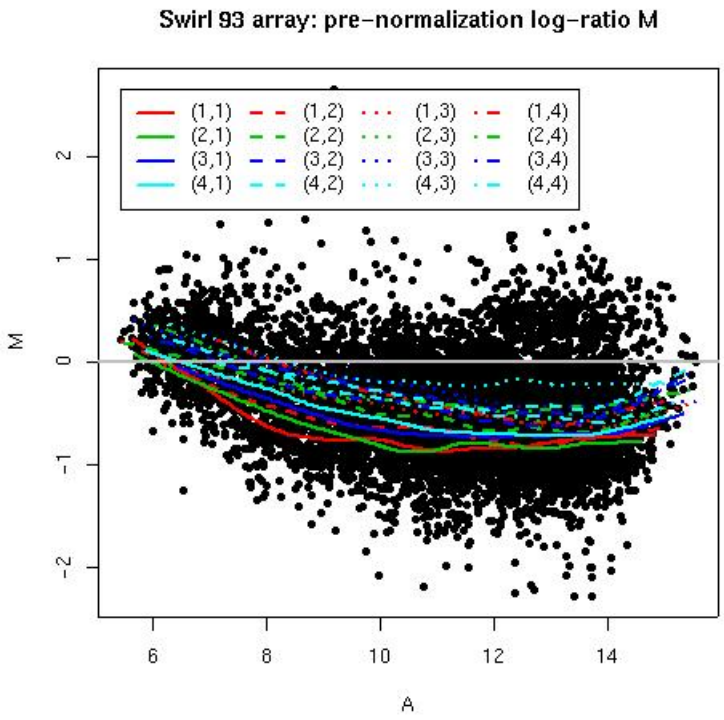
Two-channel (within)

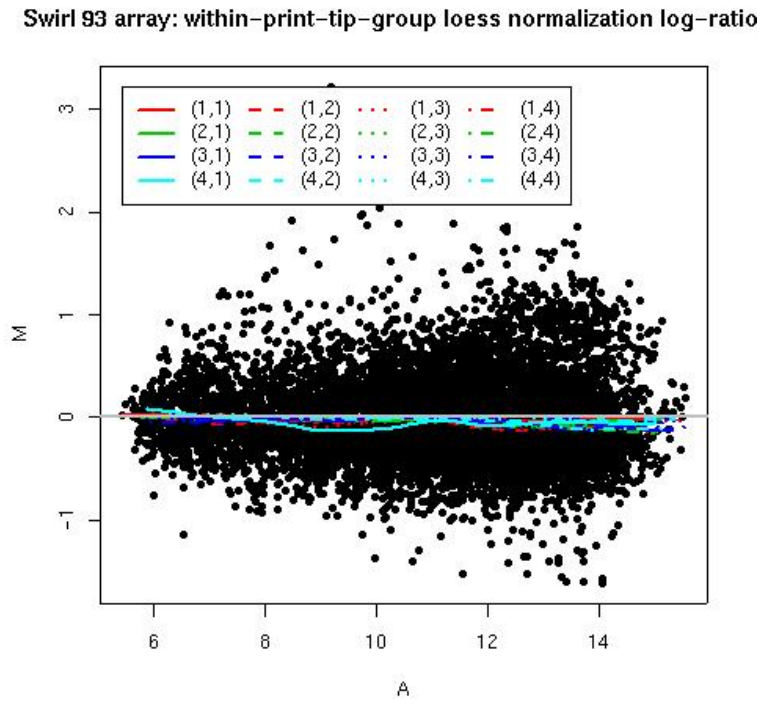
Separate-channel (between)

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

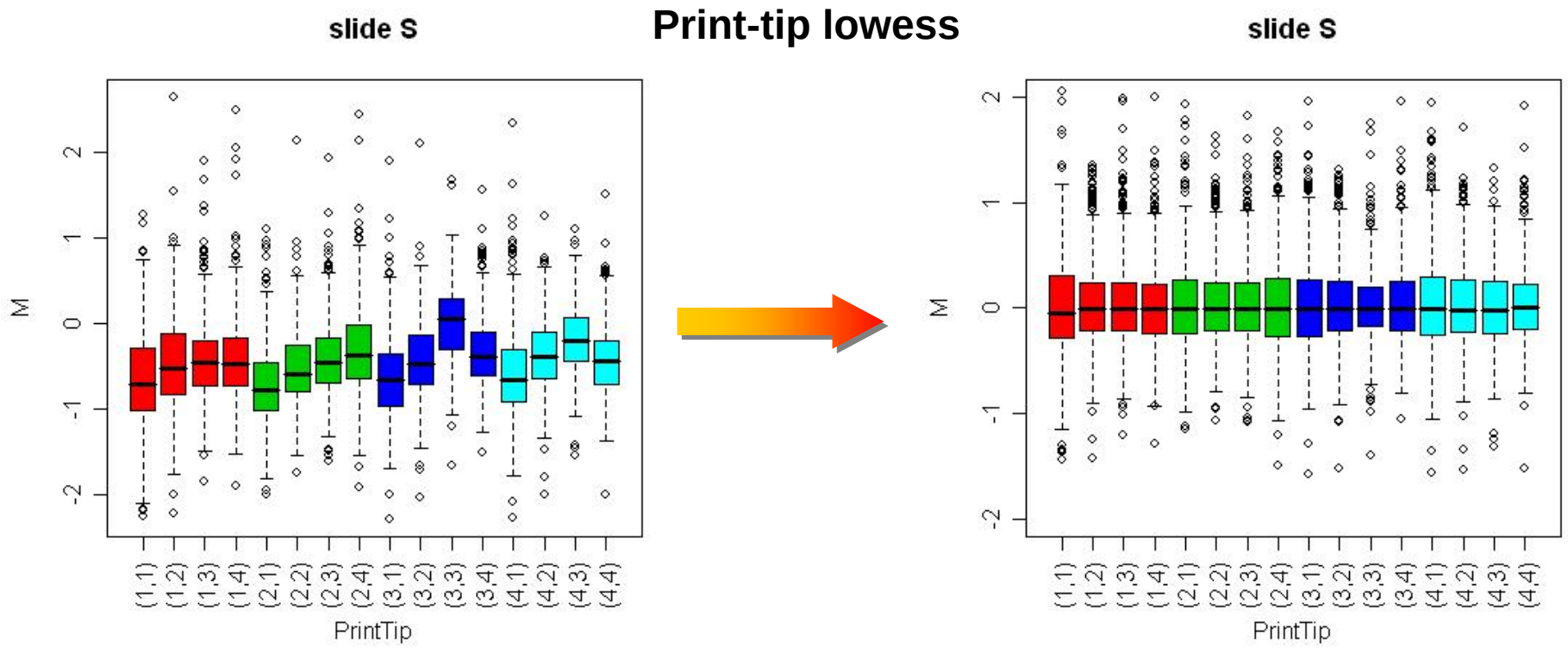
Two channel normalization

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





Two channels normalization

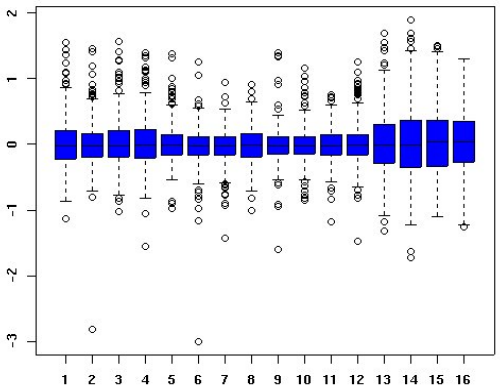


Two channels normalization

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

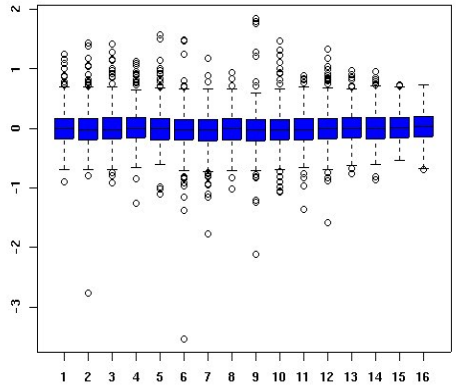
- Scale: adjust for different in scale between multiple arrays

median centered



Scaling

median centered & MAD scaled



One channel normalization

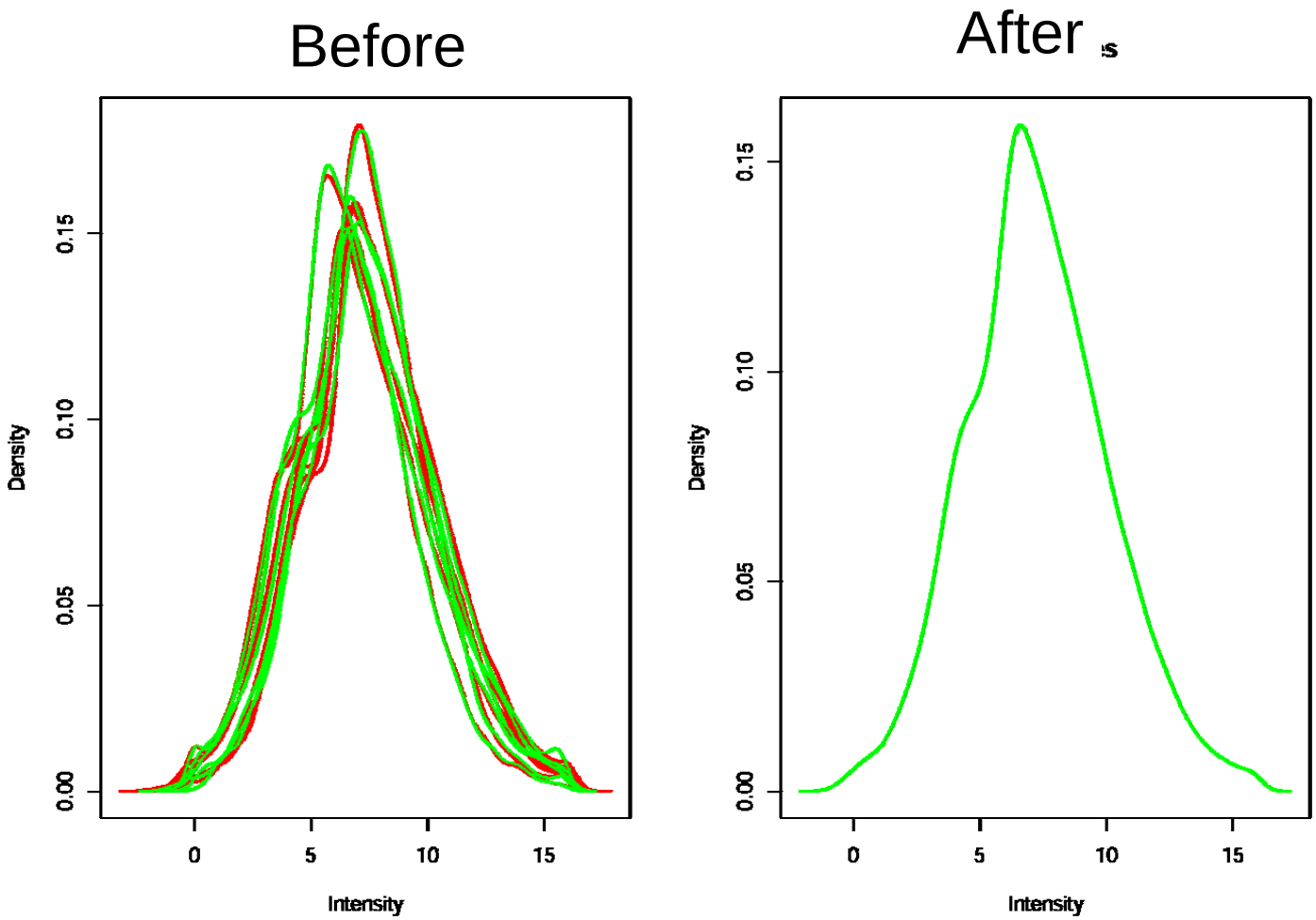
- As technology improves the spot-to-spot variation is reduced

- Development of normalization techniques that work on the absolute intensities

Ex: quantile normalization (*limma*)

variance stabilization (*vsn*)

Quantile Normalization

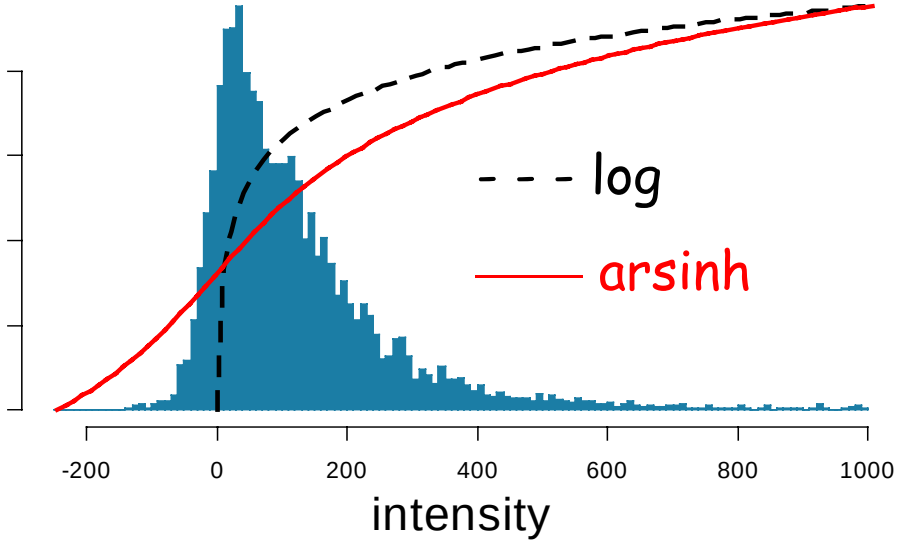


Bolstand *et al.*(2003)

Variance Stabilization Transformation

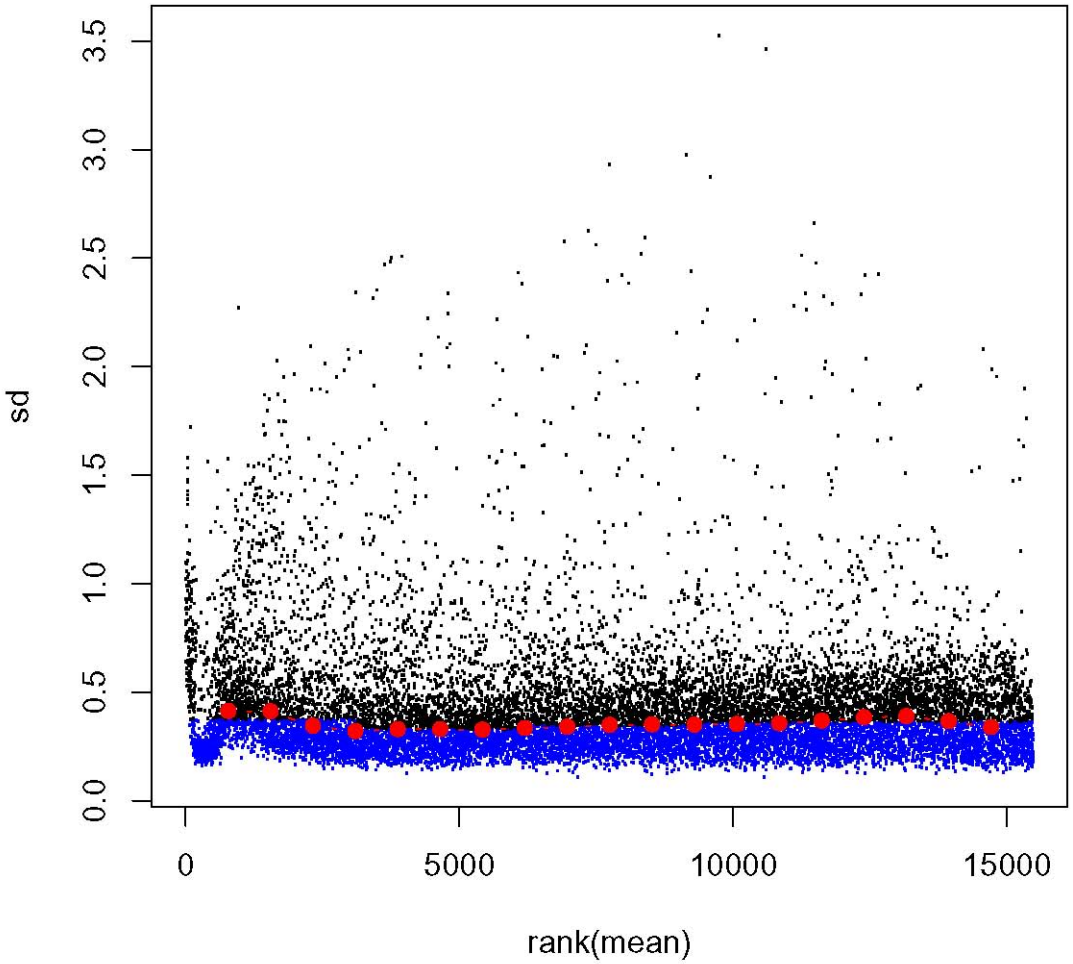
- log-transformation is replaced by a arsinh transformation
  - Meaningful around 0
  - Original intensities may be negatives

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

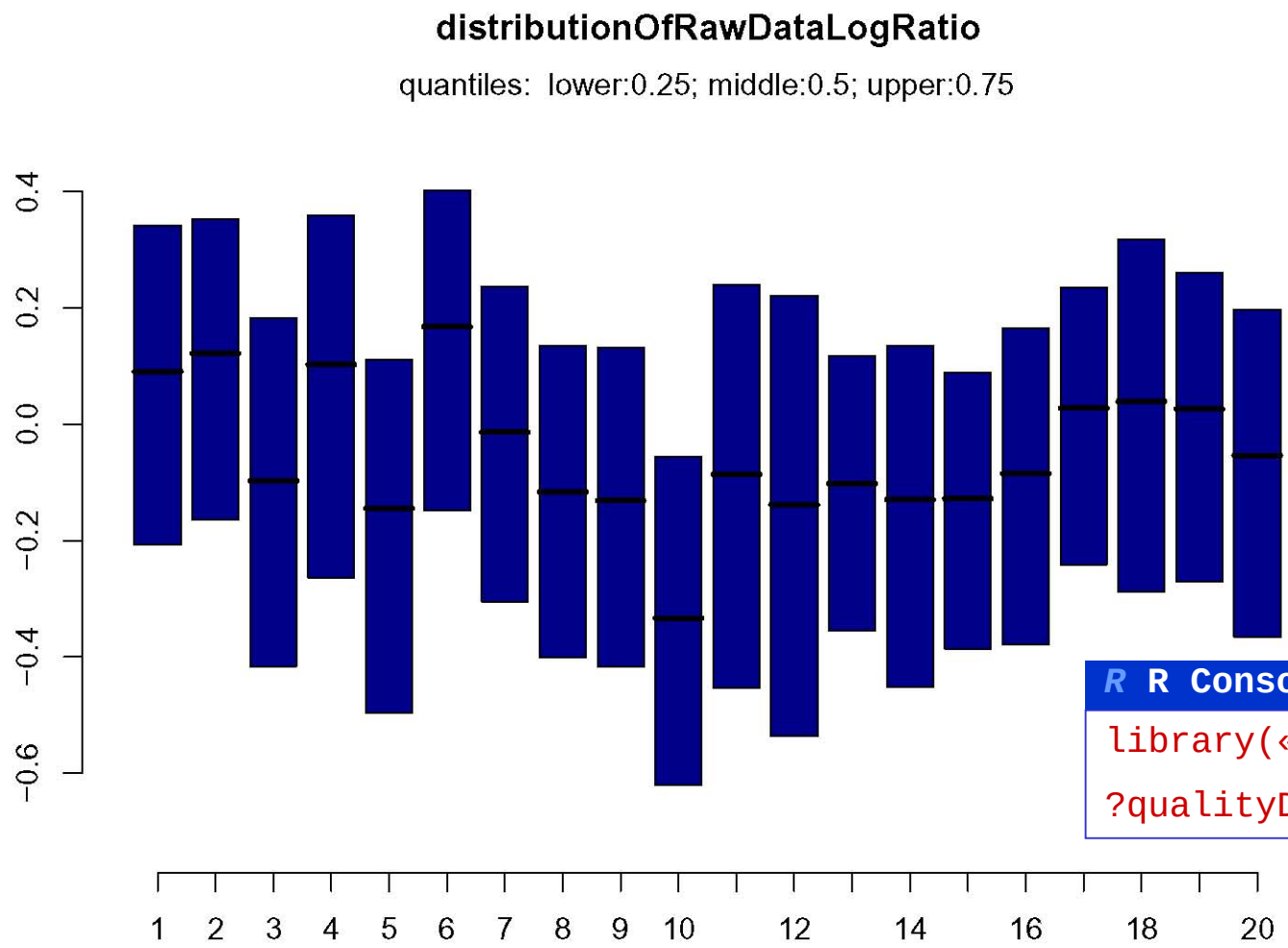


(Huber *et al.* 2004)

Variance Stabilization



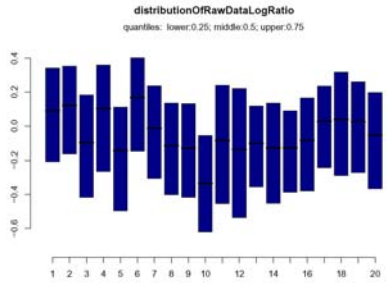
Diagnostic plot with *arrayMagic*



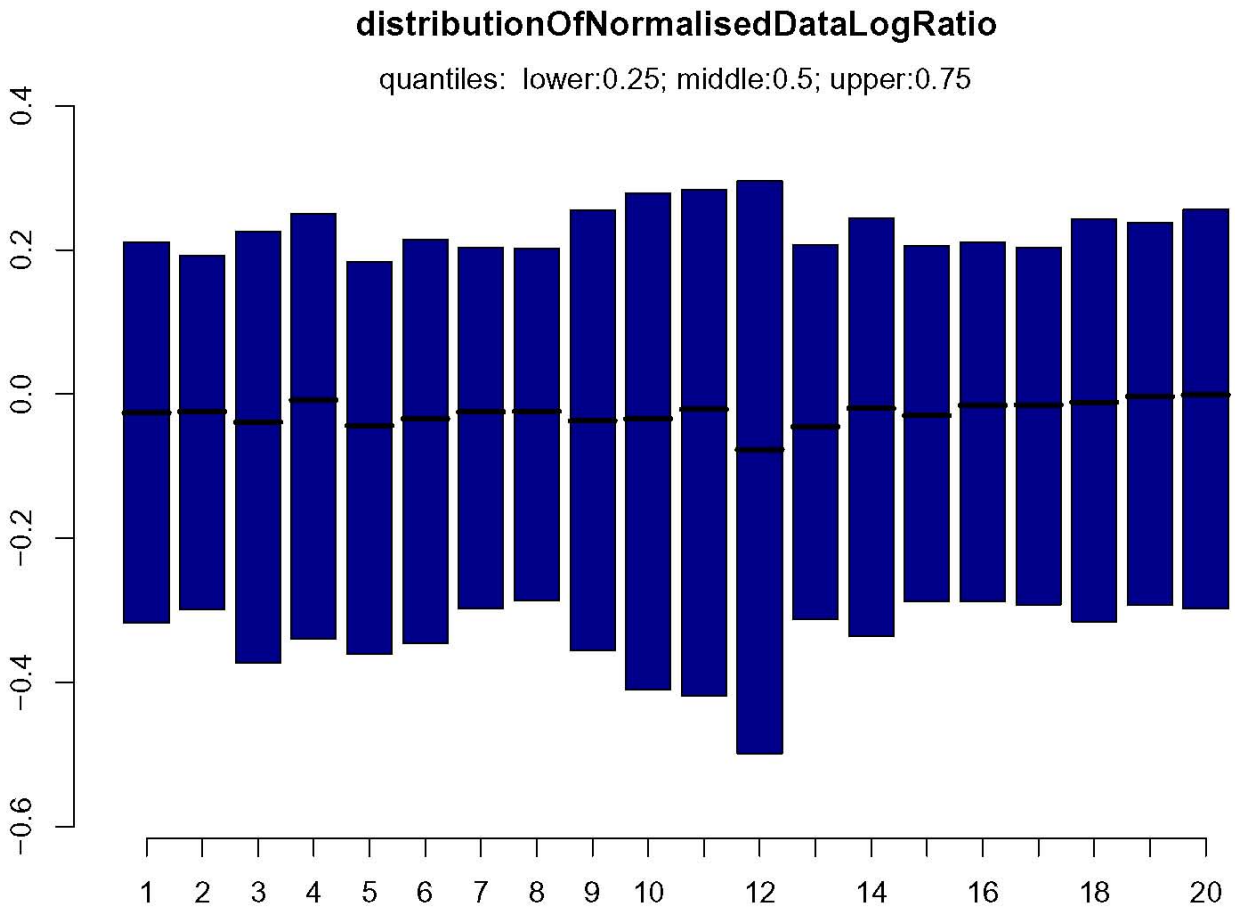
R R Console

```
library(«arrayMagic»)
?qualityDiagnostics
```

Diagnostic plot with *arrayMagic*

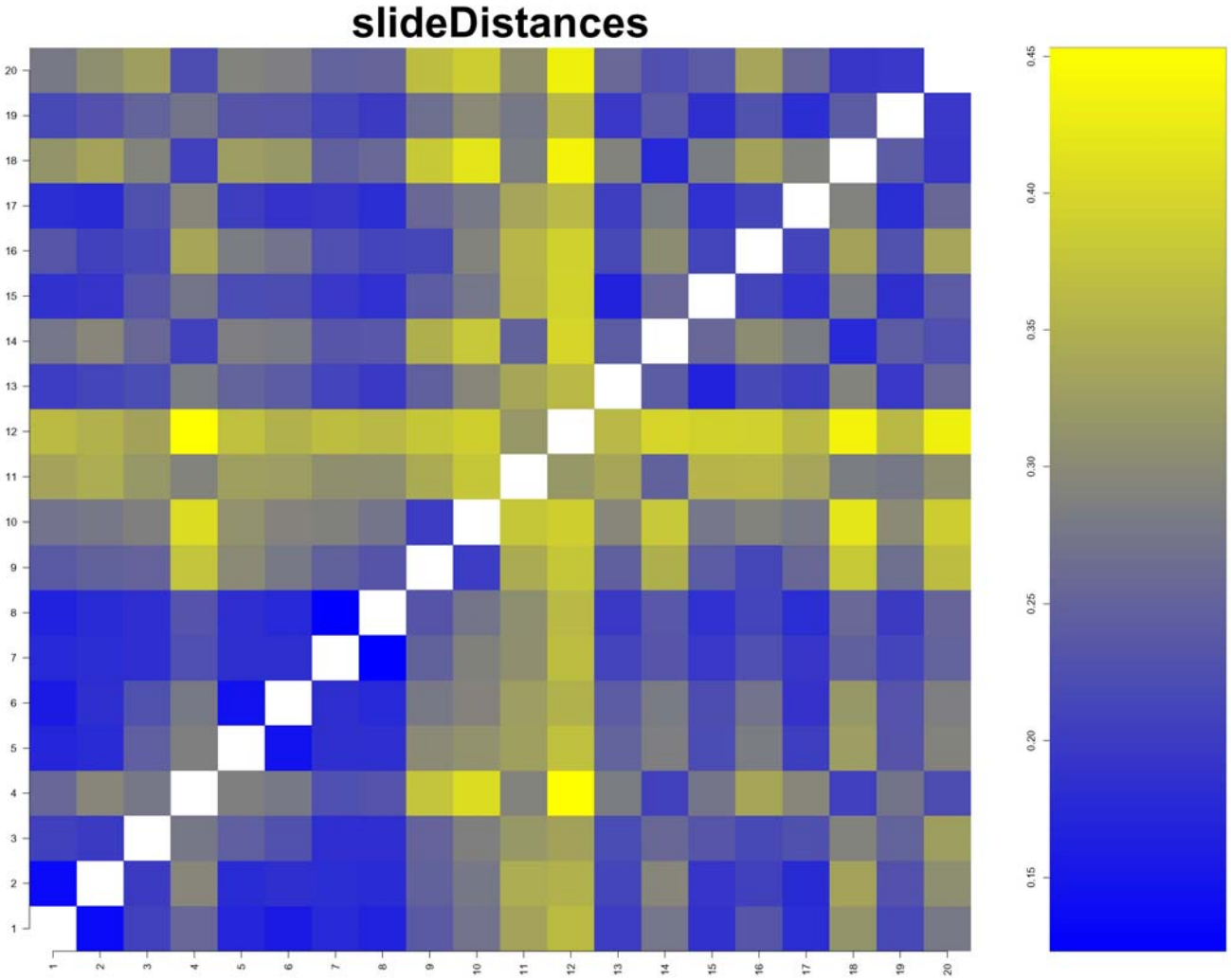


Raw data

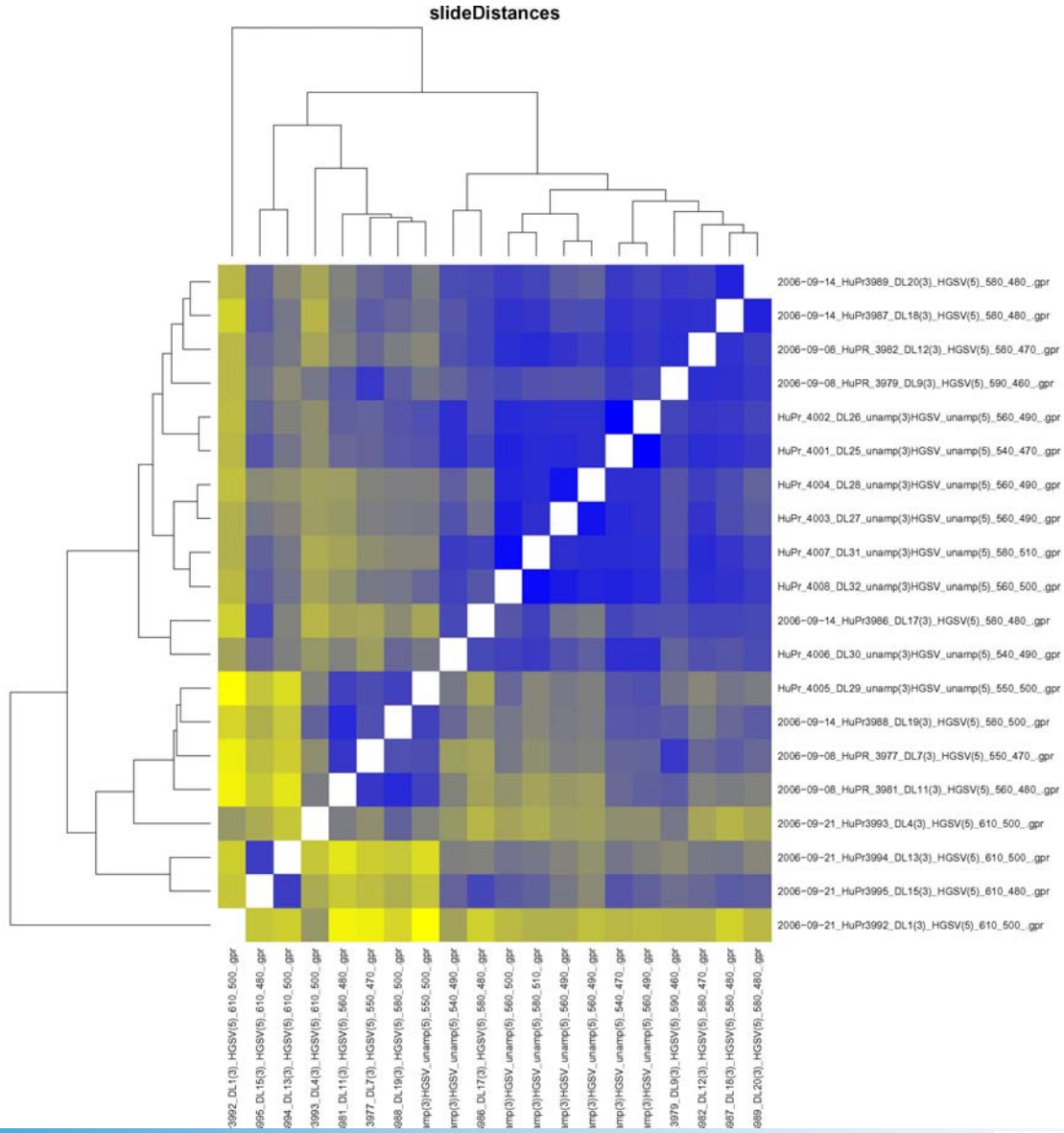


vsn normalized data

Diagnostic plot with *arrayMagic*



Diagnostic plot with *arrayMagic*



BIOCONDUCTOR

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

Outlier detection
(Technical replicates)

- Estimate the coefficient of variation
- Statistical tests (*e.g.,* Dixon test, Grubb's test
- DuplicateCorrelation() from *limma*
- Threshold (*e.g.,* max, mean)

Preprocessing : Summary

For each array:

- Background correction or not
- Normalization (within and/or between)
- Diagnostic plots QA/QC
- Replicates

BioC pacakges:

- *arrayQuality*
- *arrayMagic*
- *vsn*
- *limma*
- *…*

BioC Task View: TwoChannel

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