BIOCONDUCTOR

cDNA Microarray Analysis

with BioConductor packages

Nolwenn Le Meur

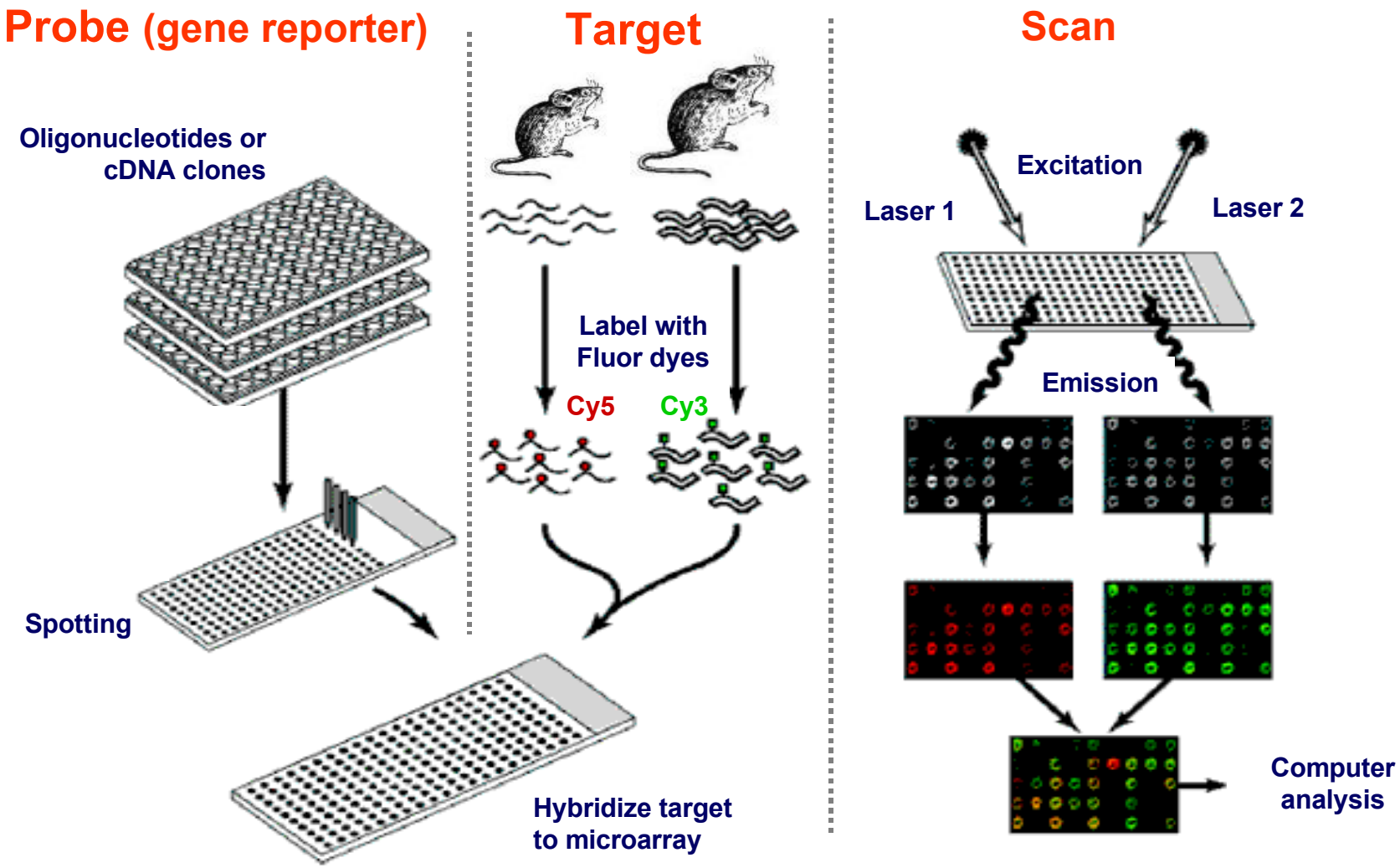
Copyright 2007

FRED HUTCHINSON
CANCER RESEARCH CENTER

Outline

- **Data acquisition**
- **Pre-processing**
  - Quality assessment
  - Pre-processing
    - background correction
    - normalization
    - summarization

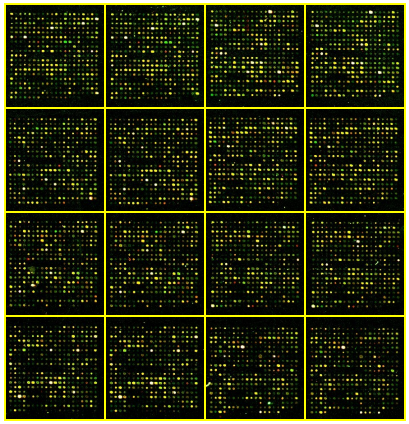
Two-color Microarray

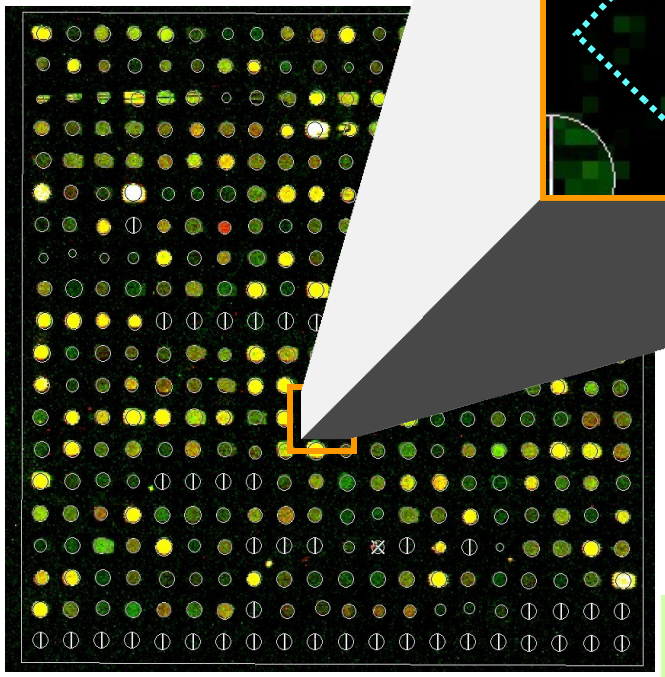


(adapted from Duggan et al., Nat. Gen. , 1999)

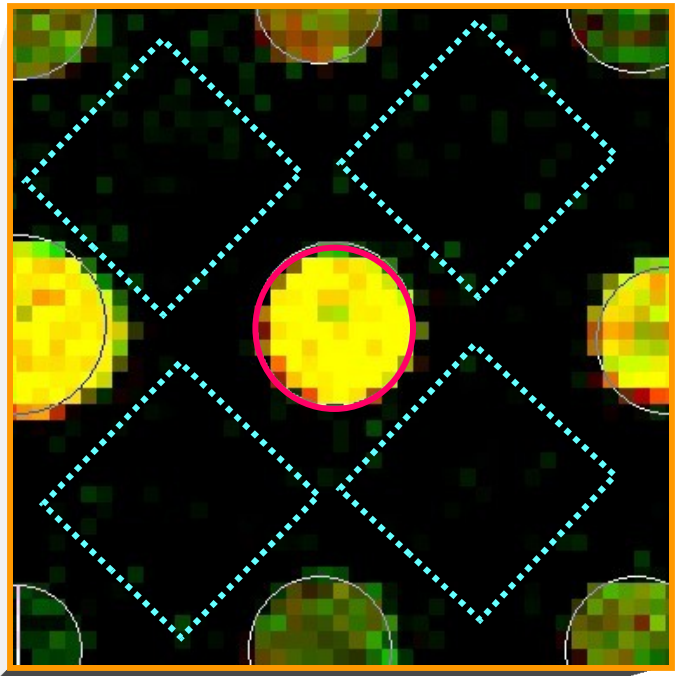
Image Analysis

1. Location

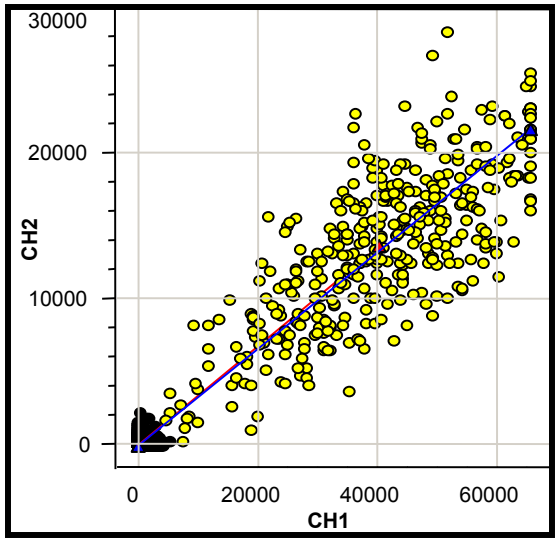


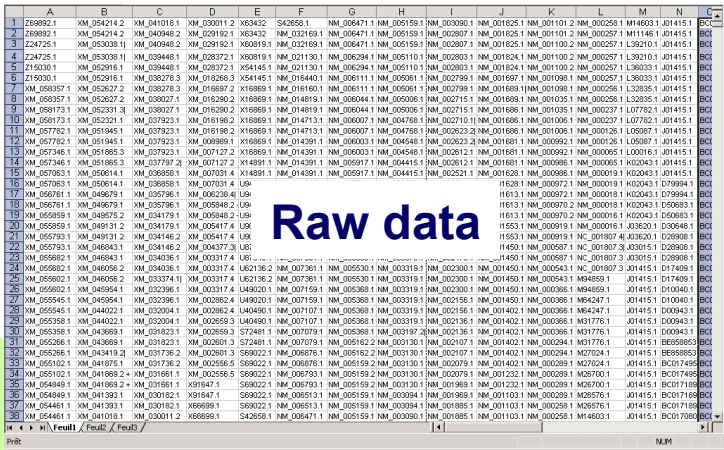


2. Segmentation



3. Quantification





Raw data

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:** Cy3 signal and background intensities
- **R, Rb:** Cy5 signal and background intensities

BioC Task View: 129 packages

http://www.bioconductor.org/packages/bioc/1.8/TwoChannel.html

Subview of

- Microarray

Packages in view

Package	Maintainer	Title
arrayQuality	A. Paquet	Assessing array quality on spotted arrays
bridge	Raphael Gottardo	Bayesian Robust Inference for Differential Gene Expression
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.
makePlatformDesign	Benilton Carvalho	Platform Design Package
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
oligo	Benilton Carvalho	Oligonucleotide Arrays
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
vsn	Wolfgang Huber	Variance stabilization and calibration for microarray data

Import data...

- *using limma* package
- limmaUsersGuide()

```
library("limma")
targets <- readTargets("Targets.txt")
RG  <- read.maimages(targets$FileName, source="genepix")
```

...into RGList

```
> names(RG)
"R"        "G"    "Rb"   "Gb"   "targets"     "genes"
"source"   "printer"

> RG$R[1,]
          s1      s2     s3     s4
[1, ]      6207  39167 6696  6000
```

Quality Assessment vs Control

- Quality Assessment: computation and interpretation of metrics that are intended to measure quality
- Quality Control: possible subsequent action, removing bad array or re-doing part of the experiment

Quality Assessment

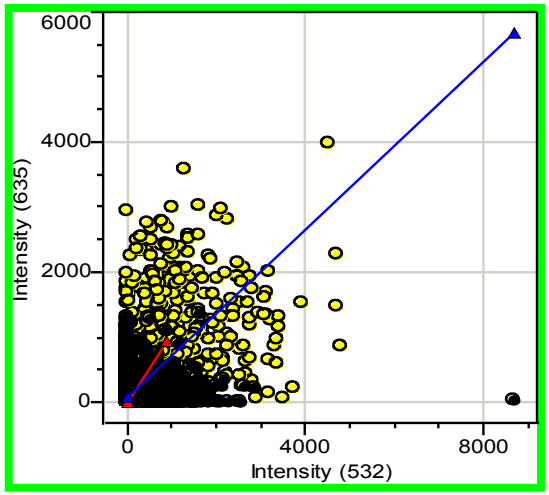
For each array:

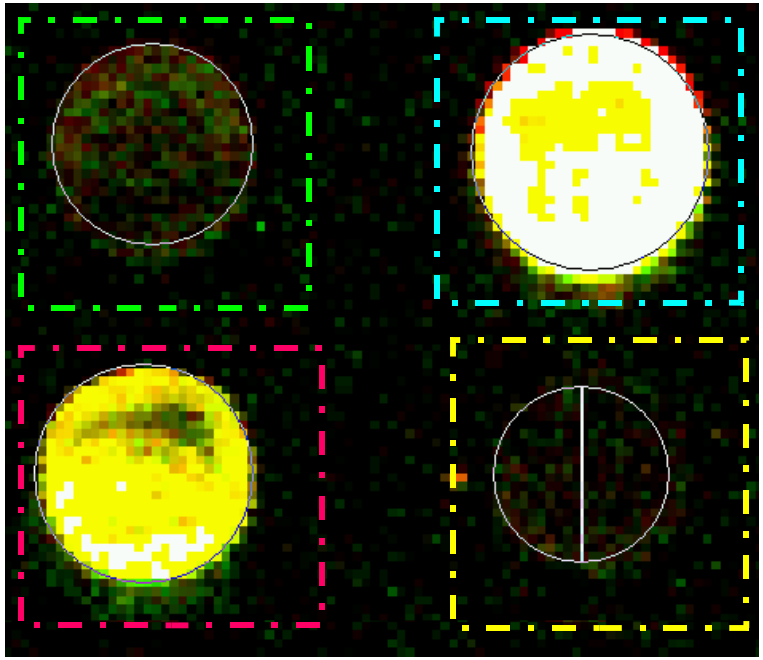
- **Diagnostics plots** of spot statistics

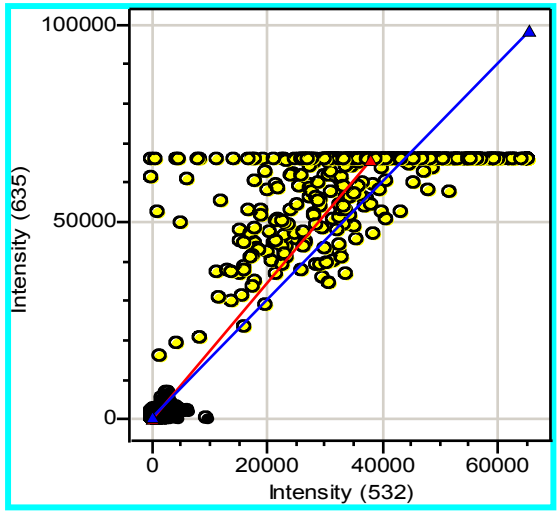
  e.g. R and G log-intensities, M, A, spot area.

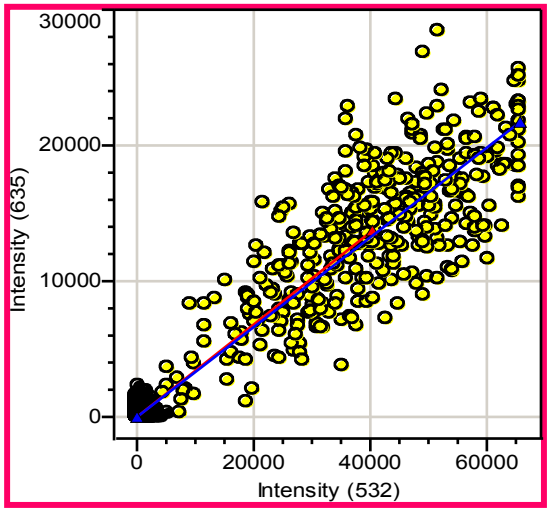
  - Boxplots;
  - 2D spatial images;
  - Scatter-plots, e.g. MA-plots;
  - Density plots.
- **Stratify** plots according to layout parameters, *e.g.* print-tip-group, plate.
- **summary statistics**

Quality Filtering

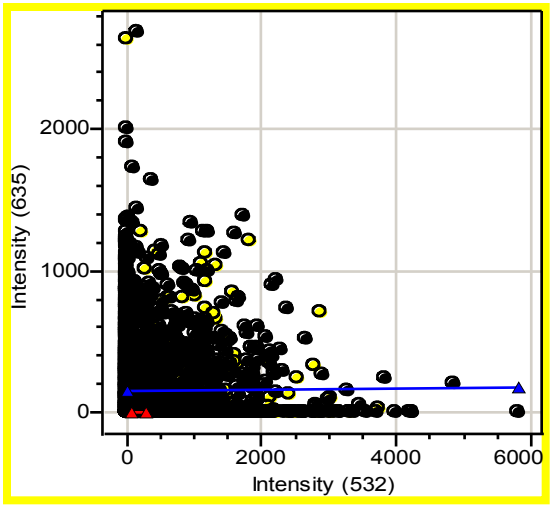




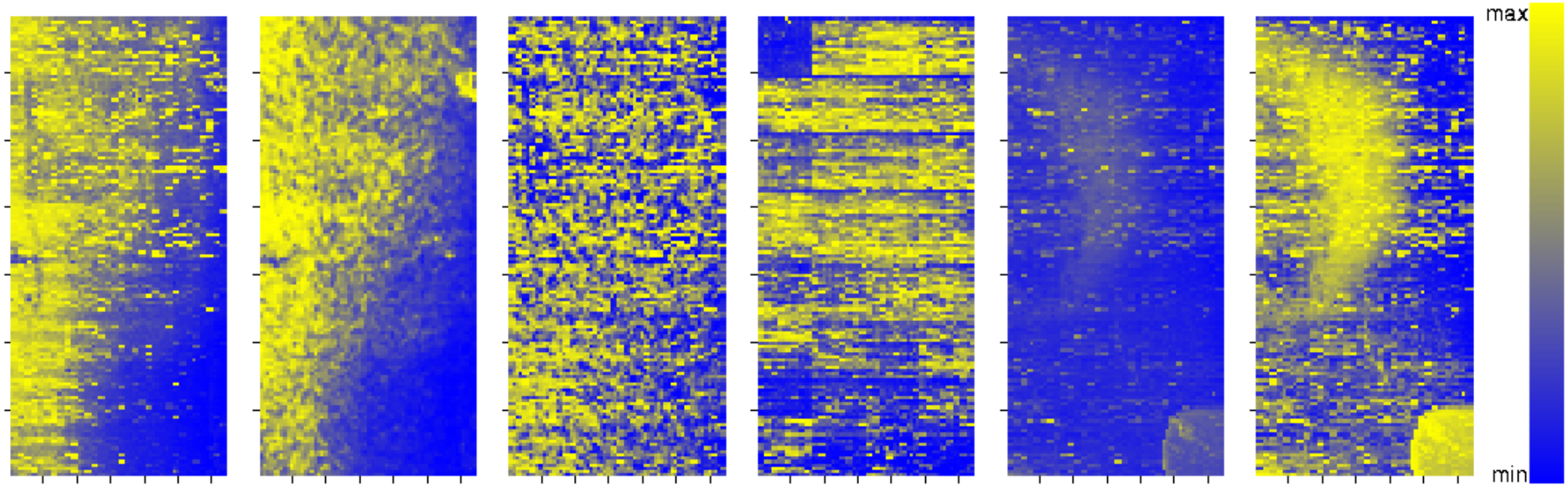




- Background
- Foreground



Spatial Effects – Image Plots



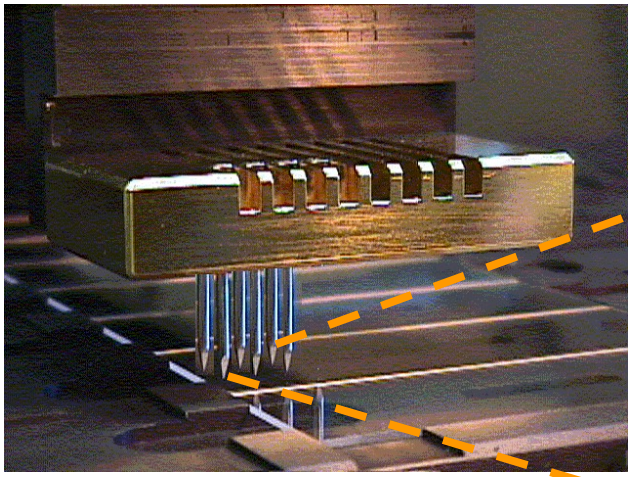
R **Rb** **R-Rb**
color scale by rank

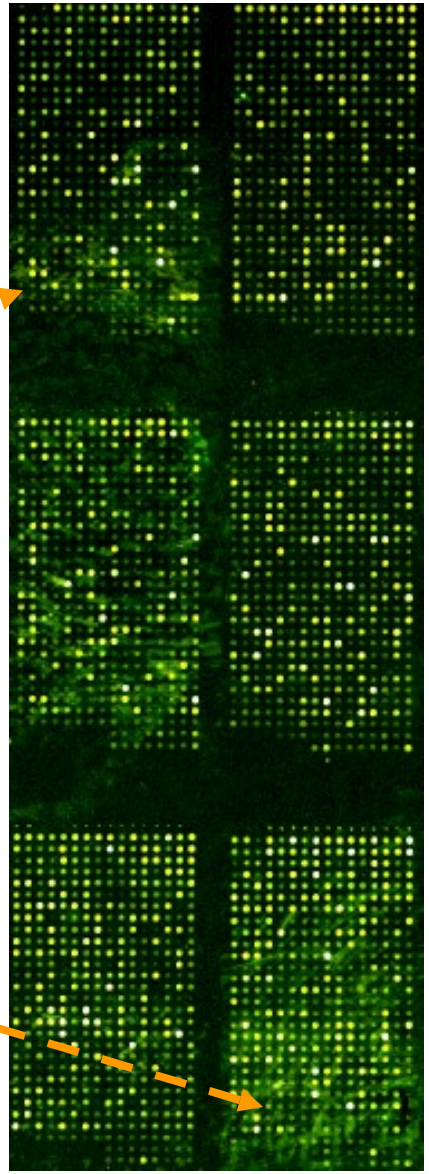
another array: print-tip

color scale ~ log(G)

color scale ~ rank(G)

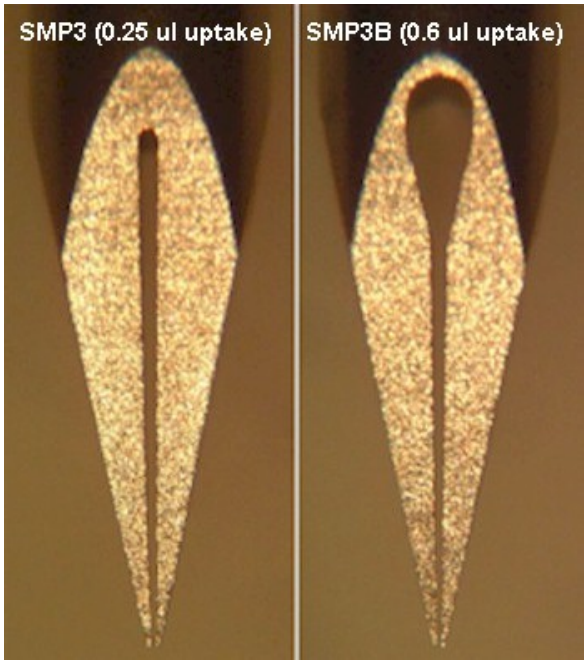
Spatial Effects

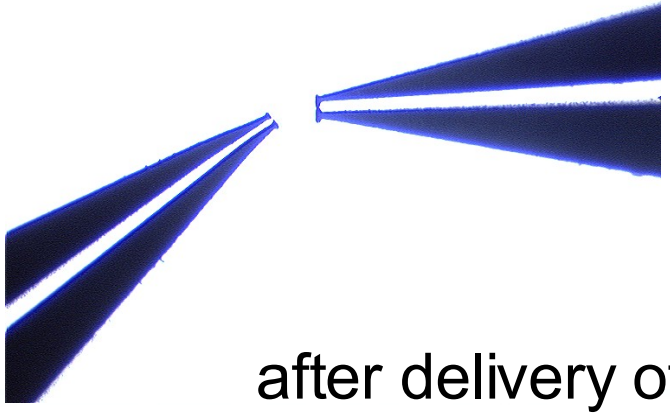




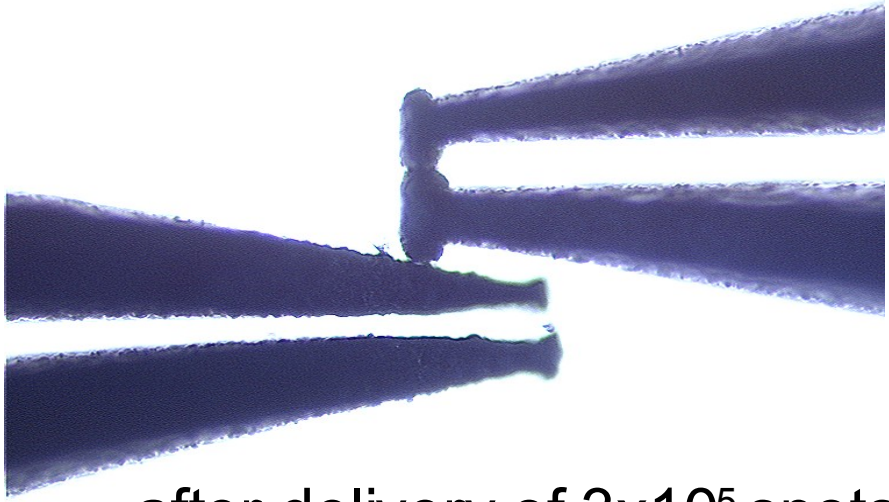
1 pin → 1 block

Spotting Pin Quality Decline





after delivery of $5x10^5$ spots

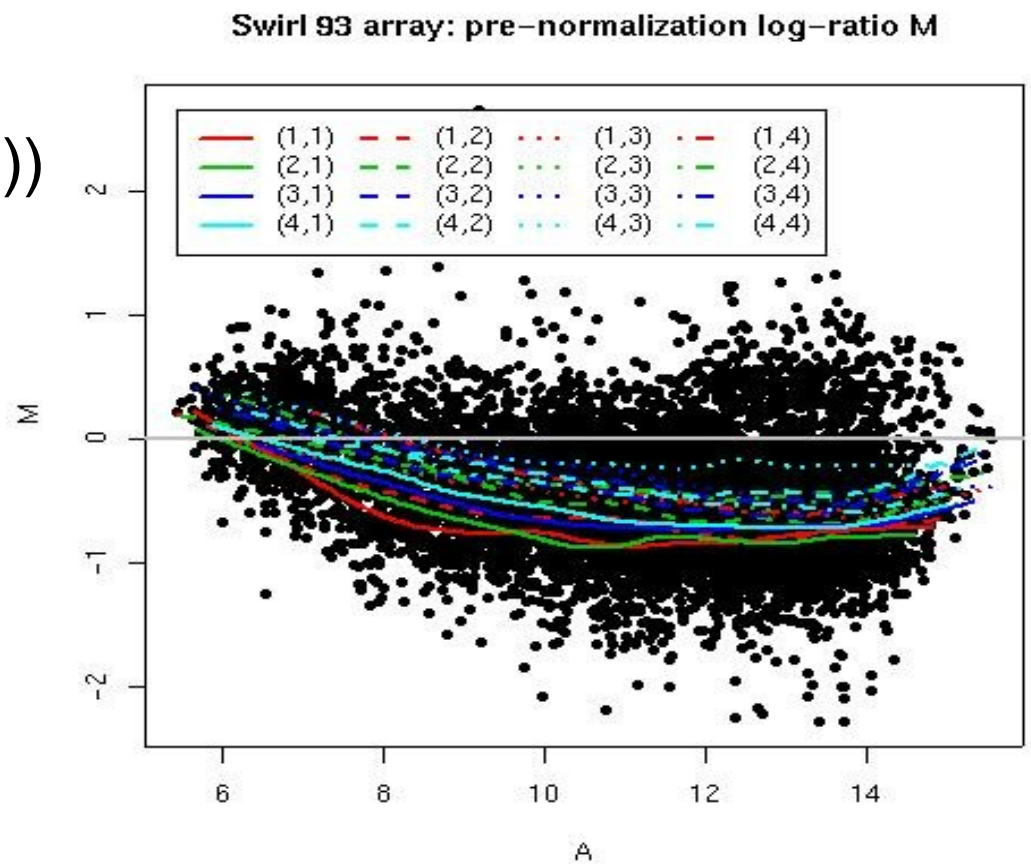


after delivery of $3x10^5$ spots

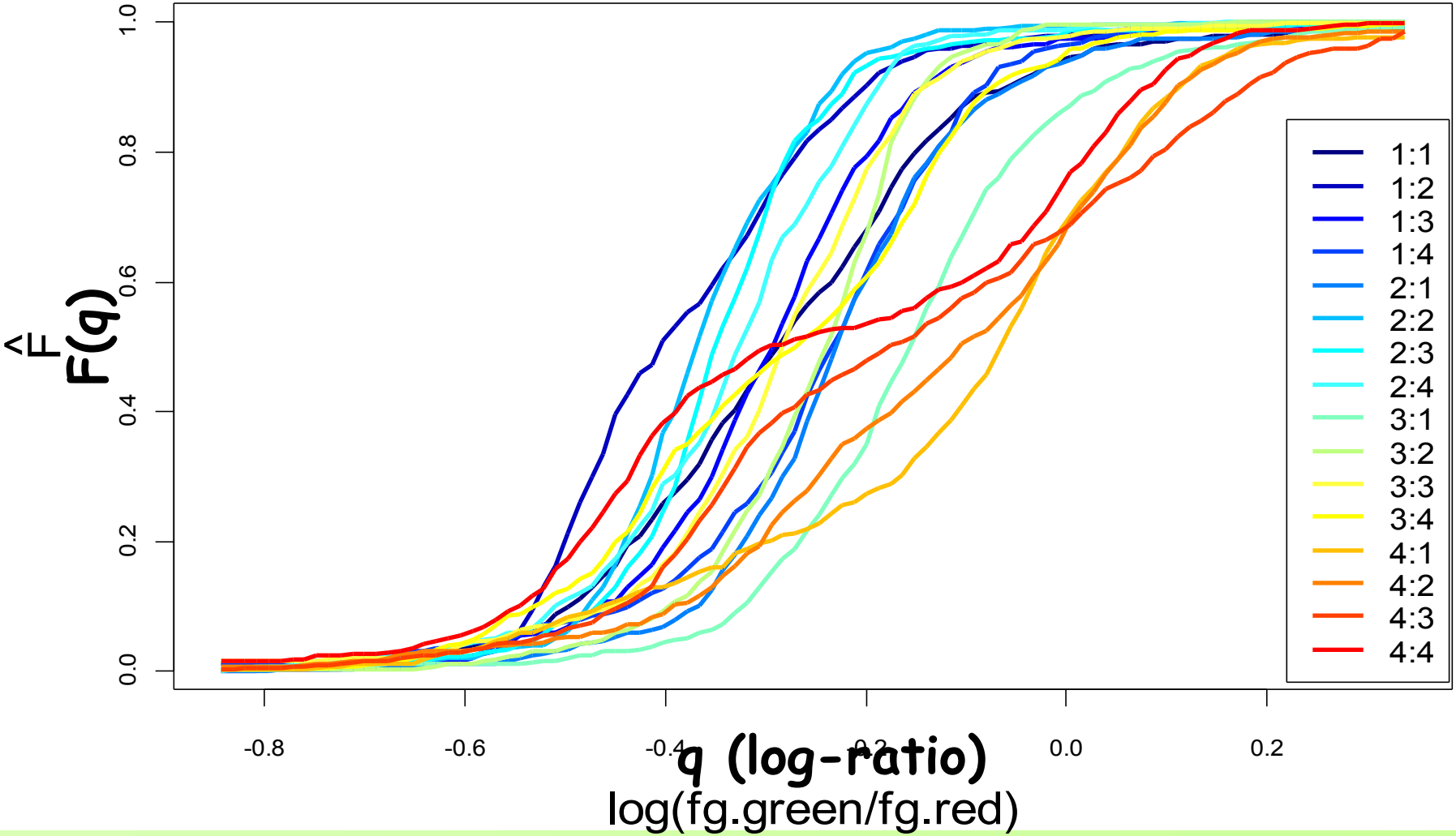
MA plot

M = log2(R) - log2(G)

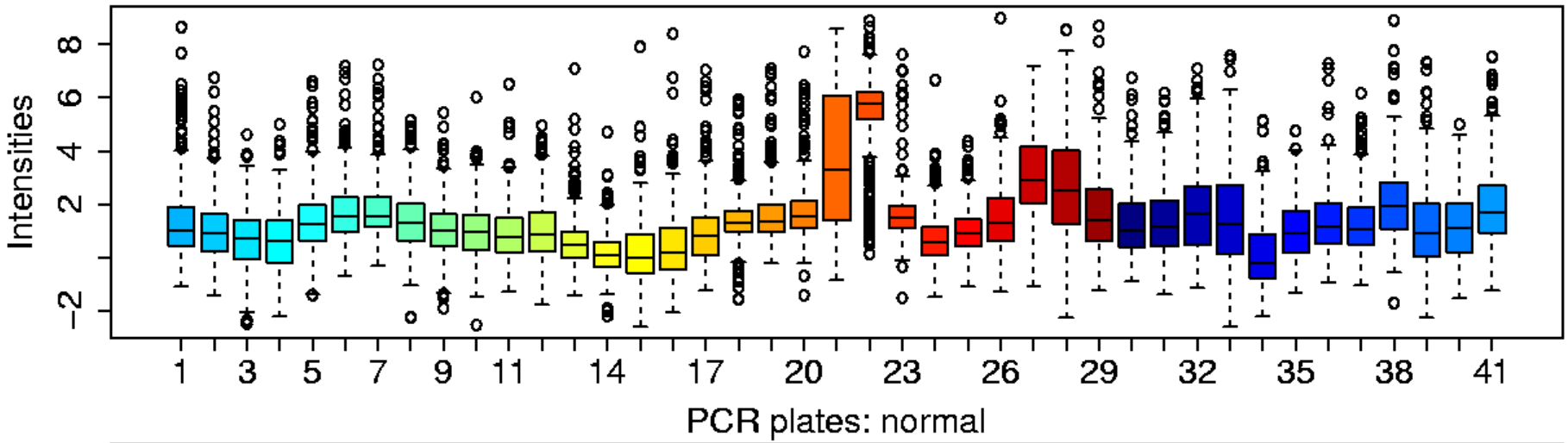
A = 1/2(log2(R) + log2(G))

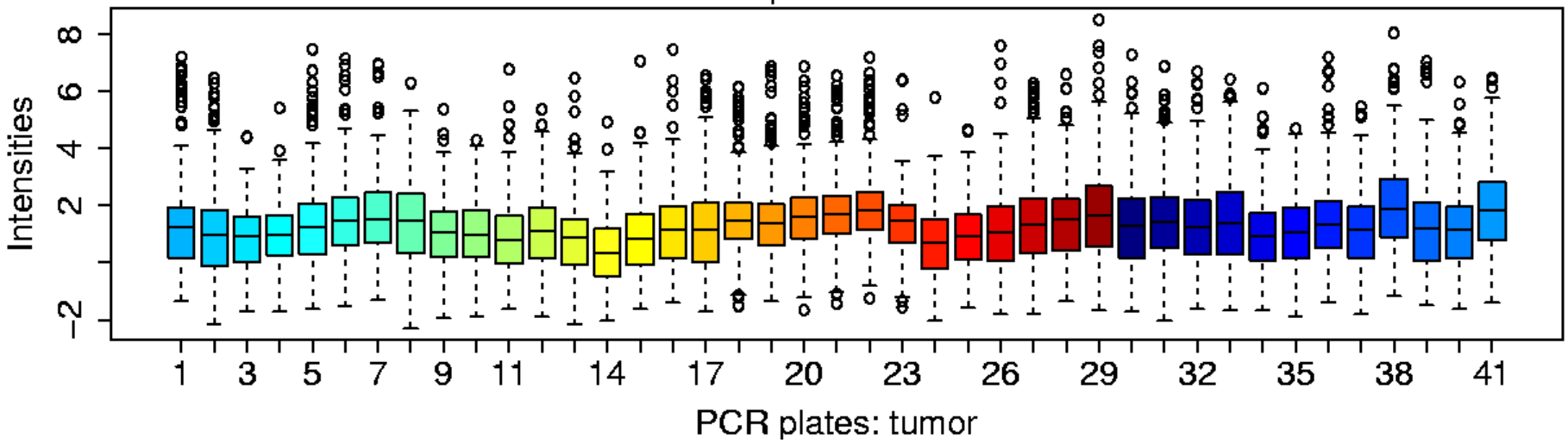


Print-tip Effects – ECDF plot



PCR Plates - Boxplots

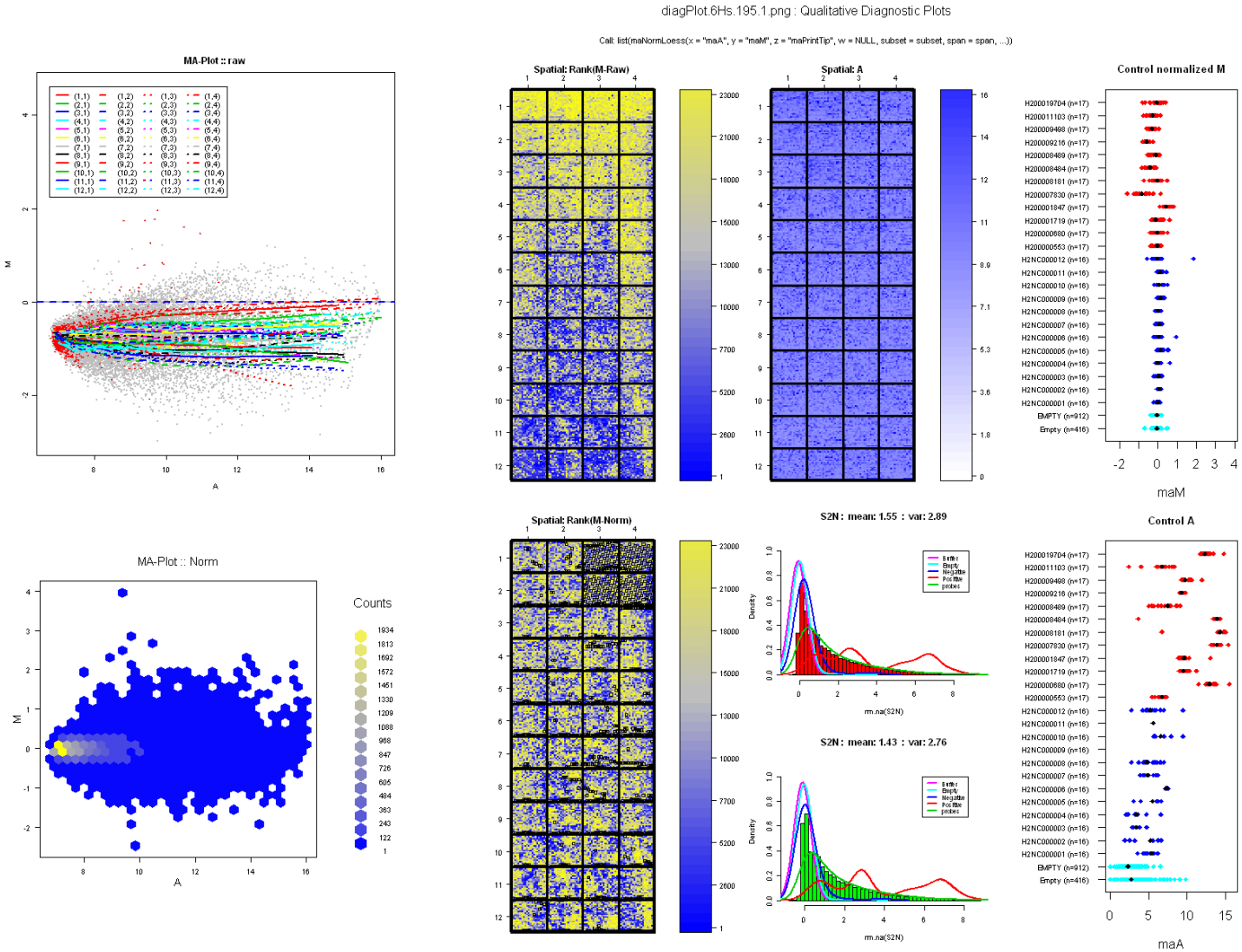




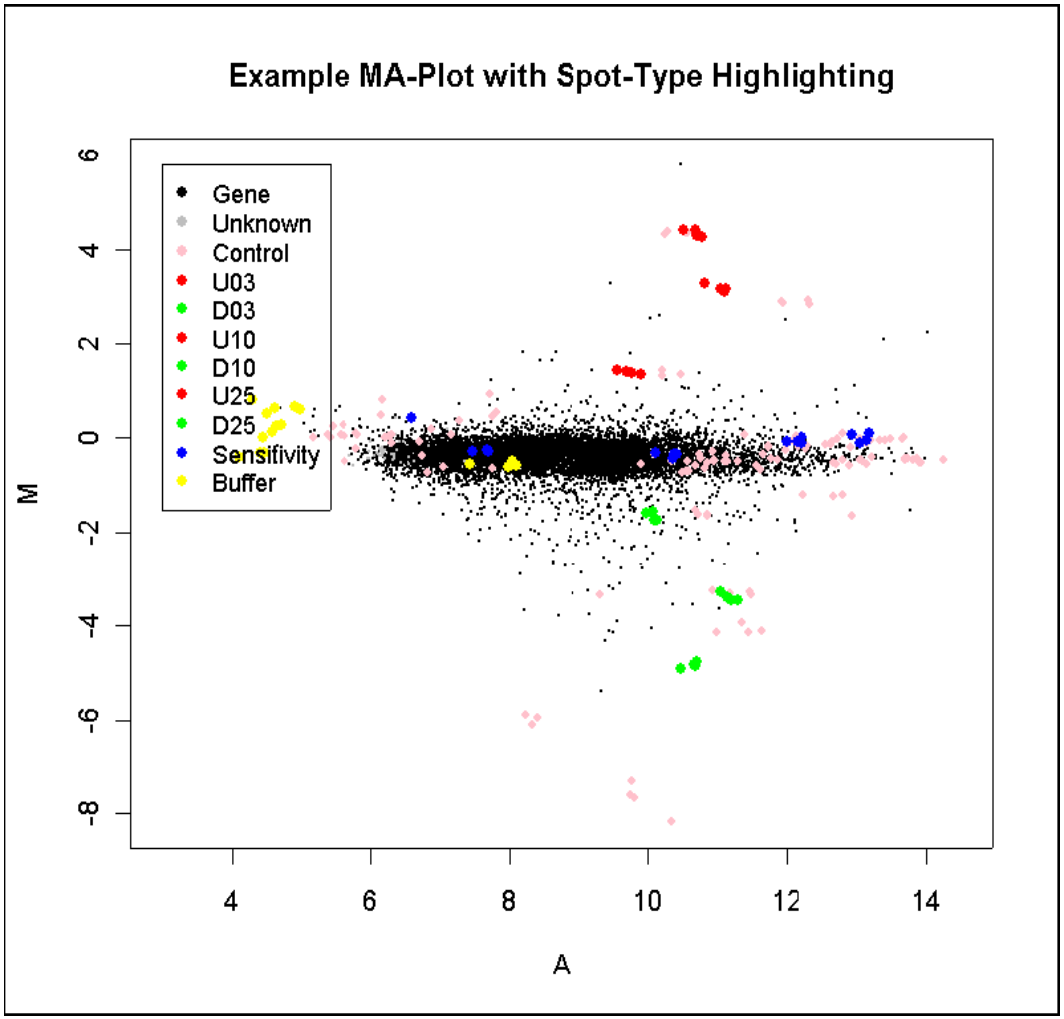
Bioconductor QA packages

- *Several packages*
  - *arrayMagic*
  - *arrayQuality*
  - *arrayQualityMetrics*
- *pdf or html*

Diagnostic plot with *arrayQuality*



Data Exploration with *limma*



(Limma user Guide)

Quality Assessment: Summary

For each array:

- Diagnostics plots
- Stratify

BioC packages:

- *arrayQuality*
- *arrayMagic*
- …

Outline

- **Data acquisition**
- **Pre-processing**
  - Quality assessment
  - Pre-processing
    - background correction
    - normalization
    - summarization

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”

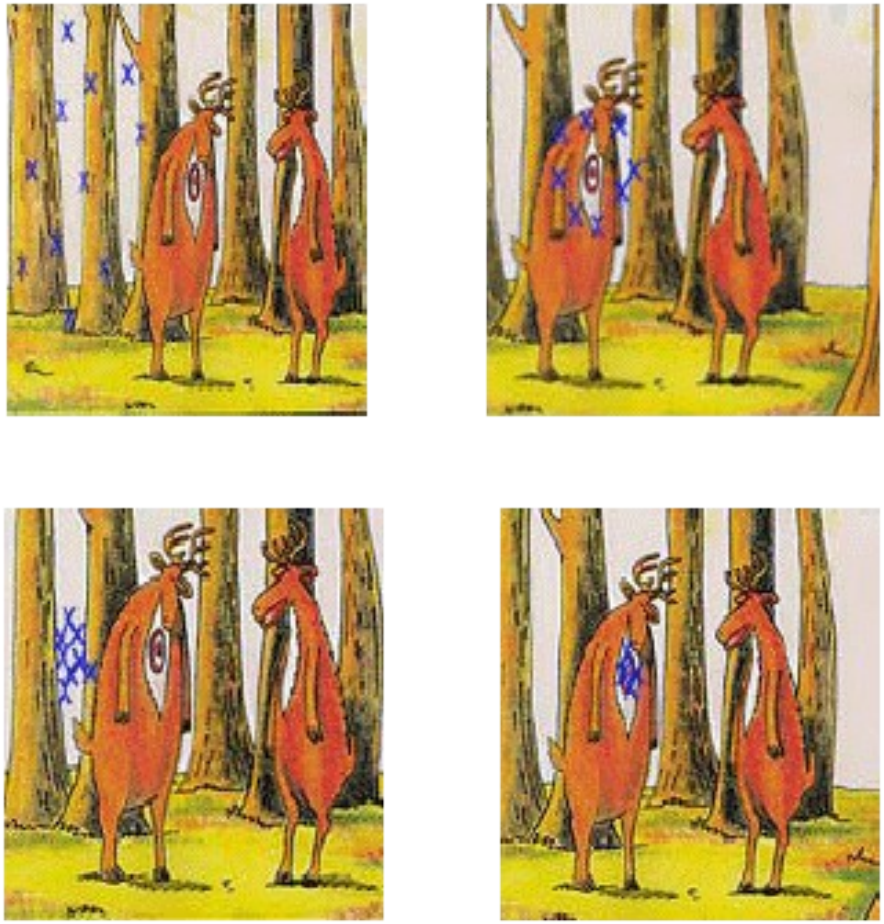
Error Model

Variance-Bias trade off

<- Bias Accuray ->

Variance ->

<- Precision

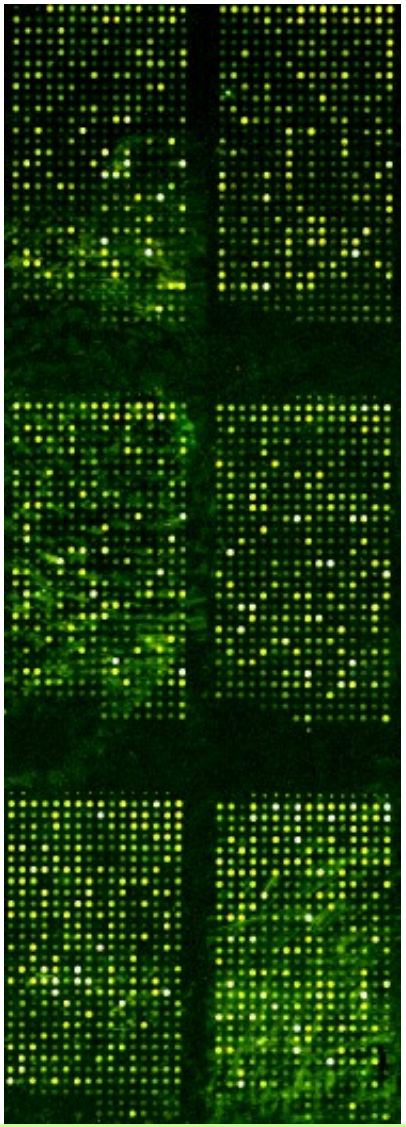


Background Correction

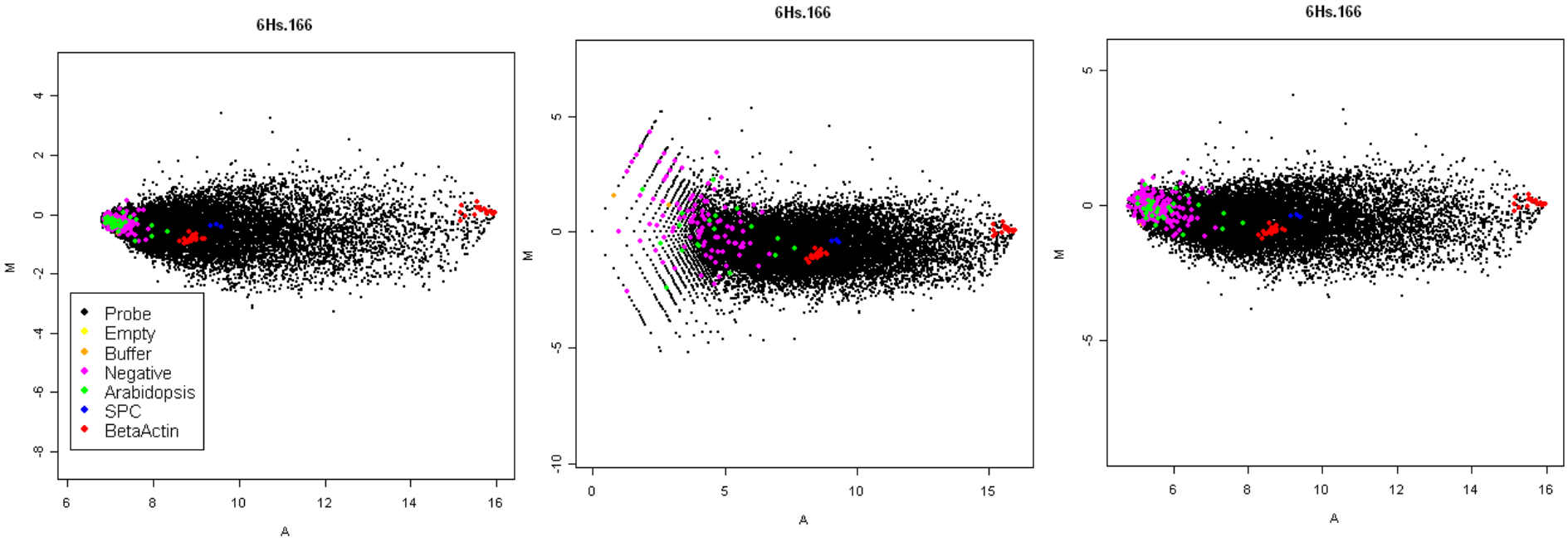
- none
- subtraction, movingmin
- *Minimun,edwards, normexp,…*

- More details … *limma*

  >?backgroundCorrect



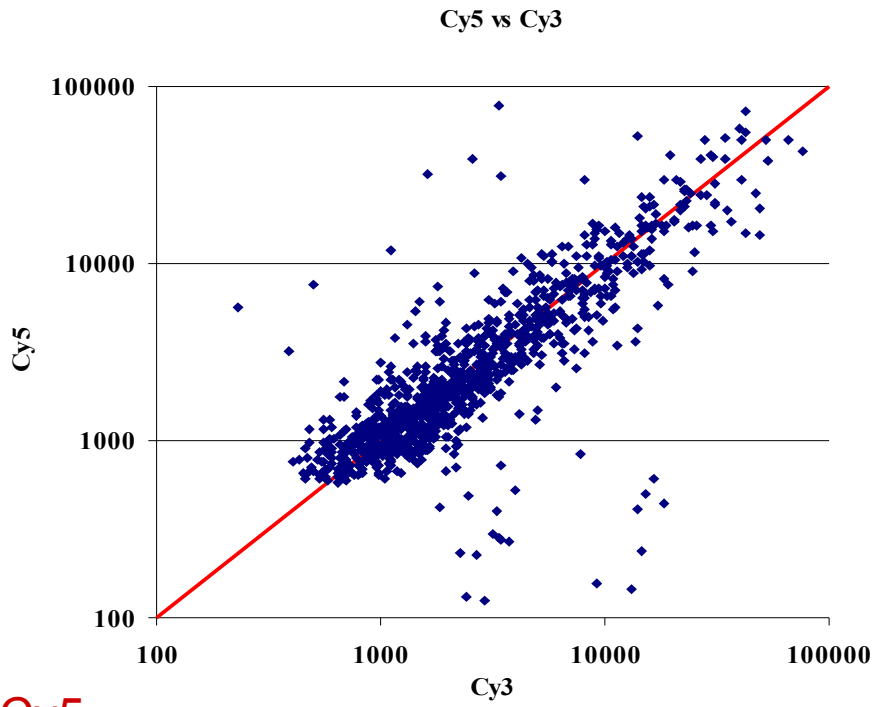
Background Correction

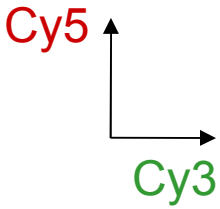


none substraction *normexp*

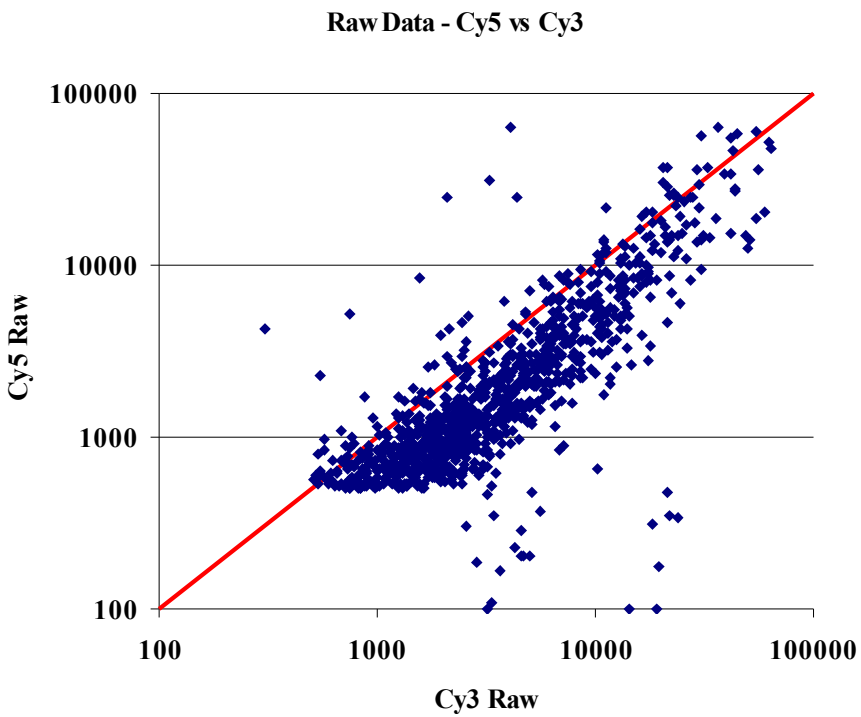
Why Normalize?

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.

Data Transformation

measured intensity = offset + gain × true abundance

$$Y_{ik} = B_{ik} + \alpha_{ik} S_k$$

Example: log transformation

- Intensity measurements adapt a distribution that is closer to the normal distribution
- Muliplicative noise becomes additive noise: variance more independent of intensity

Normalization methods

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

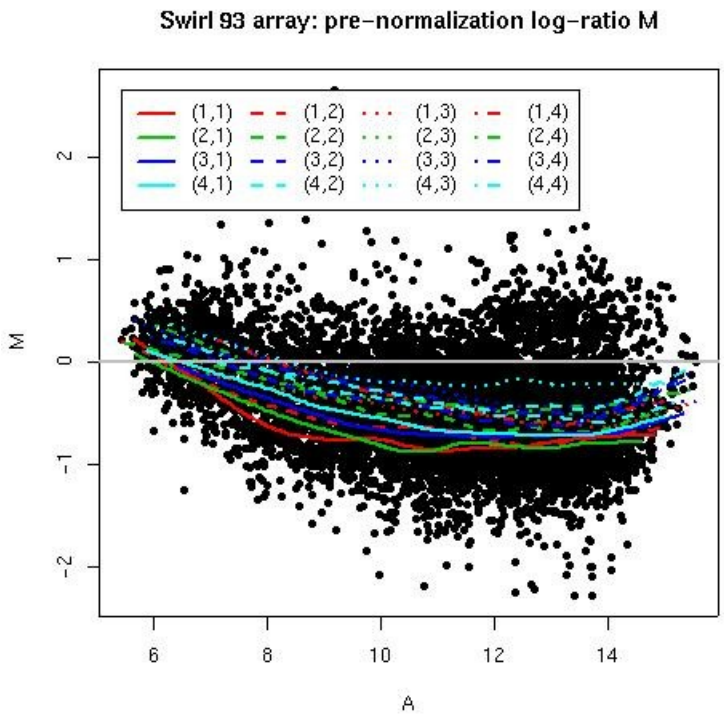
Two-channel

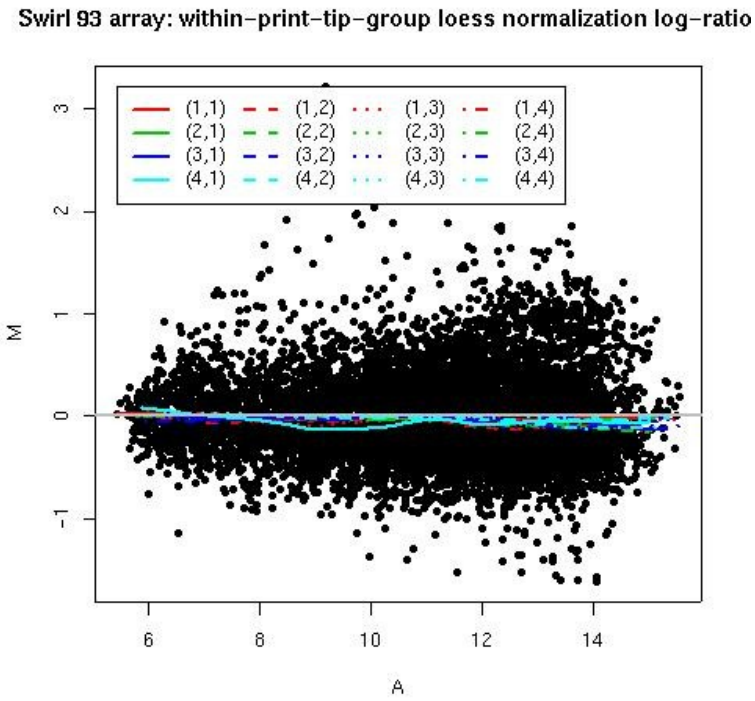
Separate-channel

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



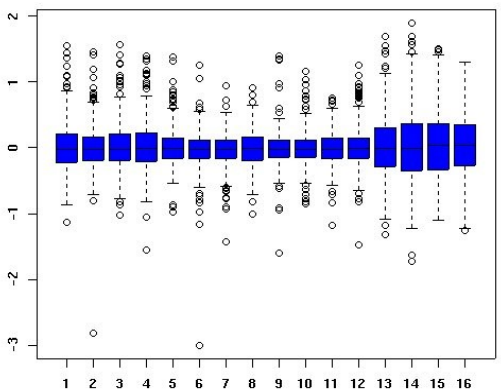


Two channel normalization

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

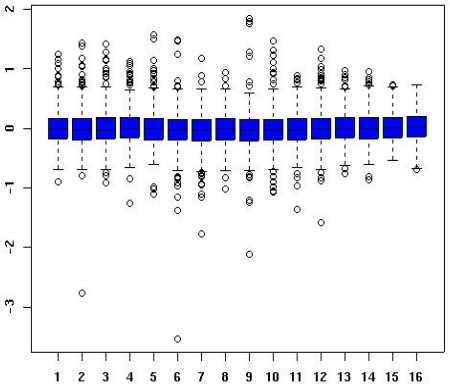
- Scale: adjust for different in scale between multiple arrays

median centered



Scaling

median centered & MAD scaled



One channel normalization

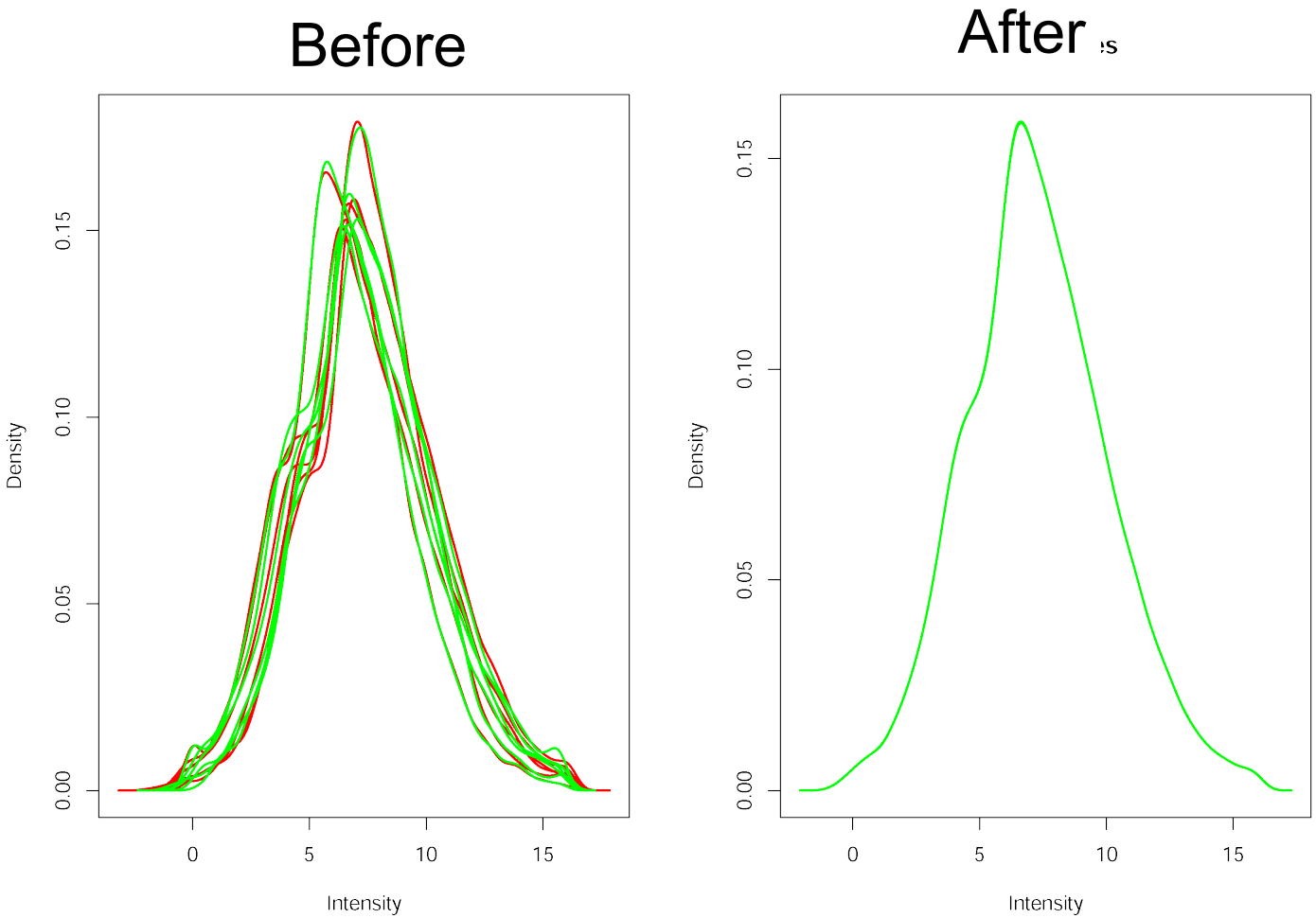
- As technology improves the spot-to-spot varation is reduced

- Development of normalization techniques that work on the absolute intensities

Ex: quantile normalization (*limma*)

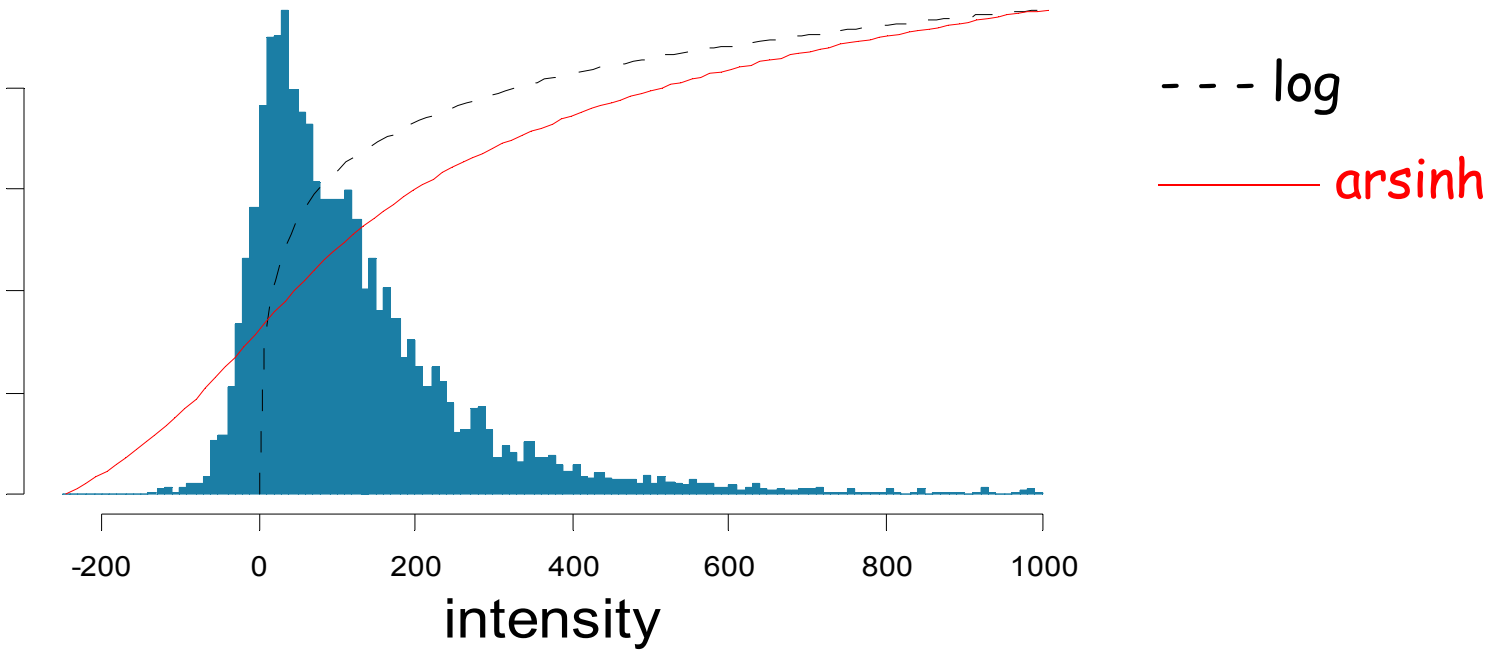
variance stabilization (*vsn*)

Quantile Normalization



Bolstand *et al*.(2003)

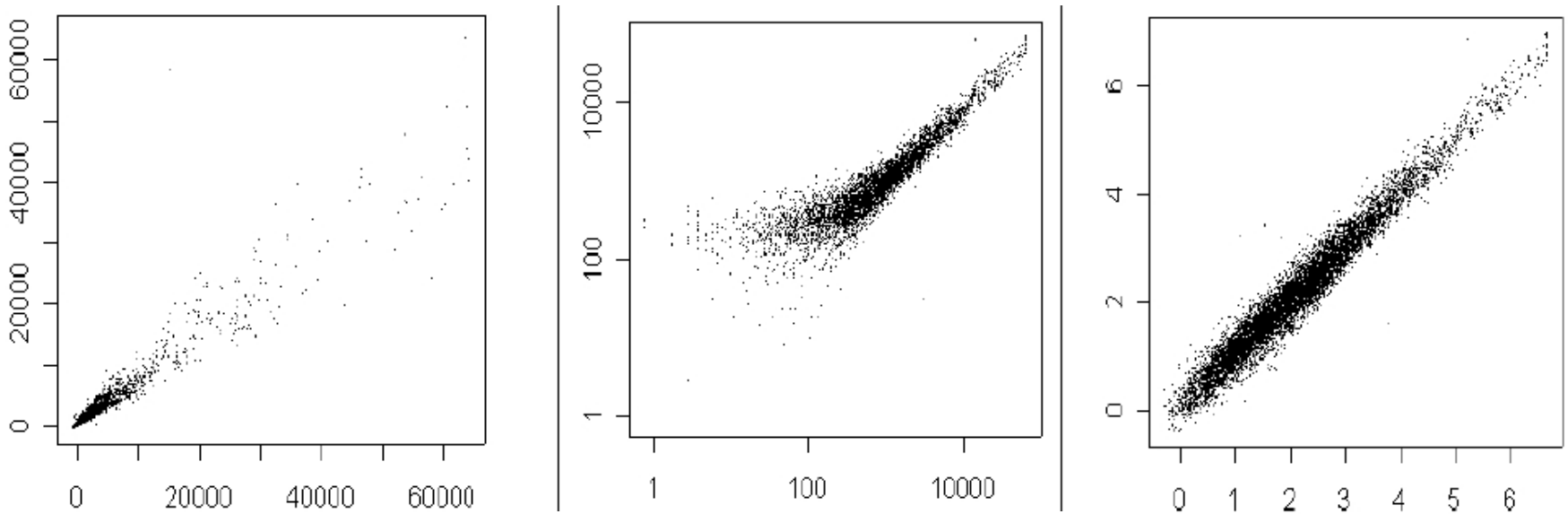
Variance Statibilizing Transformation



- Meaningful around 0
- Original intensities may be negatives

(Huber *et al.* 2004)

Variance stabilization (*vsn*)



linear log arsinh

Variance stabilization (*vsn*)

log-ratio

$$\log \frac{x_i}{x_j}$$

'glog' (generalized log-ratio)

$$\log \frac{x_i + \sqrt{x_i^2 + c_i^2}}{x_j + \sqrt{x_j^2 + c_j^2}}$$

- interpretation as "fold change"

+ interpretation even in cases where genes are off in some conditions (negative values)

+ visualization

+ can use standard statistical methods (hypothesis testing, ANOVA, clustering, classification…) without the worries about low-level variability that are often warranted on the log-scale

Preprocessing : Summary

For each array:

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

BioC pacakges:

- *marray*
- *limma*
- …