



cDNA Microarray Analysis with BioConductor packages

Nolwenn Le Meur

Copyright 2007

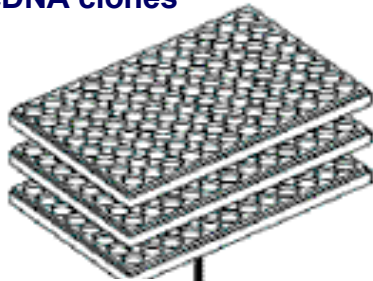
Outline

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

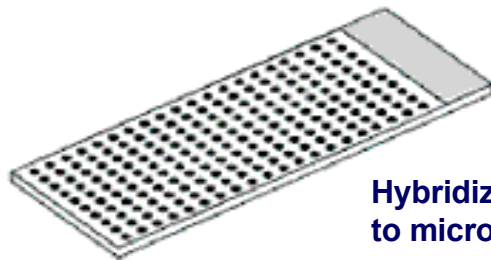
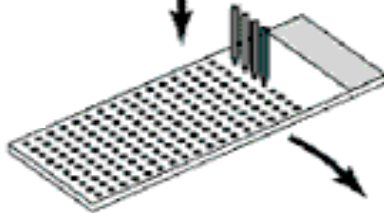
Two-color Microarray

Probe (gene reporter)

Oligonucleotides or cDNA clones

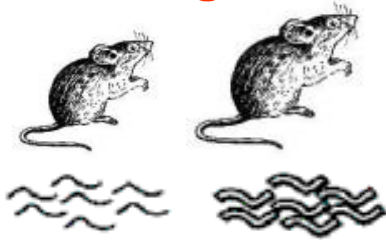


Spotting

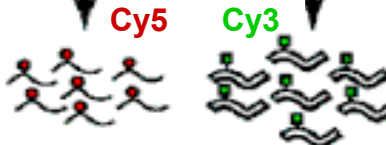


Hybridize target to microarray

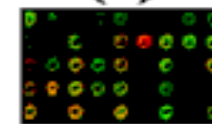
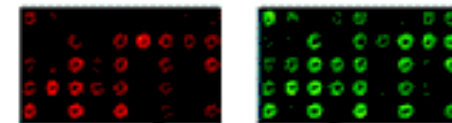
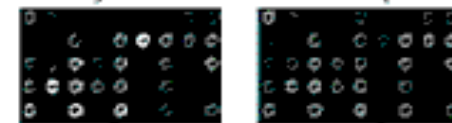
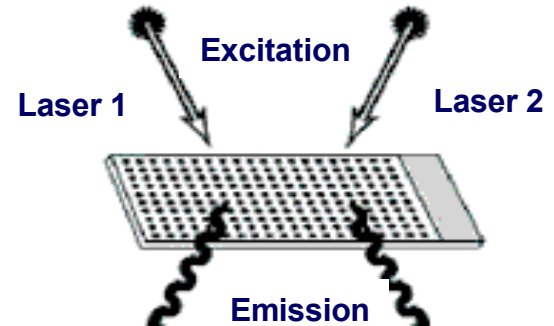
Target



Label with Fluor dyes



Scan



Computer analysis

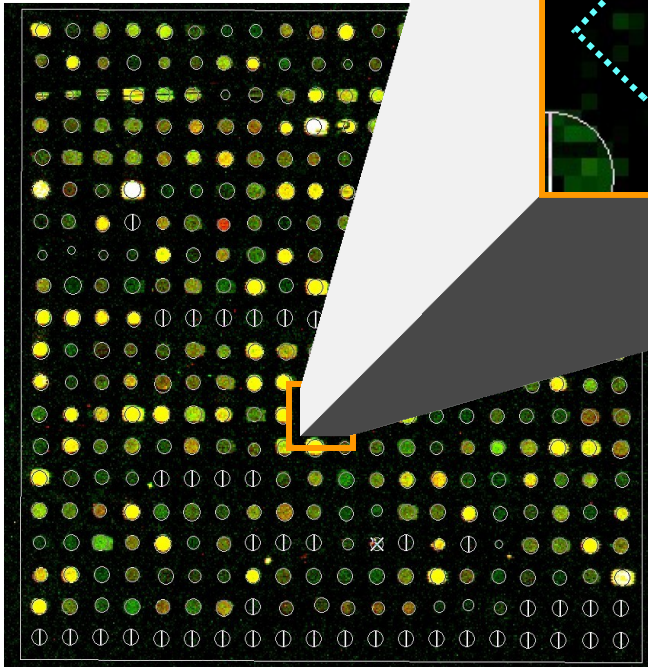
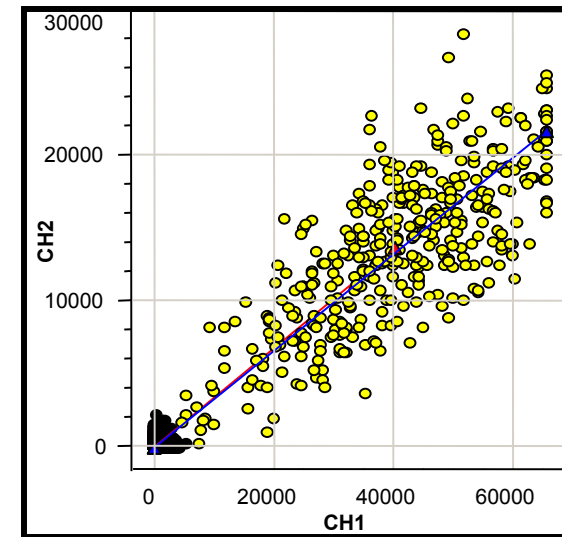
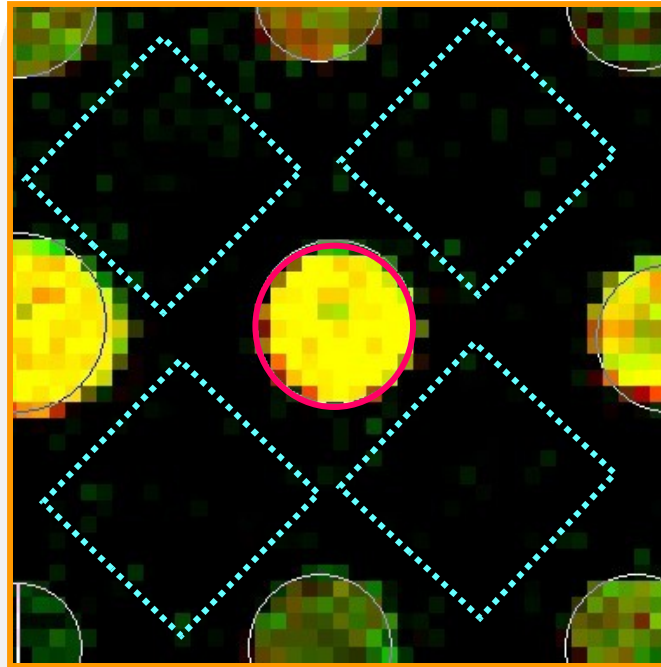
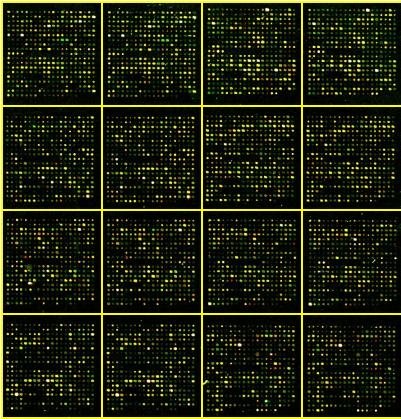
(adapted from Duggan et al., Nat. Gen., 1999)

Image Analysis

1. Location

2. Segmentation

3. Quantification



Raw data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	G
1	269822.1	XM_054214.2	XM_041018.1	XM_030011.2	863432	542658	NM_006471.1	NM_005159.1	NM_003090.1	NM_001625.1	NM_001101.2	NM_000258.1	M14603.1	J01415.1	BCC1
2	269822.1	XM_042474.2	NM_040448.2	XM_020191.1	163432	NM_023169.1	NM_006471.1	NM_000207.1	NM_001625.1	NM_001101.2	NM_000257.1	M11148.1	J01415.1	BCC1	
3	224725.1	XM_050308.1	NM_040948.2	NM_020191.1	163432	NM_023169.1	NM_006471.1	NM_000207.1	NM_001625.1	NM_001101.2	NM_000257.1	M11148.1	J01415.1	BCC1	
4	224725.1	XM_050308.1	NM_039448.1	XM_020372.1	X60819.1	NM_021130.1	NM_006294.1	NM_005110.1	NM_002003.1	NM_001824.1	NM_001100.2	NM_000257.1	L39210.1	J01415.1	BCC1
5	210300.1	XM_052916.1	NM_039448.1	XM_020372.1	X64145.1	NM_021130.1	NM_006294.1	NM_005110.1	NM_002003.1	NM_001824.1	NM_001100.2	NM_000257.1	L36033.1	J01415.1	BCC1
6	210300.1	XM_052916.1	NM_030278.2	XM_016266.3	X64145.1	NM_016460.1	NM_006111.1	NM_005081.1	NM_002799.1	NM_001691.1	NM_001098.1	NM_000257.1	L36033.1	J01415.1	BCC1
7	XM_056357.1	XM_050627.2	XM_038278.3	XM_016887.2	X18989.1	NM_016160.1	NM_006111.1	NM_005081.1	NM_002799.1	NM_001691.1	NM_001098.1	NM_000256.1	L32835.1	J01415.1	BCC1
8	XM_050357.1	XM_050627.2	XM_038027.1	XM_016262.2	X18989.1	NM_016160.1	NM_006111.1	NM_005081.1	NM_002799.1	NM_001691.1	NM_001098.1	NM_000256.1	L32835.1	J01415.1	BCC1
9	XM_056173.1	XM_052331.3	XM_016262.2	X18989.1	NM_016262.2	X18989.1	NM_005444.1	NM_005081.1	NM_002799.1	NM_001691.1	NM_001098.1	NM_000257.1	L07762.1	J01415.1	BCC1
10	XM_056173.1	XM_052321.1	XM_037923.1	XM_016186.2	X18989.1	NM_014713.1	NM_006007.1	NM_004788.1	NM_002710.1	NM_001698.1	NM_001098.1	NM_000257.1	L07762.1	J01415.1	BCC1
11	XM_057762.1	XM_051945.1	XM_037923.1	XM_016186.2	X18989.1	NM_014713.1	NM_006007.1	NM_004788.1	NM_002710.1	NM_001698.1	NM_001098.1	NM_000257.1	L07762.1	J01415.1	BCC1
14	XM_057346.1	XM_051895.3	XM_037762.1	XM_007127.2	X14891.1	NM_014391.1	NM_006003.1	NM_004548.1	NM_002621.1	NM_001691.1	NM_000999.1	NM_000065.1	L00016.1	J01415.1	BCC1
13	XM_057346.1	XM_051895.3	XM_037923.1	XM_007127.2	X14891.1	NM_014391.1	NM_006003.1	NM_004548.1	NM_002621.1	NM_001691.1	NM_000999.1	NM_000065.1	L00016.1	J01415.1	BCC1
15	XM_057093.1	XM_050611.4	XM_036558.1	XM_007031.4	L9										BCC1
16	XM_057093.1	XM_050611.4	XM_036558.1	XM_007031.4	L9										BCC1
17	XM_056761.1	XM_048978.1	XM_035786.1	XM_005928.4	L9										BCC1
18	XM_056761.1	XM_048978.1	XM_035786.1	XM_005928.4	L9										BCC1
19	XM_056989.1	XM_048575.2	XM_034179.1	XM_005946.2	L9										BCC1
20	XM_056989.1	XM_048575.2	XM_034179.1	XM_005946.2	L9										BCC1
21	XM_055793.1	XM_049131.2	XM_034146.2	XM_005417.4	L9										BCC1
22	XM_055793.1	XM_048843.1	XM_034146.2	XM_004377.3	L9										BCC1
23	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
24	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
25	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
26	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
27	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
28	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
29	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
30	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
31	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
32	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
33	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
34	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
35	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
36	XM_055821.1	XM_048843.1	XM_034036.1	XM_003917.4	L9										BCC1
37	XM_054461.1	XM_041393.1	XM_030162.1	X68899.1	S89022.1	NM_009513.1	NM_005159.1	NM_000394.1	NM_001895.1	NM_001103.1	NM_000209.1	M26576.1	J01415.1	BCC1	
38	XM_054461.1	XM_041393.1	XM_030162.1	X68899.1	S89022.1	NM_009513.1	NM_005159.1	NM_000394.1	NM_001895.1	NM_001103.1	NM_000209.1	M26576.1	J01415.1	BCC1	
39	XM_054461.1	XM_041018.1	X68899.1	S42605.1	NM_004878.1	NM_005159.1	NM_000390.1	NM_001895.1	NM_001103.1	NM_000209.1	M47603.1	J01415.1	BCC1		

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



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
mmNorm	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

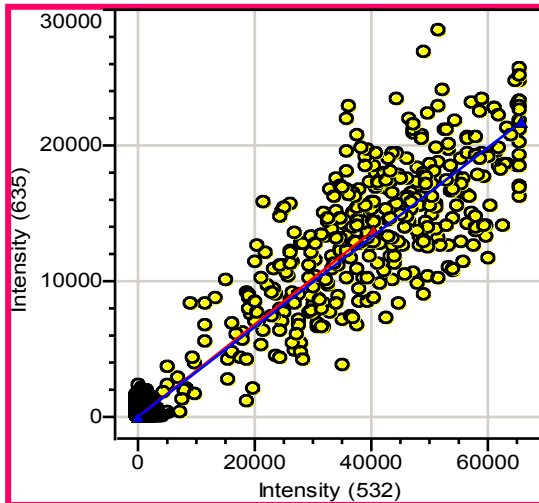
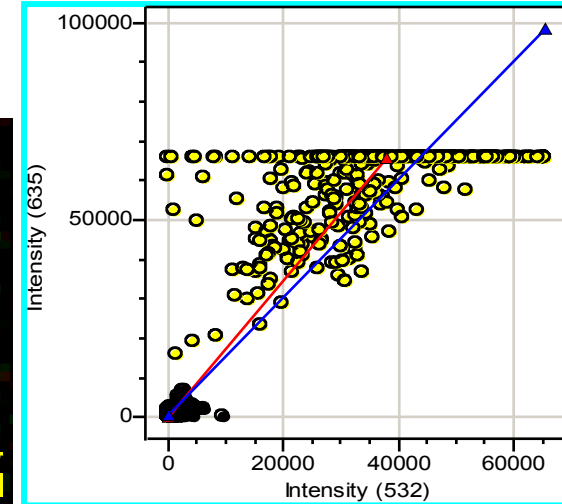
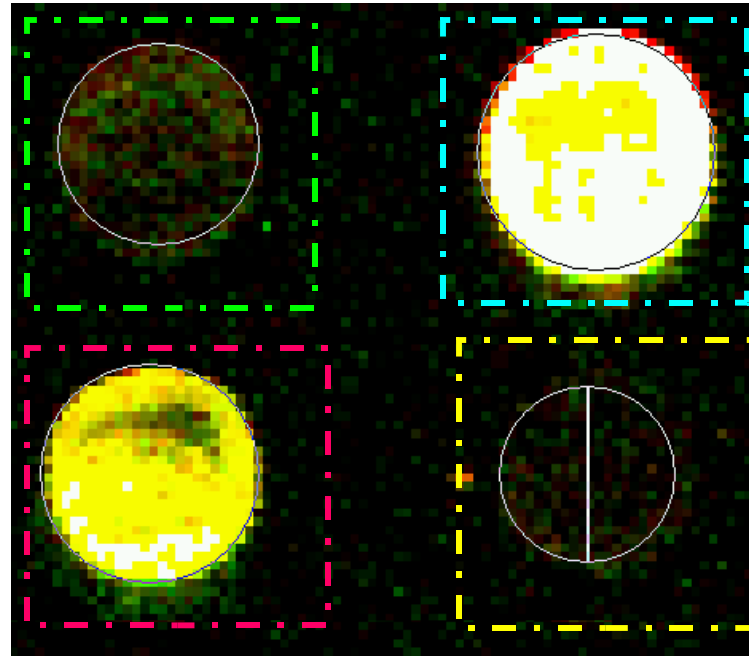
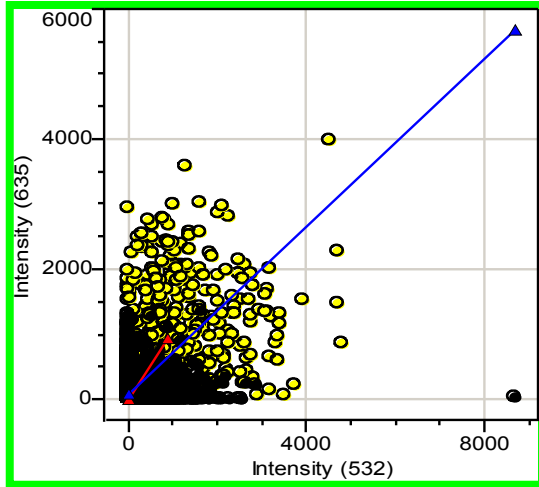
- **Quality Assessment:** computation and interpretation of metrics that are intended to measure quality
- **Quality Control:** possible subsequent action, removing bad array or re-doing part of the experiment

Quality Assessment

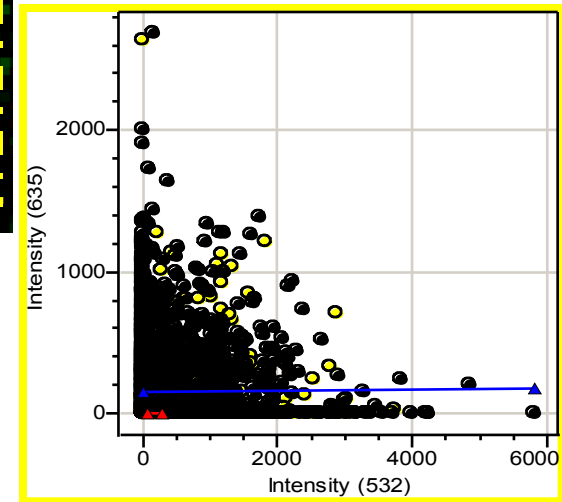
For each array:

- **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**

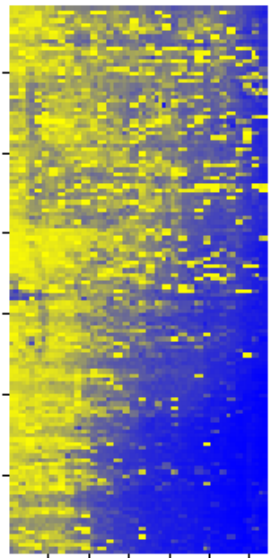
Quality Filtering



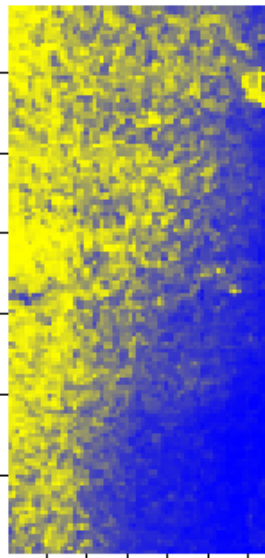
- Background
- Foreground



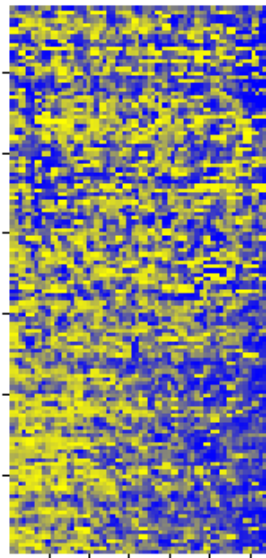
Spatial Effects – Image Plots



R

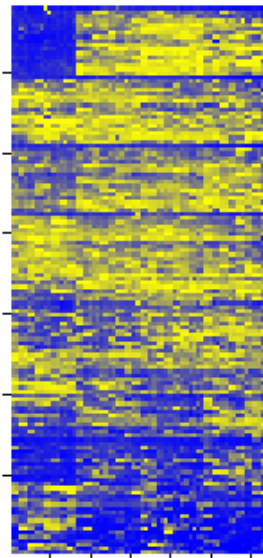


Rb

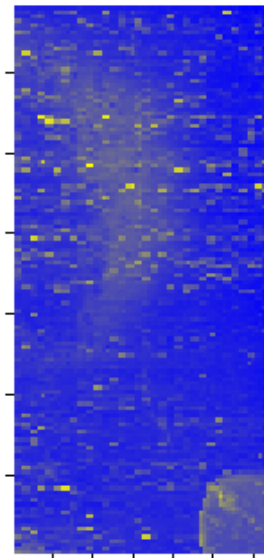


R-Rb

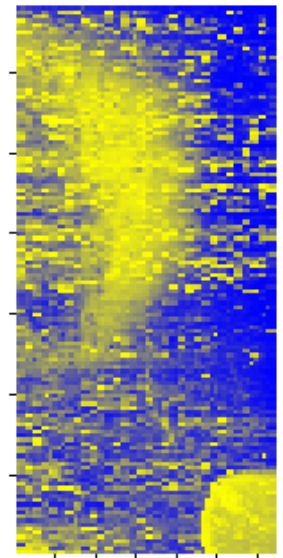
color scale by rank



another array:
print-tip



color scale ~
log(G)

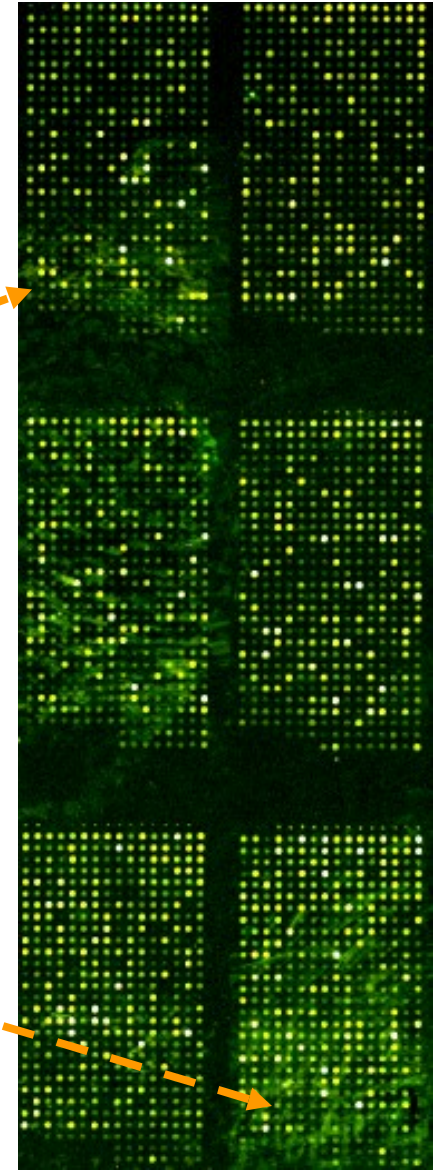
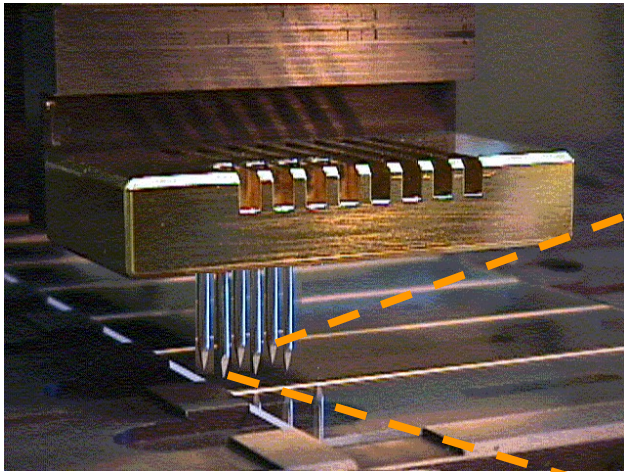



color scale ~
rank(G)

max

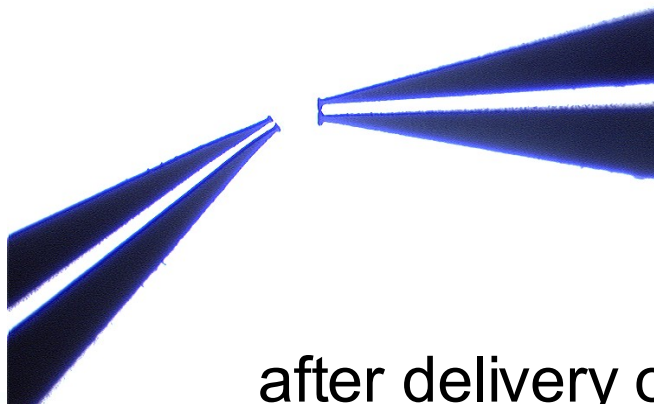
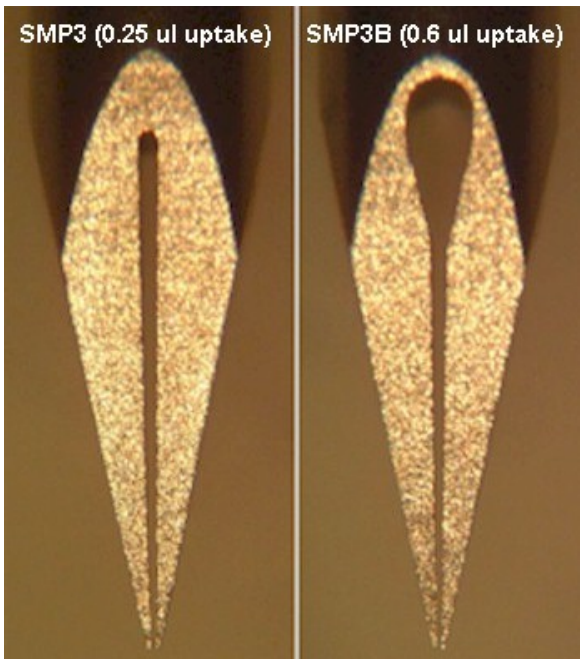
min

Spatial Effects

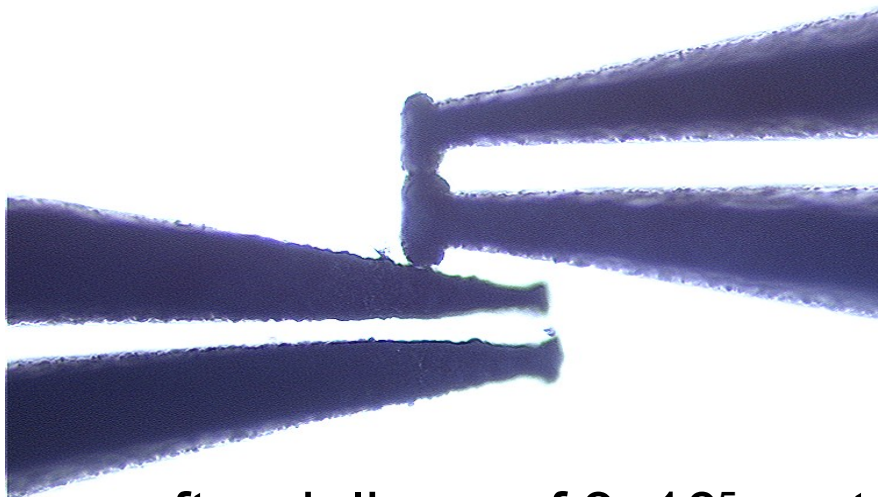


1 pin  1 block

Spotting Pin Quality Decline



after delivery of 5×10^5 spots

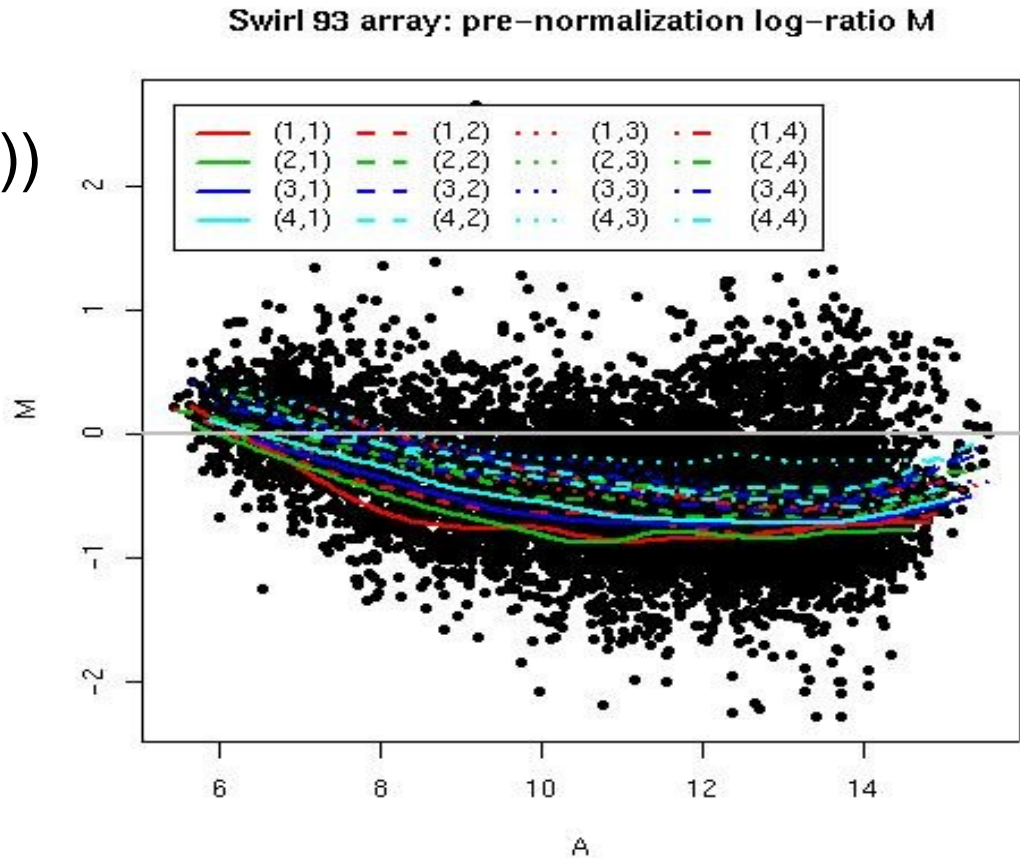


after delivery of 3×10^5 spots

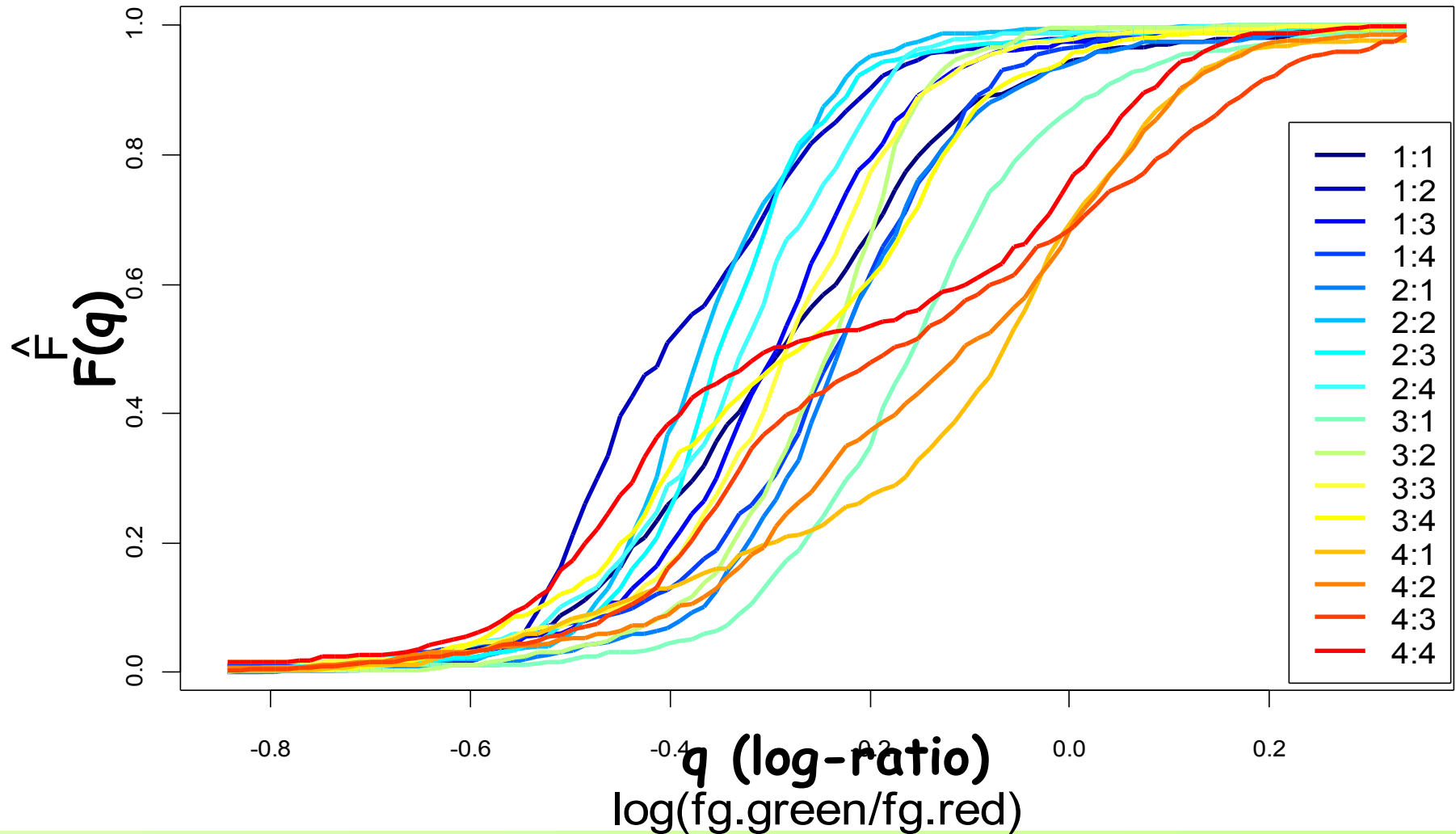
MA plot

$$M = \log_2(R) - \log_2(G)$$

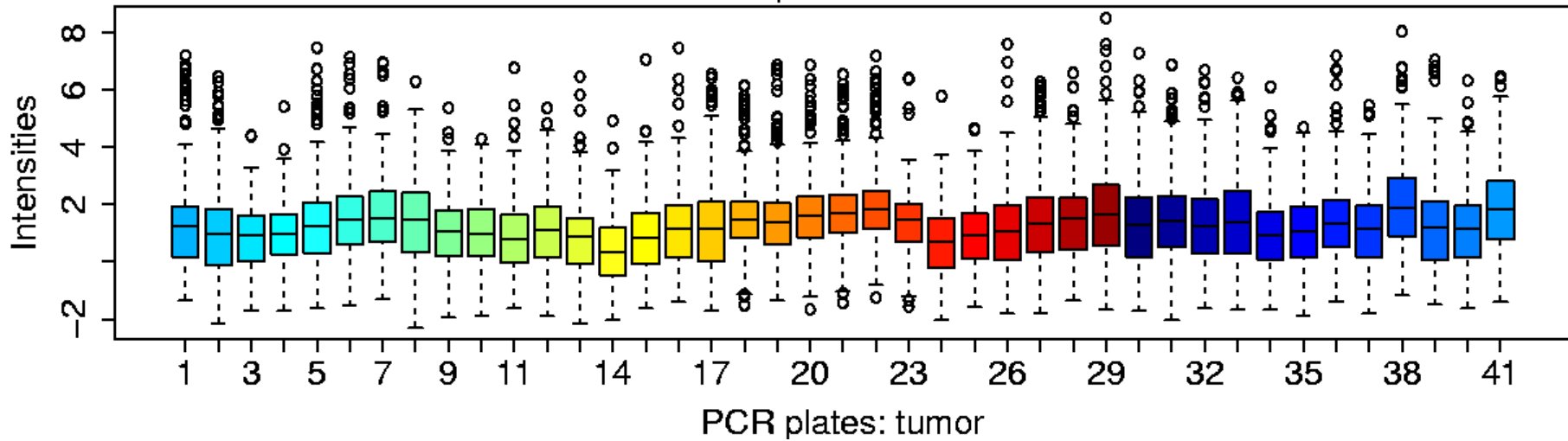
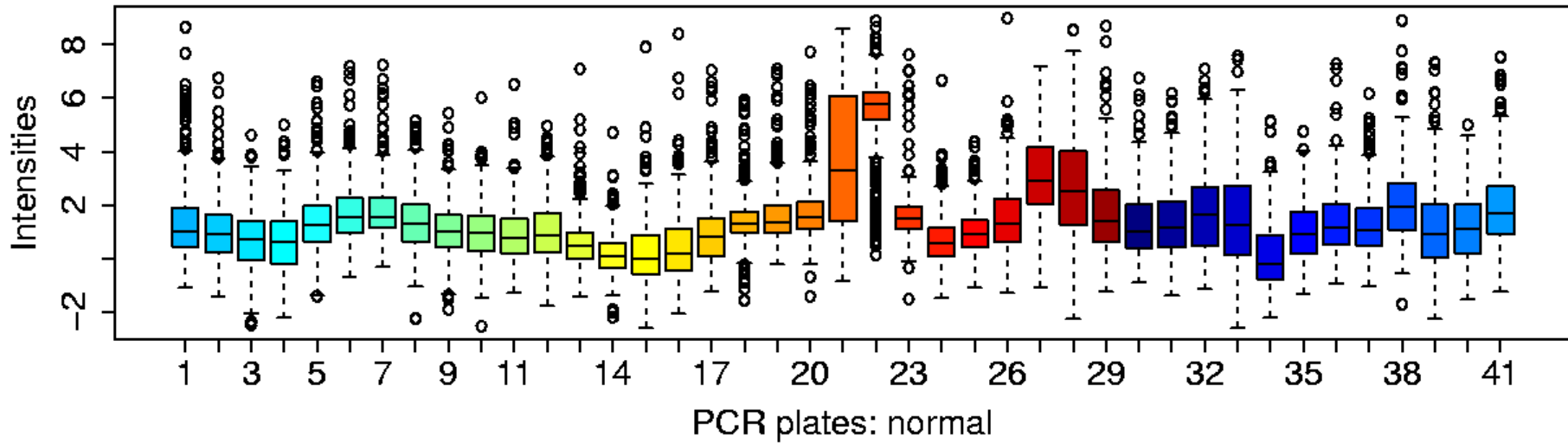
$$A = 1/2(\log_2(R) + \log_2(G))$$



Print-tip Effects – ECDF plot



PCR Plates - Boxplots



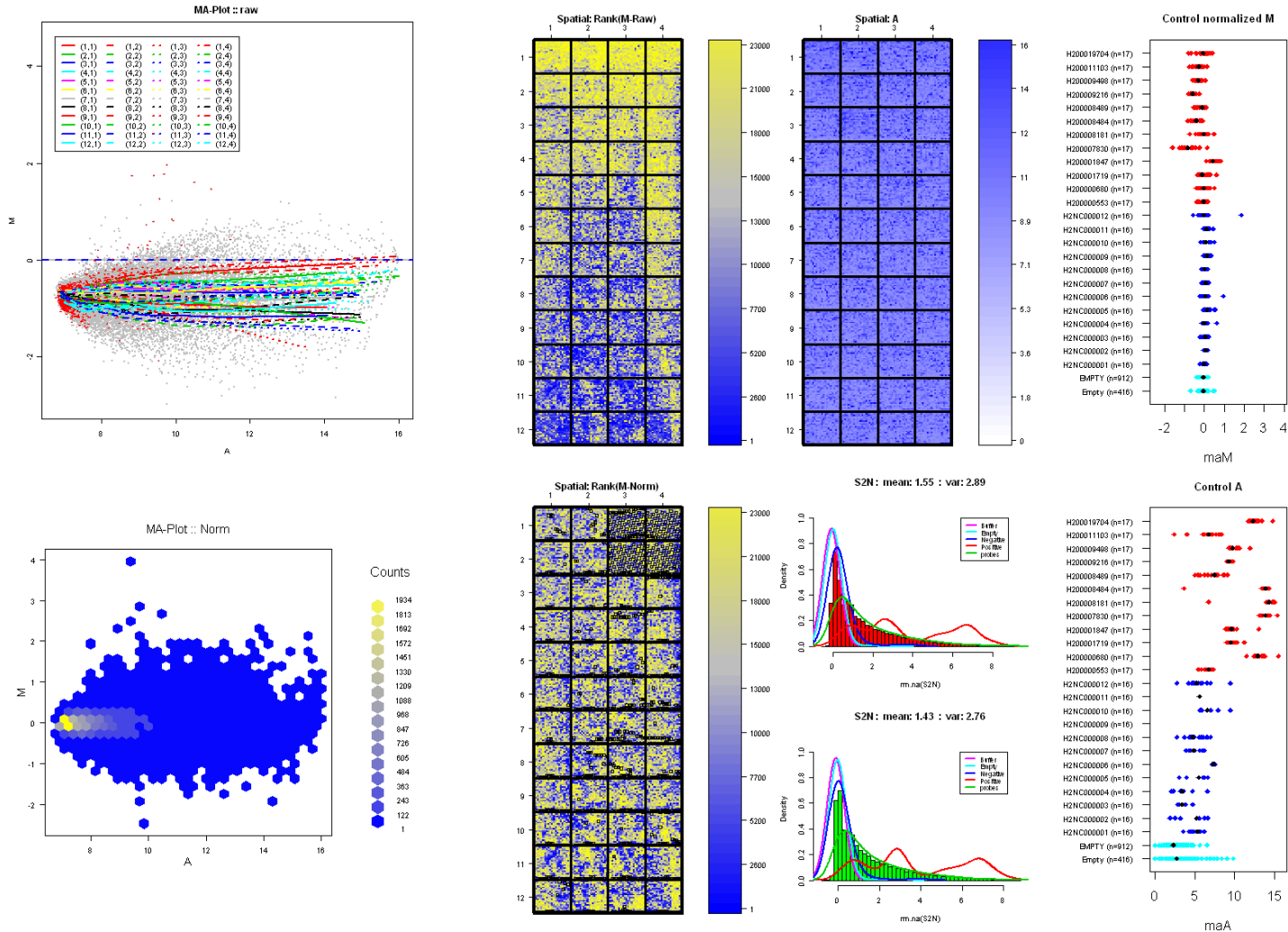
Bioconductor QA packages

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

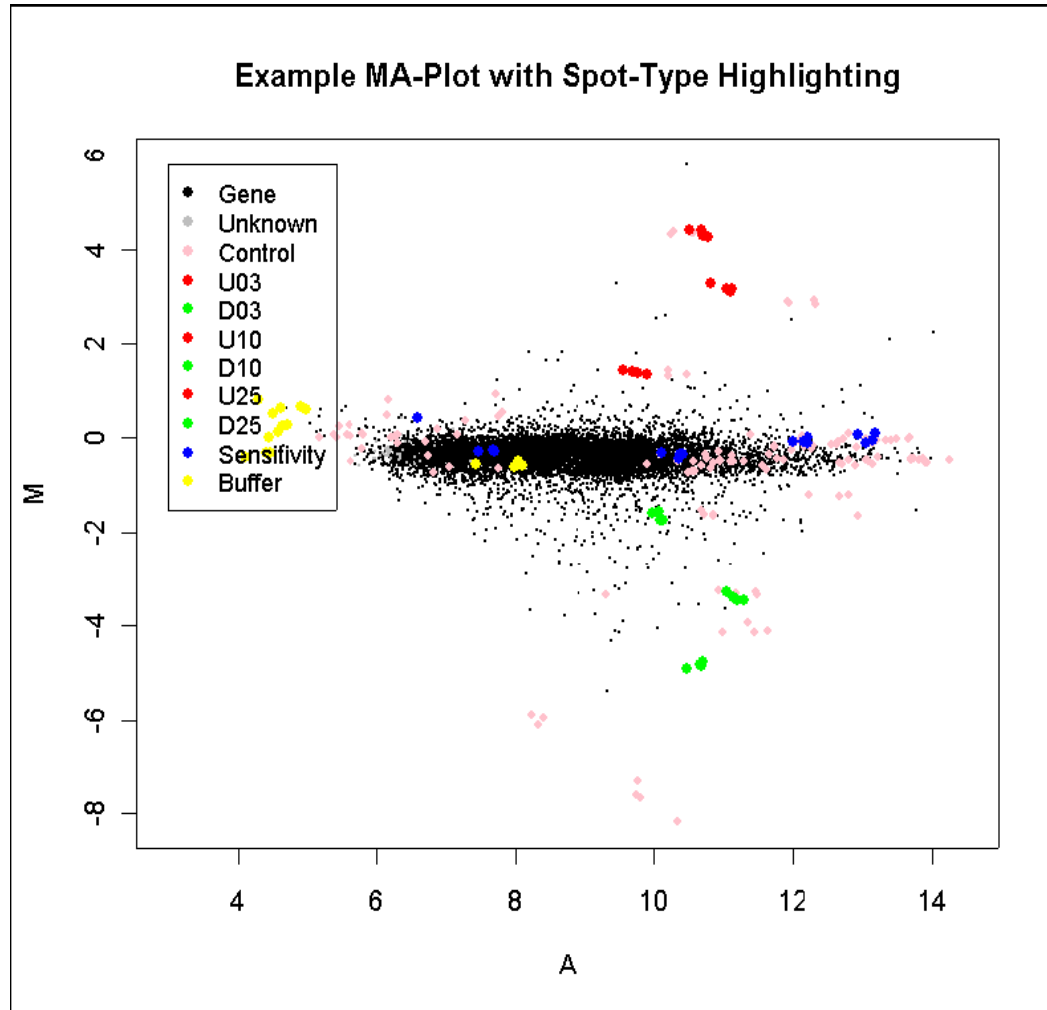
Diagnostic plot with *arrayQuality*

diagPlot6Hs.195.1.png : Qualitative Diagnostic Plots

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



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

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

■ ...

Systematic

- similar effect on many measurements
- corrections can be estimated from data

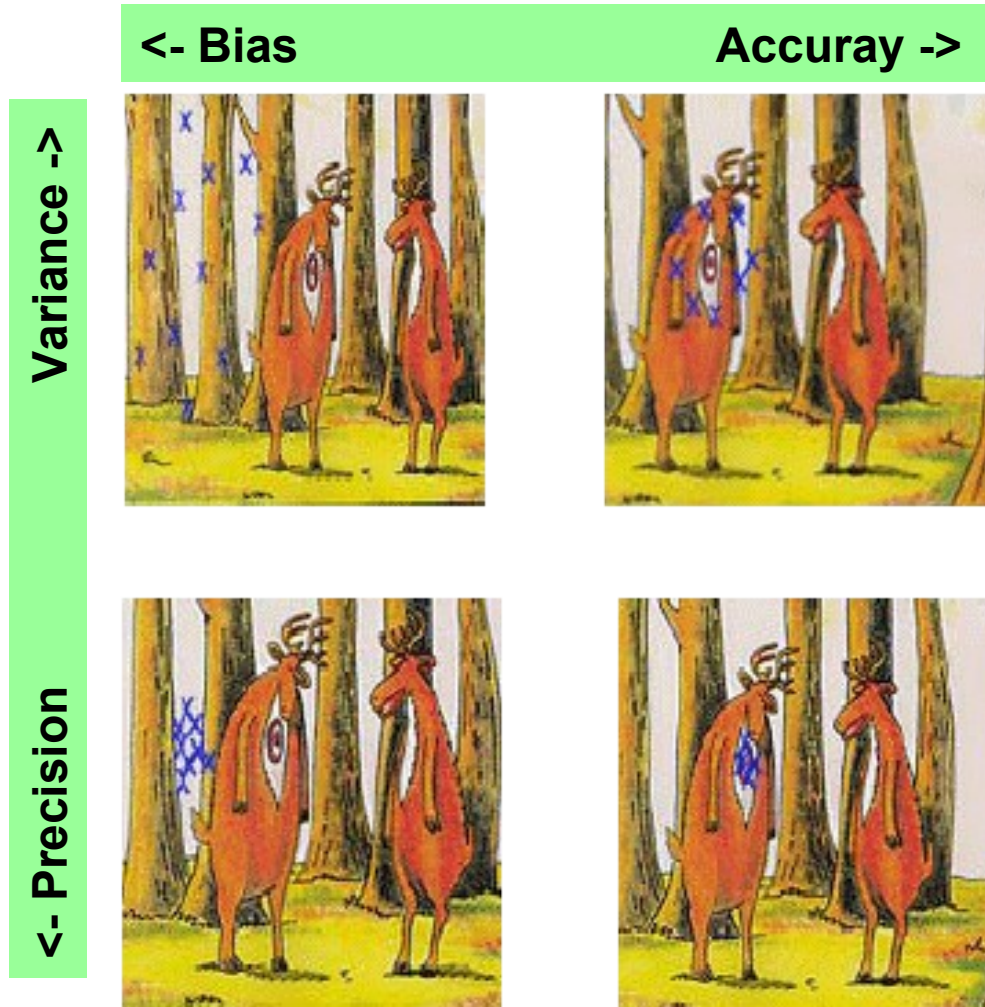
Stochastic

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

Calibration

Error Model

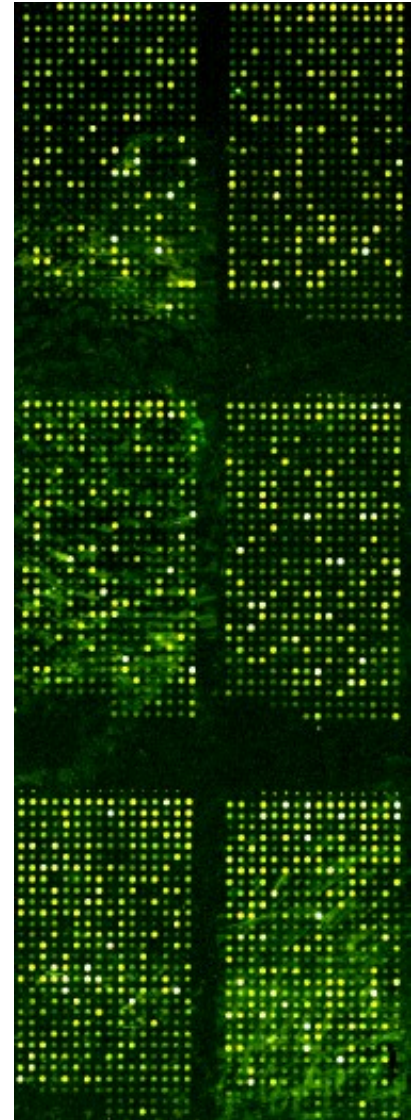
Variance-Bias trade off



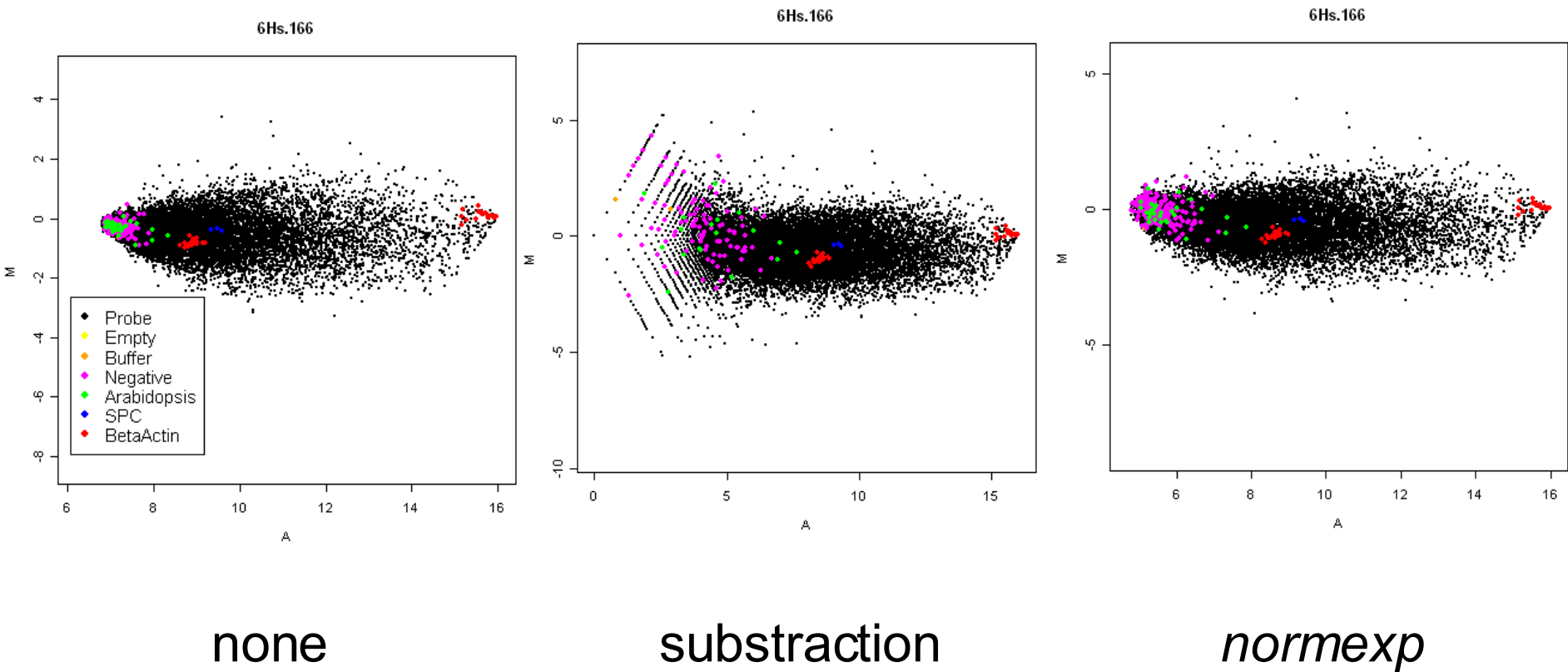
Background Correction

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

- More details ... *limma*
 >?backgroundCorrect



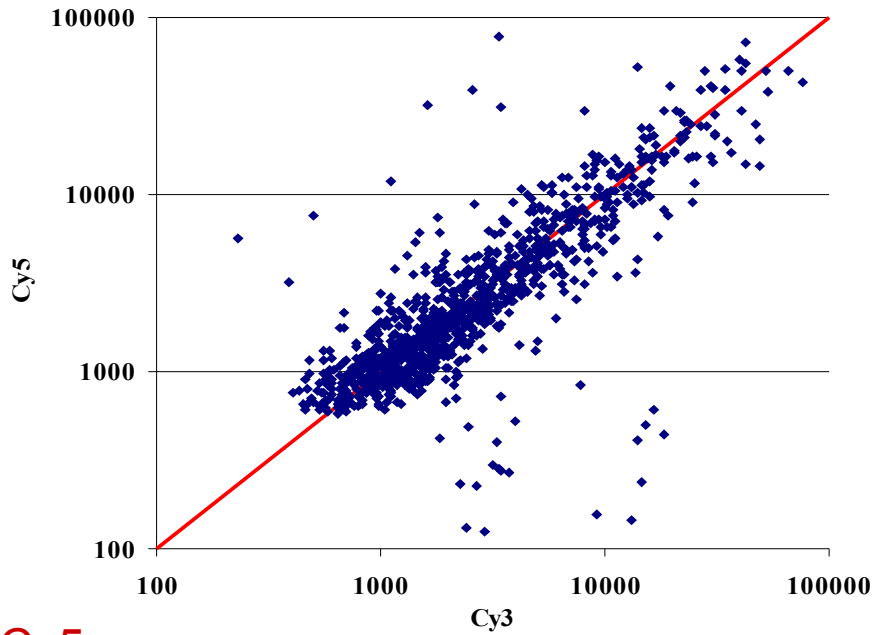
Background Correction



Why Normalize?

Theory

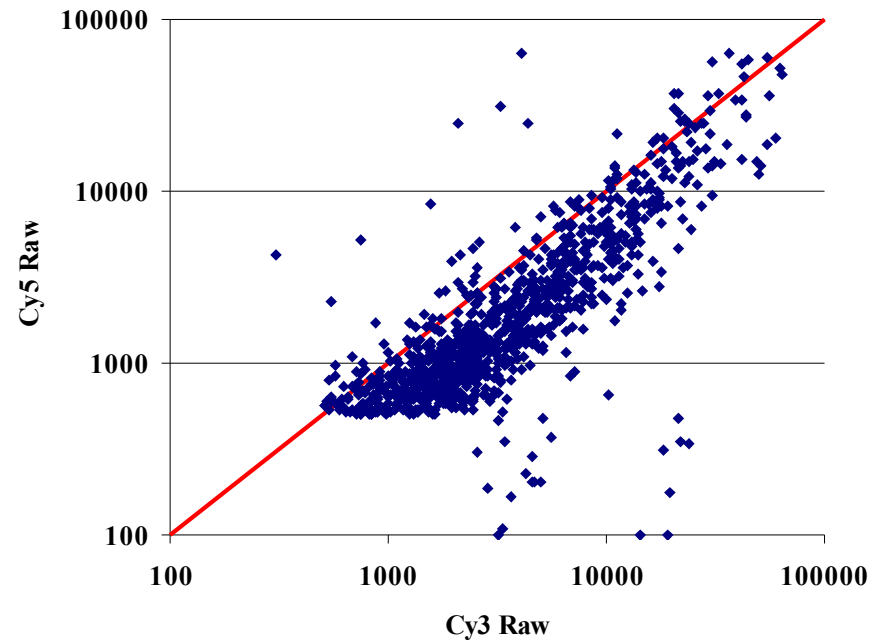
Cy5 vs Cy3



Cy5
Cy3

Reality

Raw Data - Cy5 vs Cy3



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
- Multiplicative noise becomes additive noise: variance more independent of intensity

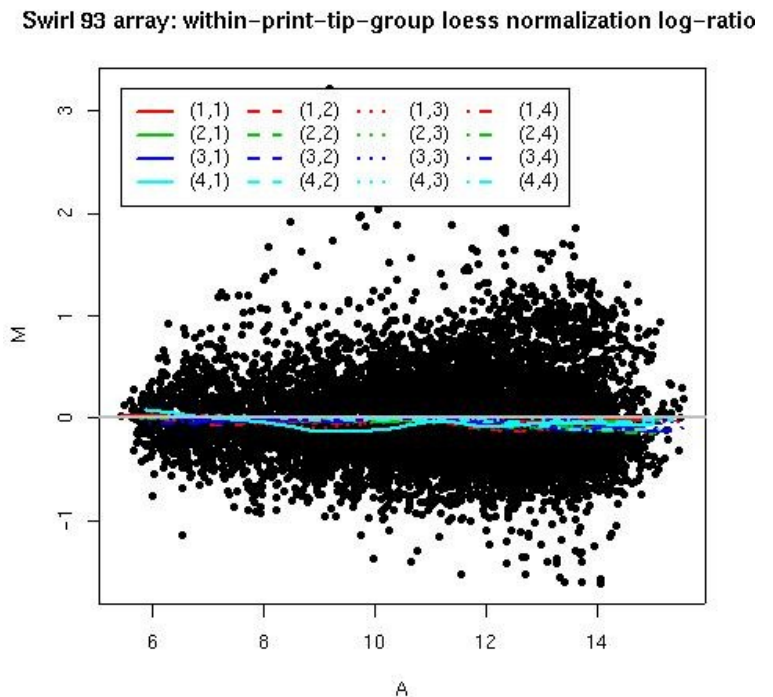
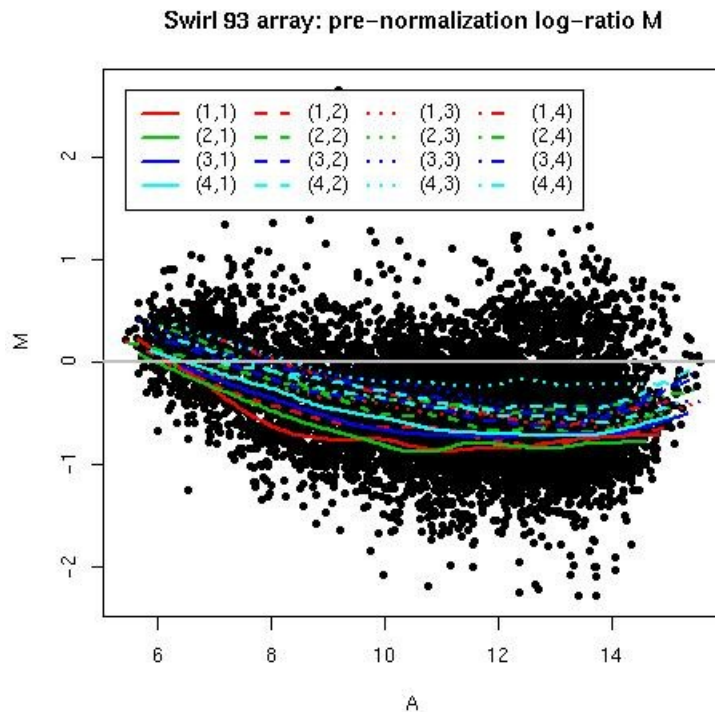
Normalization methods

- 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*

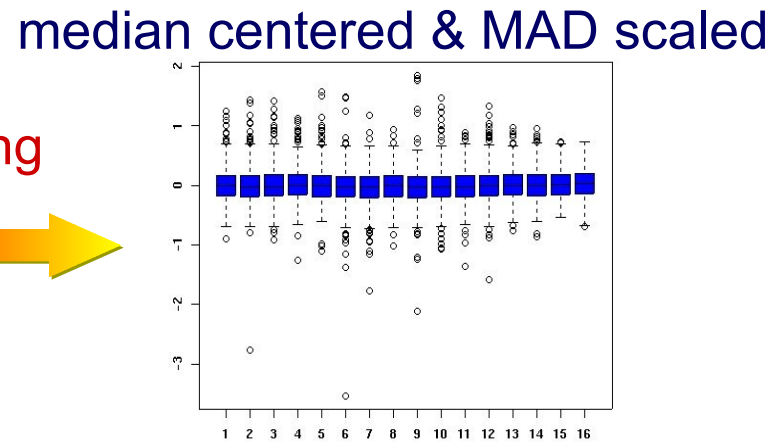
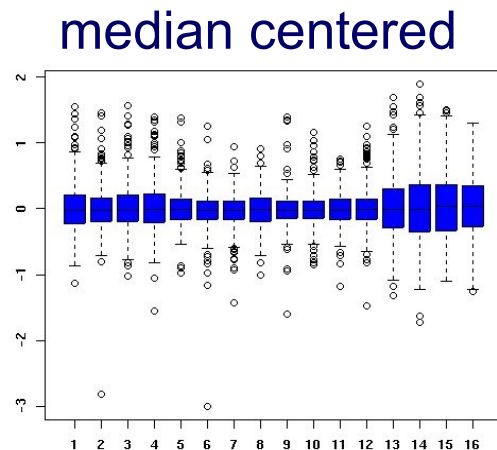
Two channel normalization

- **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



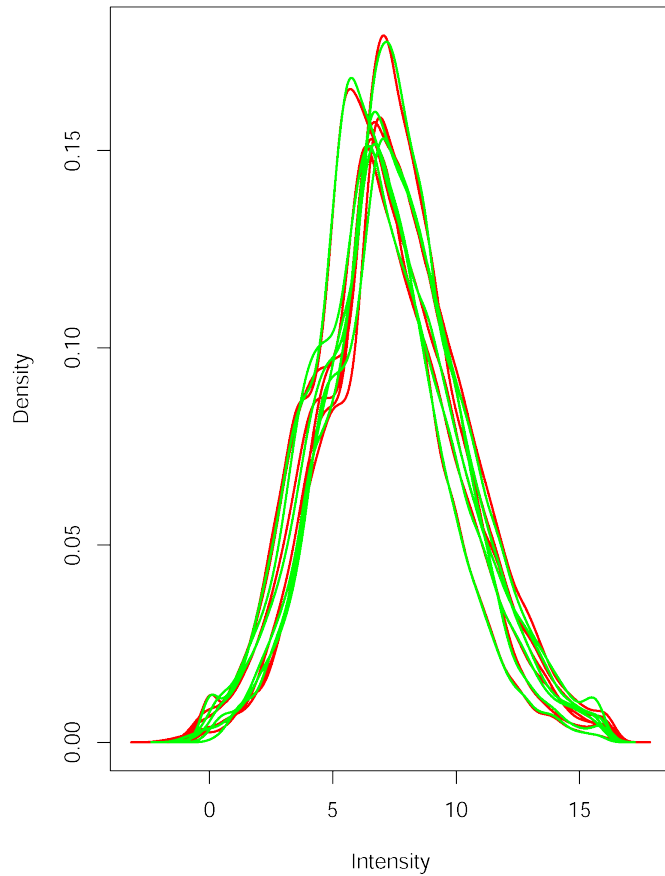
One channel normalization

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

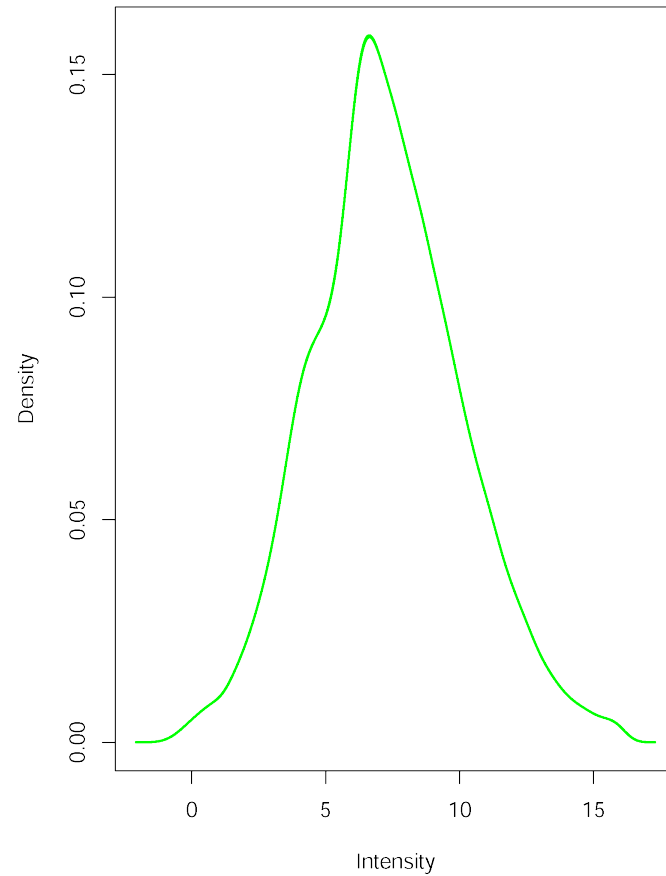
Ex: quantile normalization (*limma*)
variance stabilization (*vsn*)

Quantile Normalization

Before

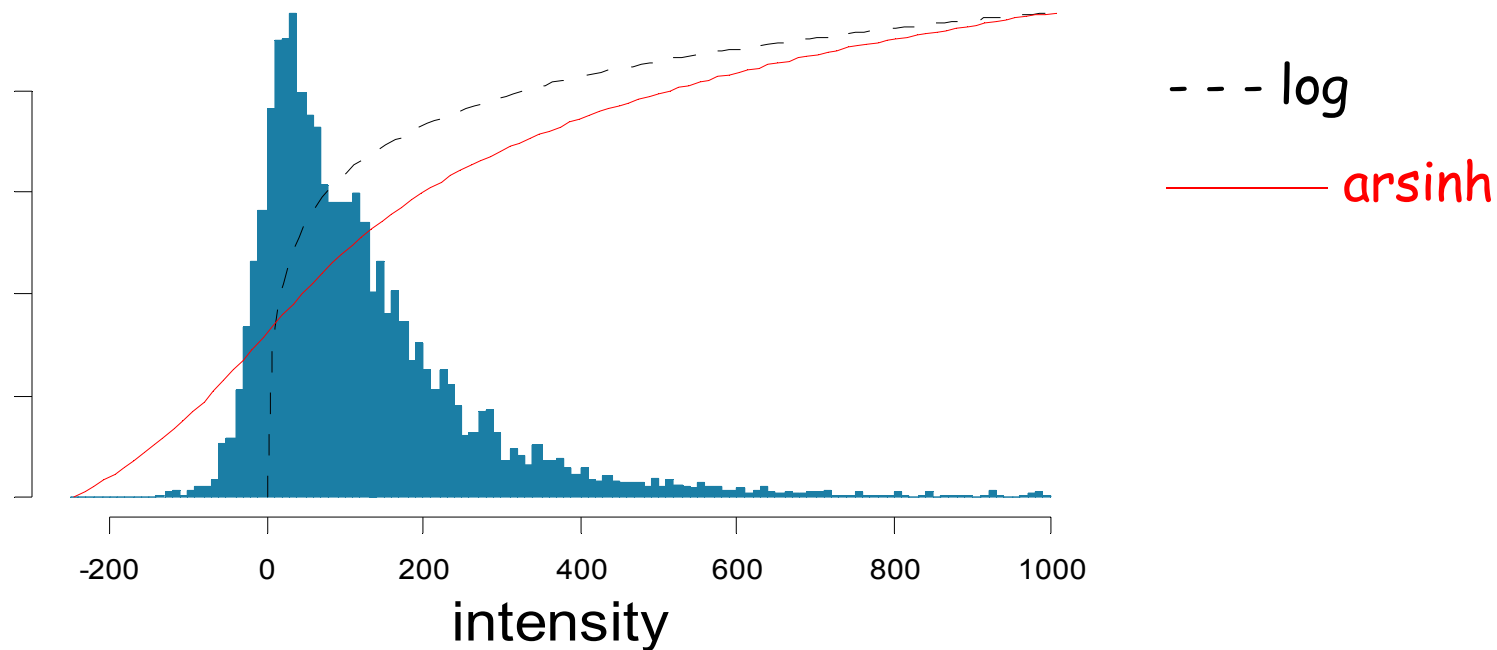


After s



Bolstand *et al.*(2003)

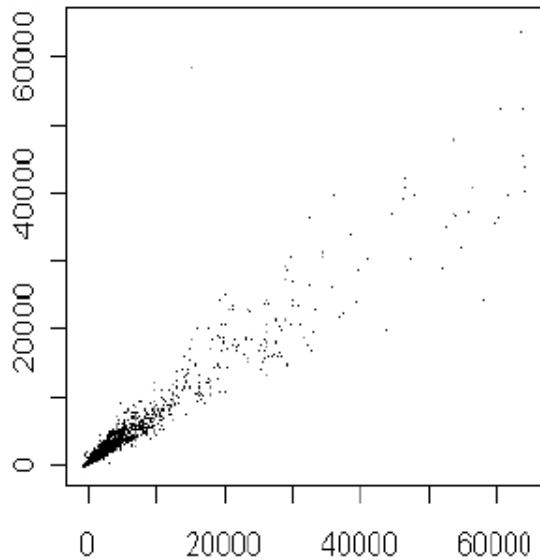
Variance Stabilizing Transformation



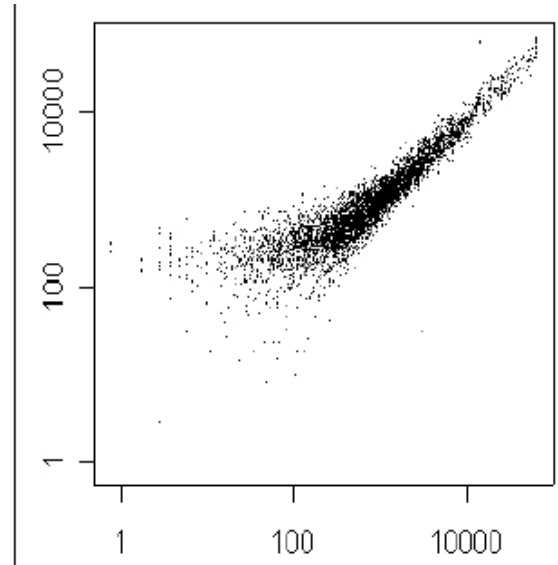
- Meaningful around 0
- Original intensities may be negatives

(Huber *et al.* 2004)

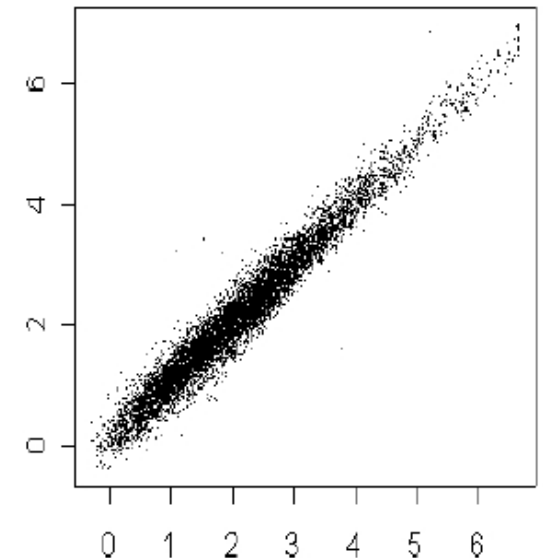
Variance stabilization (*vsn*)



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 packages:

- *marray*
- *limma*
- ...

