BIOCONDUCTOR

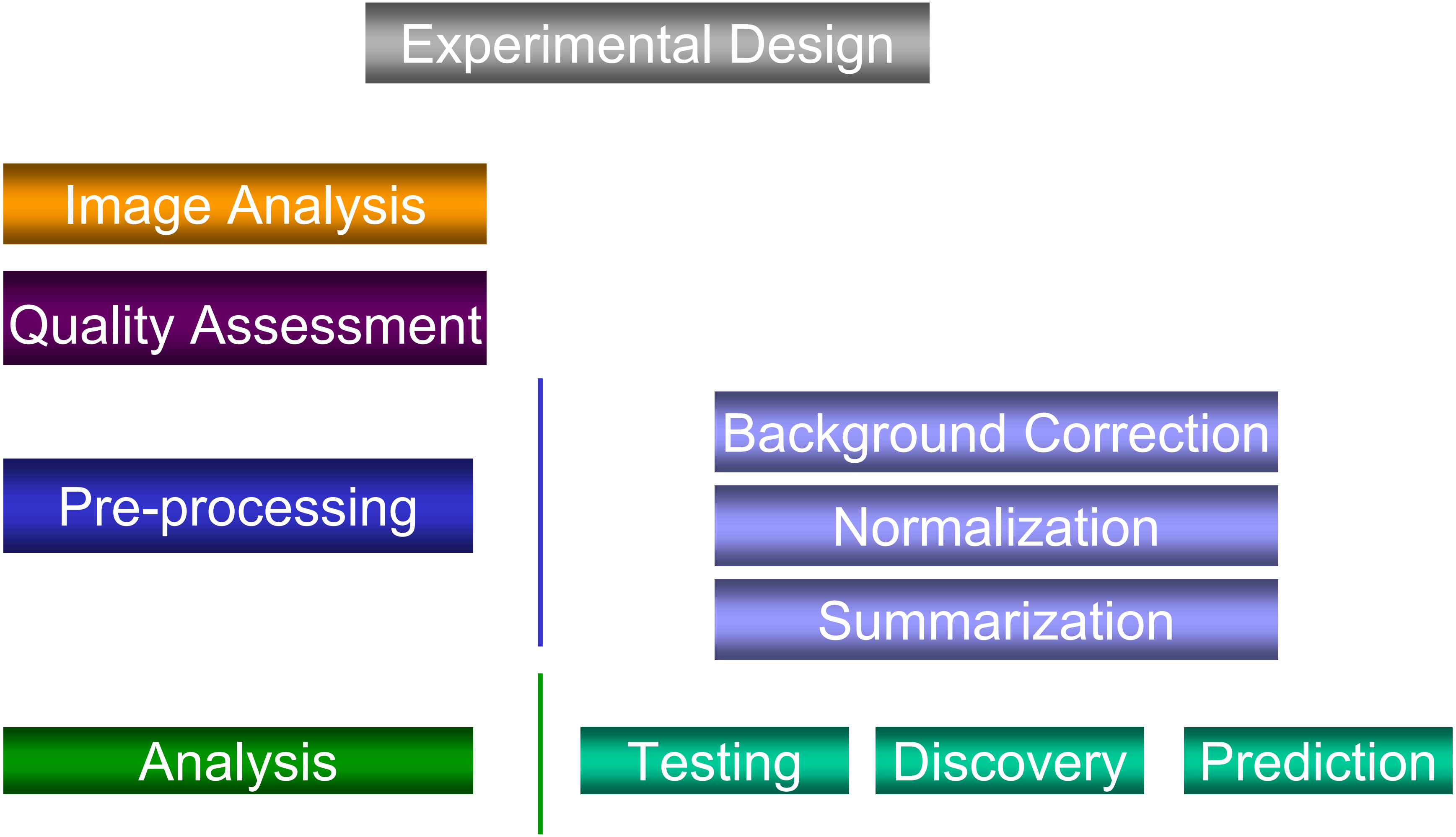
cDNA Microarray Analysis

with BioConductor packages

Nolwenn Le Meur

FRED HUTCHINSON
CANCER RESEARCH CENTER

Data Analysis of Microarrays



Experimental Design

Image Analysis

Quality Assessment

Pre-processing

Background Correction

Normalization

Summarization

Analysis

Testing

Discovery

Prediction

Outline

- **Data acquisition & Pre-processing (chap. 4)**
  - Image analysis
  - Quality assessment
  - Pre-processing
- **Differential expression (chap. 14, 15 & 23)**

- **Lab : case studies (chap 4 & 23)**
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)
  - limma (G.K Symth)

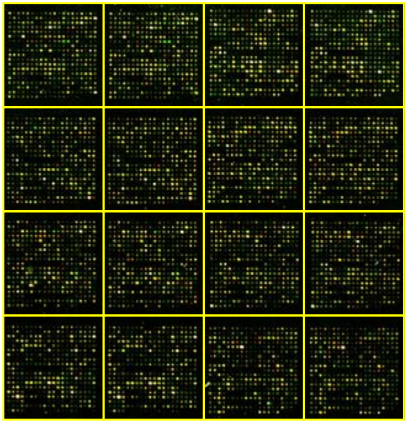
Terminology

- **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.

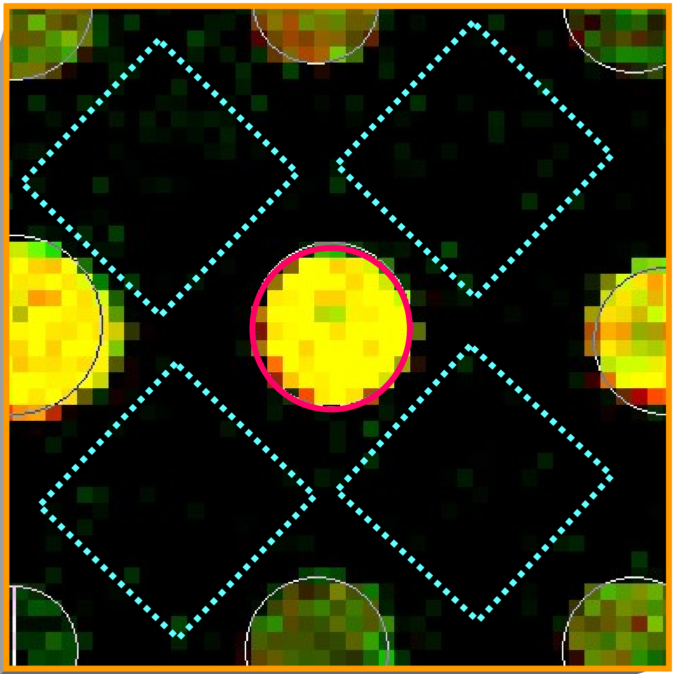
- **G, Gb:** Cy 3 signal and background intensities
- **R, Rb:** Cy5 signal and background intensities
- **M** = log2(R) - log2(G)
- **A** = 1/2(log2(R) - log2(G))

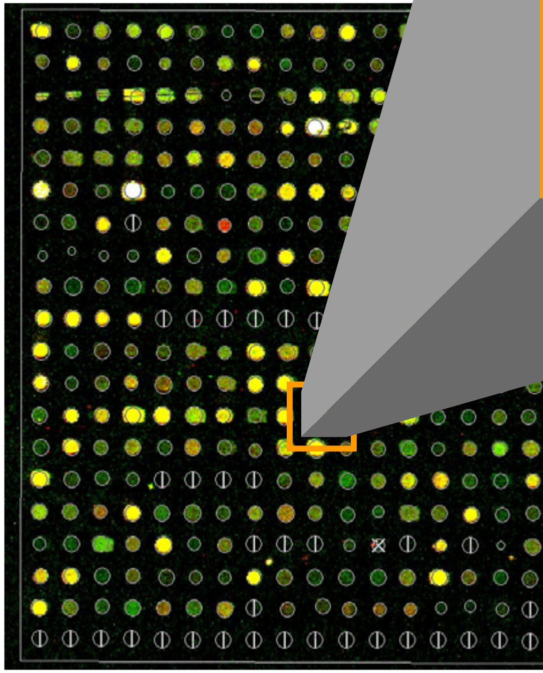
Image Analysis

1. Location

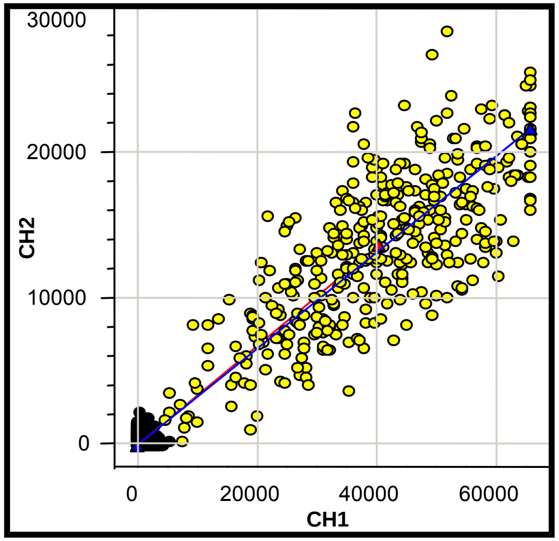


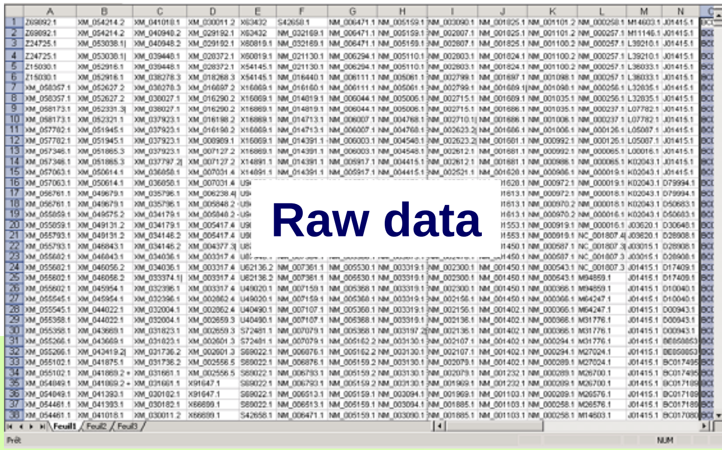
2. Segmentation





3. Quantification

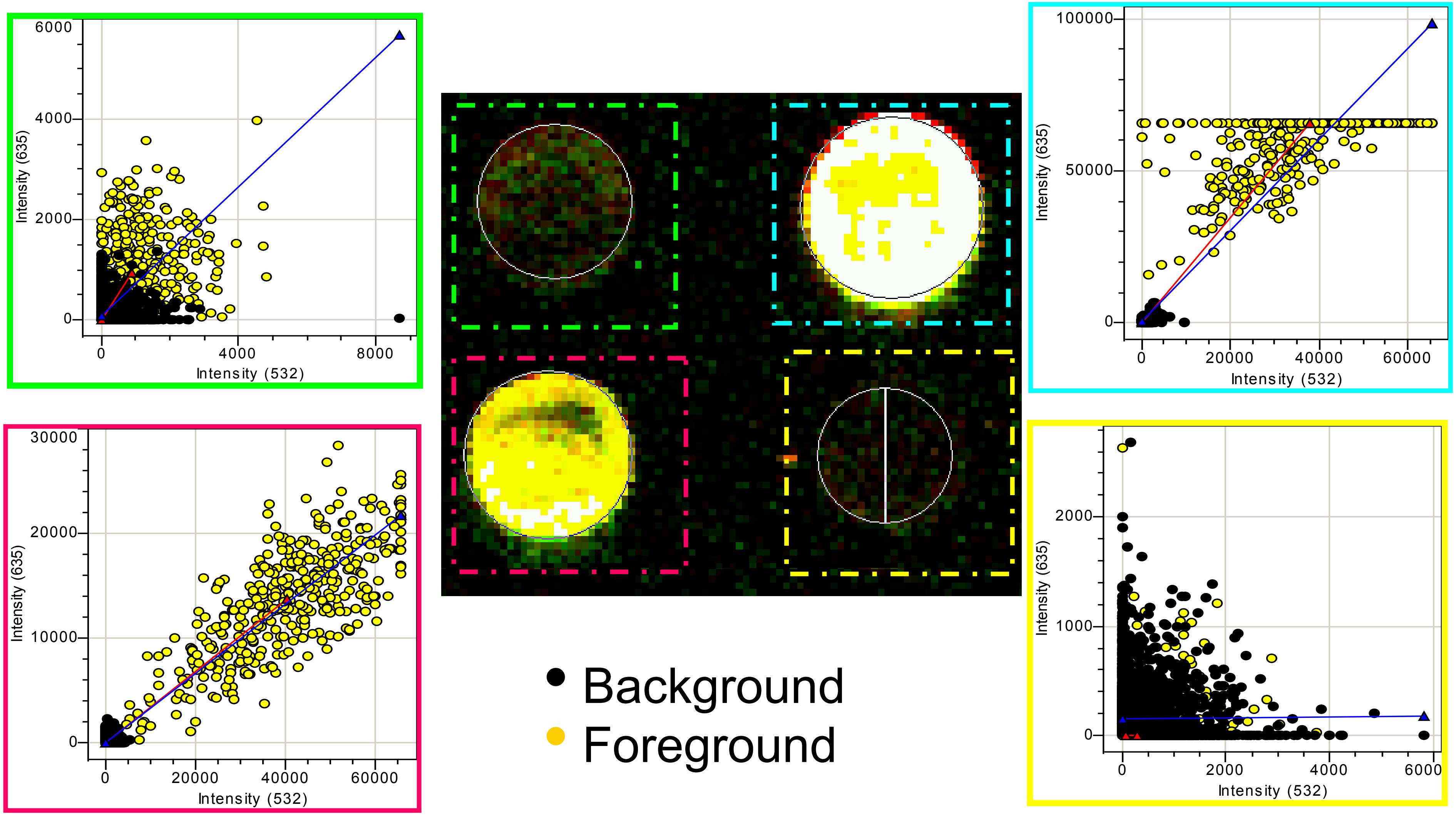




Raw data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Z69092.1	XM_054214.2	XM_041018.1	XM_030011.2	X63432	S42658.1	NM_006471.1	NM_005159.1	NM_003090.1	NM_001825.1	NM_001101.2	NM_000258.1	M14603.1	J01415.1
2	Z69092.1	XM_054214.2	XM_040948.2	XM_029192.1	X63432	NM_032169.1	NM_006471.1	NM_005159.1	NM_002807.1	NM_001825.1	NM_001101.2	NM_000257.1	M11146.1	J01415.1
3	Z24725.1	XM_053038.1|	XM_040948.2	XM_029192.1	X60819.1	NM_032169.1	NM_006471.1	NM_005159.1	NM_002807.1	NM_001825.1	NM_001100.2	NM_000257.1	L39210.1	J01415.1
4	Z24725.1	XM_053038.1|	XM_039448.1	XM_028372.1	X60819.1	NM_021130.1	NM_006294.1	NM_005110.1	NM_002803.1	NM_001824.1	NM_001100.2	NM_000257.1	L39210.1	J01415.1
5	Z15030.1	XM_052916.1	XM_039448.1	XM_028372.1	X54145.1	NM_021130.1	NM_006294.1	NM_005110.1	NM_002803.1	NM_001824.1	NM_001100.2	NM_000257.1	L36033.1	J01415.1
6	Z15030.1	XM_052916.1	XM_038278.3	XM_018268.3	X54145.1	NM_016440.1	NM_006111.1	NM_005061.1	NM_002799.1	NM_001697.1	NM_001098.1	NM_000257.1	L36033.1	J01415.1
7	XM_058357.1	XM_052627.2	XM_038278.3	XM_016697.2	X16869.1	NM_016160.1	NM_006111.1	NM_005061.1	NM_002799.1	NM_001689.1|	NM_001098.1	NM_000256.1	L32035.1	J01415.1
8	XM_058357.1	XM_052627.2	XM_038027.1	XM_016290.2	X16869.1	NM_014819.1	NM_006044.1	NM_005006.1	NM_002715.1	NM_001689.1	NM_001035.1	NM_000256.1	L32035.1	J01415.1
9	XM_058173.1	XM_052331.3|	XM_038027.1	XM_016290.2	X16869.1	NM_014819.1	NM_006044.1	NM_005006.1	NM_002715.1	NM_001686.1	NM_001035.1	NM_000237.1	L07782.1	J01415.1
10	XM_058173.1	XM_052321.1	XM_037923.1	XM_016198.2	X16869.1	NM_014713.1	NM_006007.1	NM_004768.1	NM_002710.1|	NM_001686.1	NM_001006.1	NM_000237.1	L07782.1	J01415.1
11	XM_057782.1	XM_051945.1	XM_037923.1	XM_016198.2	X16869.1	NM_014713.1	NM_006007.1	NM_004768.1	NM_002623.2|	NM_001686.1	NM_001006.1	NM_000126.1	L05087.1	J01415.1
12	XM_057782.1	XM_051945.1	XM_037923.1	XM_008989.1	X16869.1	NM_014391.1	NM_006003.1	NM_004548.1	NM_002623.2|	NM_001681.1	NM_000992.1	NM_000126.1	L05087.1	J01415.1
13	XM_057346.1	XM_051865.3	XM_037923.1	XM_007127.2	X16869.1	NM_014391.1	NM_006003.1	NM_004548.1	NM_002612.1	NM_001681.1	NM_000992.1	NM_000065.1	L00016.1	J01415.1
14	XM_057346.1	XM_051865.3	XM_037797.2|	XM_007127.2	X14891.1	NM_014391.1	NM_005917.1	NM_004415.1	NM_002612.1	NM_001681.1	NM_000986.1	NM_000065.1	K02043.1	J01415.1
15	XM_057063.1	XM_050614.1	XM_036858.1	XM_007031.4	X14891.1	NM_014391.1	NM_005917.1	NM_004415.1	NM_002521.1	NM_001620.1	NM_000986.1	NM_000019.1	K02043.1	J01415.1
16	XM_057063.1	XM_050614.1	XM_036858.1	XM_007031.4	U9					01620.1	NM_000972.1	NM_000019.1	K02043.1	D79994.1
17	XM_056761.1	XM_049679.1	XM_035796.1	XM_006238.4|	U9					01613.1	NM_000972.1	NM_000018.1	K02043.1	D79994.1
18	XM_056761.1	XM_049679.1	XM_035796.1	XM_005848.2	U9					01613.1	NM_000970.2	NM_000018.1	K02043.1	D50683.1
19	XM_055859.1	XM_049675.2	XM_034179.1	XM_005848.2	U9					01613.1	NM_000970.2	NM_000016.1	K02043.1	D50683.1
20	XM_055859.1	XM_049131.2	XM_034179.1	XM_005417.4	U9					01553.1	NM_000919.1	NM_000016.1	J03620.1	D30648.1
21	XM_055793.1	XM_049131.2	XM_034146.2	XM_005417.4	U9					01553.1	NM_000919.1	NC_001807.4|	J03620.1	D28908.1
22	XM_055793.1	XM_046843.1	XM_034146.2	XM_004377.3|	U8					01450.1	NM_000587.1	NC_001807.3|	J03015.1	D28908.1
23	XM_055602.1	XM_046843.1	XM_034036.1	XM_003317.4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	01450.1	NM_000587.1	NC_001807.3	J03015.1	D28908.1
24	XM_055602.1	XM_046056.2	XM_034036.1	XM_003317.4	U62136.2	NM_007361.1	NM_005530.1	NM_003319.1	NM_002300.1	NM_001450.1	NM_000543.1	NC_001807.3	J01415.1	D17409.1
25	XM_055602.1	XM_046056.2	XM_033374.1|	XM_003317.4	U62136.2	NM_007361.1	NM_005530.1	NM_003319.1	NM_002300.1	NM_001450.1	NM_000543.1	M94859.1	J01415.1	D17409.1
26	XM_055602.1	XM_045954.1	XM_032396.1	XM_003317.4	U49020.1	NM_007159.1	NM_005368.1	NM_003319.1	NM_002300.1	NM_001450.1	NM_000366.1	M94859.1	J01415.1	D10040.1
27	XM_055545.1	XM_045954.1	XM_032396.1	XM_002862.4	U49020.1	NM_007159.1	NM_005368.1	NM_003319.1	NM_002156.1	NM_001450.1	NM_000366.1	M64247.1	J01415.1	D10040.1
28	XM_055545.1	XM_044022.1	XM_032004.1	XM_002862.4	U40490.1	NM_007107.1	NM_005368.1	NM_003319.1	NM_002156.1	NM_001402.1	NM_000366.1	M64247.1	J01415.1	D00943.1
29	XM_055358.1	XM_044022.1	XM_032004.1	XM_002659.3	U40490.1	NM_007107.1	NM_005368.1	NM_003319.1	NM_002136.1	NM_001402.1	NM_000366.1	M31776.1	J01415.1	D00943.1
30	XM_055358.1	XM_043669.1	XM_031823.1	XM_002659.3	S72481.1	NM_007079.1	NM_005368.1	NM_003197.2|	NM_002136.1	NM_001402.1	NM_000366.1	M31776.1	J01415.1	D00943.1
31	XM_055266.1	XM_043669.1	XM_031823.1	XM_002601.3	S72481.1	NM_007079.1	NM_005162.2	NM_003130.1	NM_002107.1	NM_001402.1	NM_000294.1	M31776.1	J01415.1	BE050853
32	XM_055266.1	XM_043419.2|	XM_031736.2	XM_002601.3	S69022.1	NM_006876.1	NM_005162.2	NM_003130.1	NM_002107.1	NM_001402.1	NM_000294.1	M27024.1	J01415.1	BE050853
33	XM_055102.1	XM_041875.1	XM_031736.2	XM_002556.5	S69022.1	NM_006876.1	NM_005159.2	NM_003130.1	NM_002079.1	NM_001402.1	NM_000289.1	M27024.1	J01415.1	BC017495
34	XM_055102.1	XM_041869.2 *	XM_031681.1	XM_002556.5	S69022.1	NM_006793.1	NM_005159.2	NM_003130.1	NM_002079.1	NM_001232.1	NM_000289.1	M26700.1	J01415.1	BC017495
35	XM_054849.1	XM_041869.2 *	XM_031681.1	X91647.1	S69022.1	NM_006793.1	NM_005159.2	NM_003130.1	NM_001969.1	NM_001232.1	NM_000289.1	M26700.1	J01415.1	BC017189
36	XM_054849.1	XM_041393.1	XM_030182.1	X91647.1	S69022.1	NM_006513.1	NM_005159.1	NM_003094.1	NM_001969.1	NM_001103.1	NM_000289.1	M26576.1	J01415.1	BC017189
37	XM_054461.1	XM_041393.1	XM_030182.1	X66899.1	S69022.1	NM_006513.1	NM_005159.1	NM_003094.1	NM_001885.1	NM_001103.1	NM_000258.1	M26576.1	J01415.1	BC017189
38	XM_054461.1	XM_041018.1	XM_030011.2	X66899.1	S42658.1	NM_006471.1	NM_005159.1	NM_003090.1	NM_001885.1	NM_001103.1	NM_000258.1	M14603.1	J01415.1	BC017080

Quality Filtering



● Background

● Foreground

Quality Assessment

For each array:

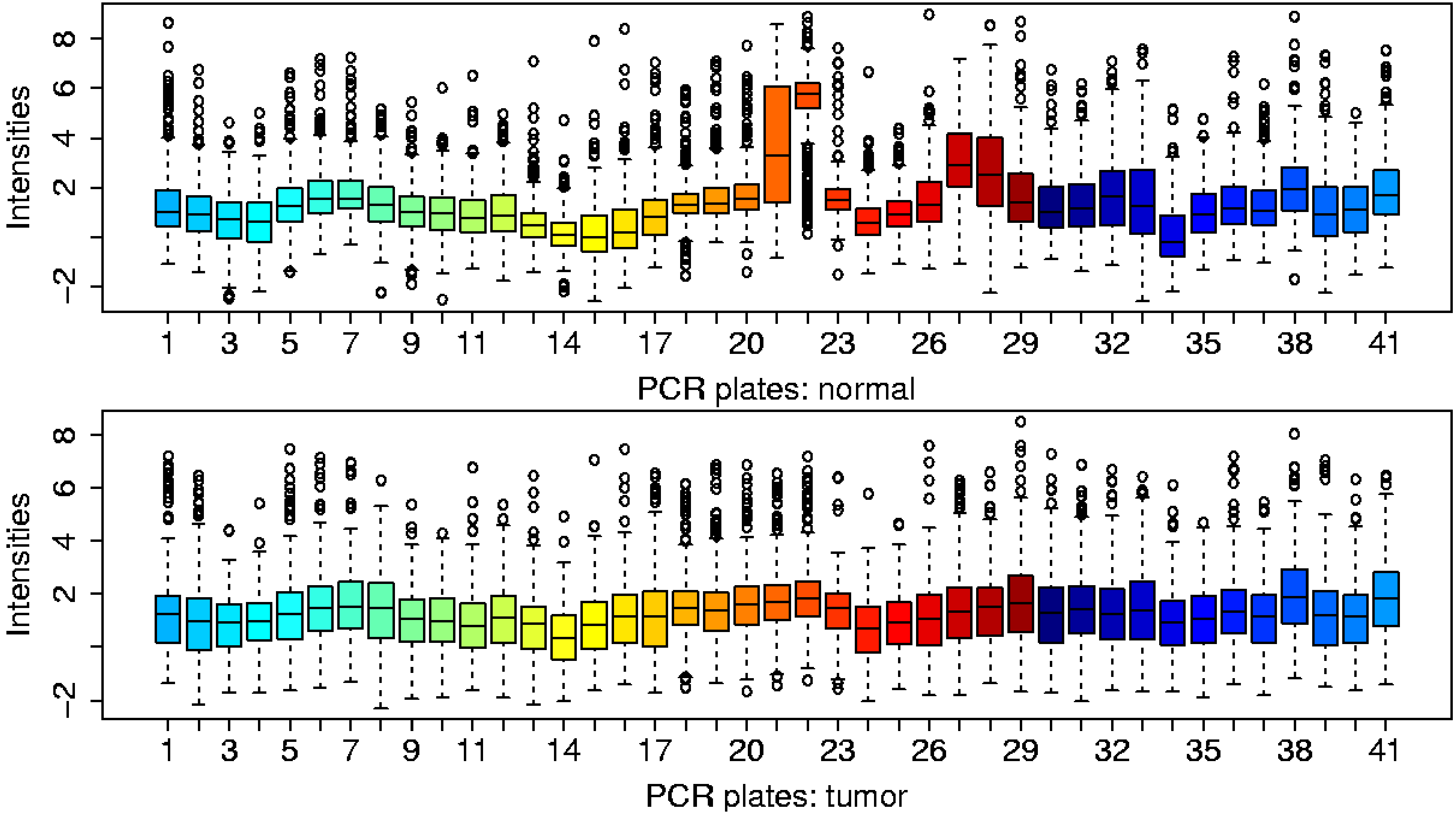
- **Diagnostics plots** of spot statistics

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

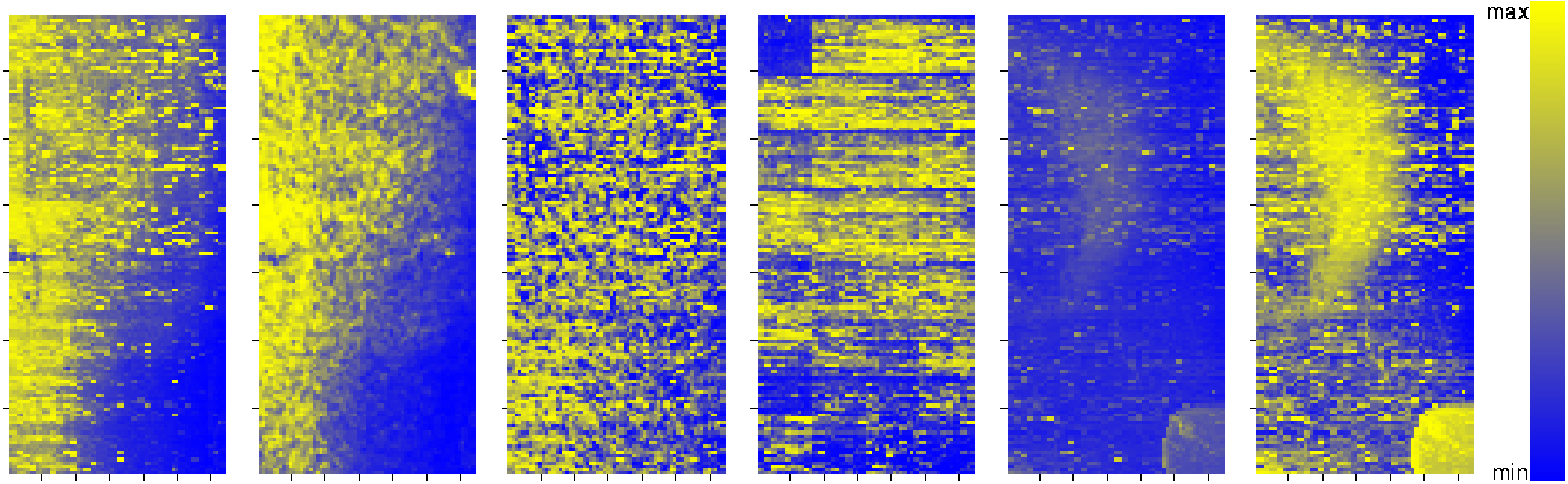
  - Boxplots;
  - 2D spatial images;
  - Scatter-plots, e.g. MA-plots;
  - Density plots.

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

PCR Plates - Boxplots



Spatial Effects – Image Plots



R Rb R-Rb

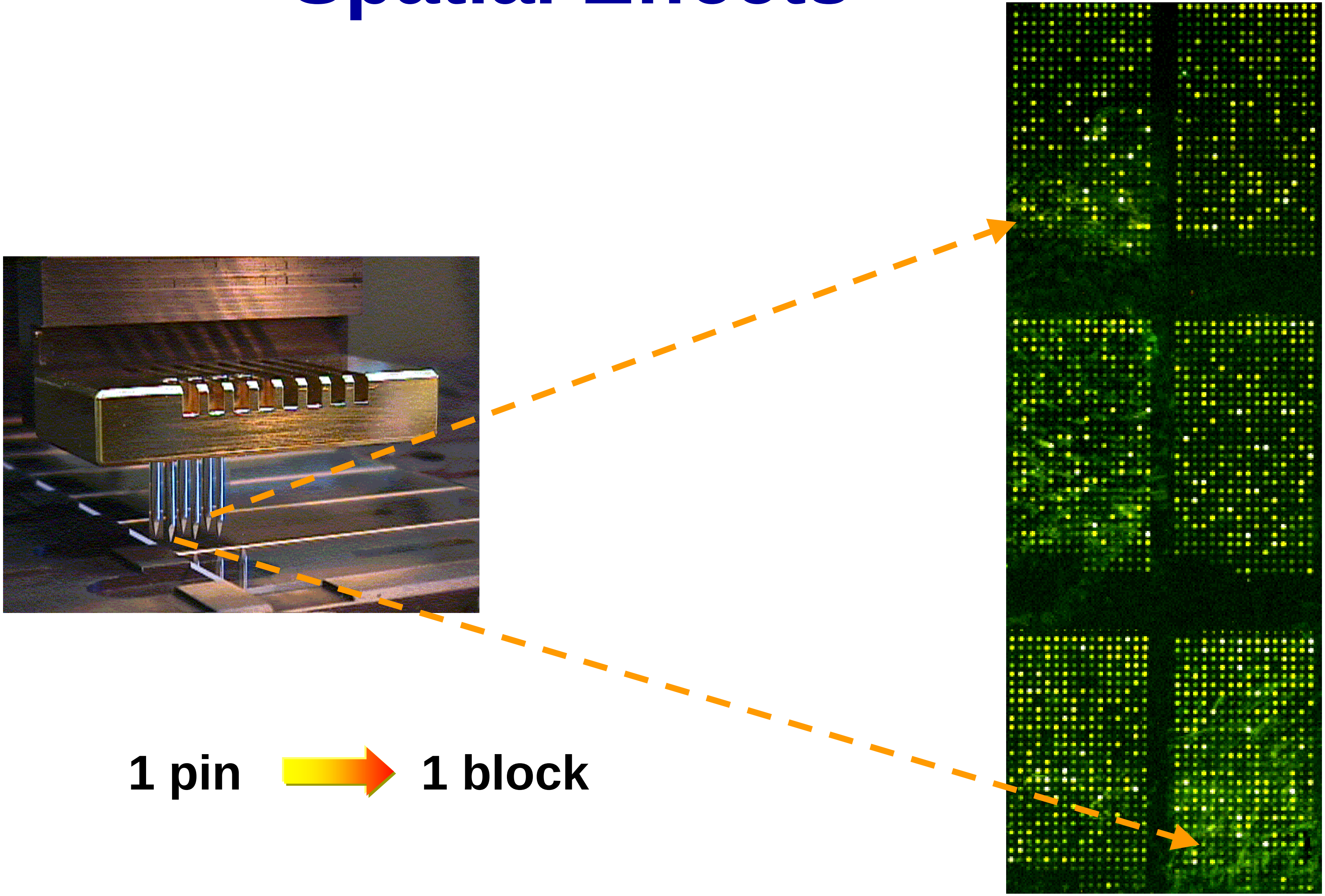
color scale by rank

another array: print-tip

color scale ~ log(G)

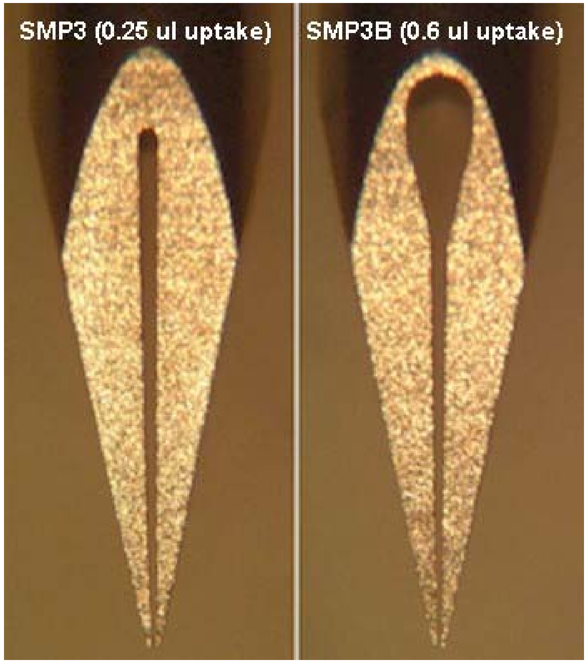
color scale ~ rank(G)

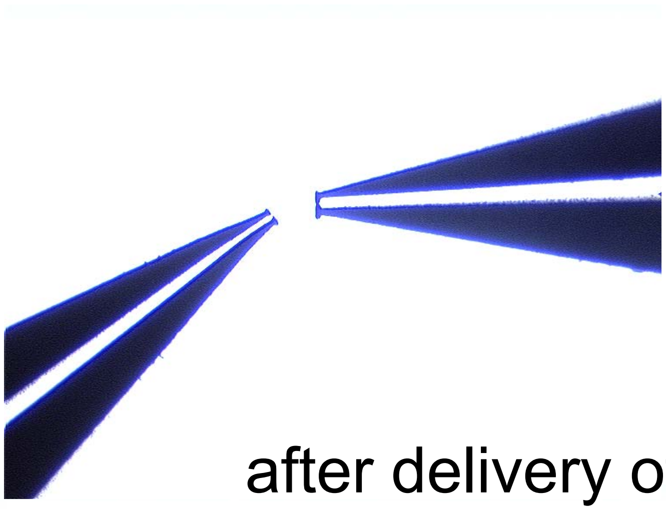
Spatial Effects



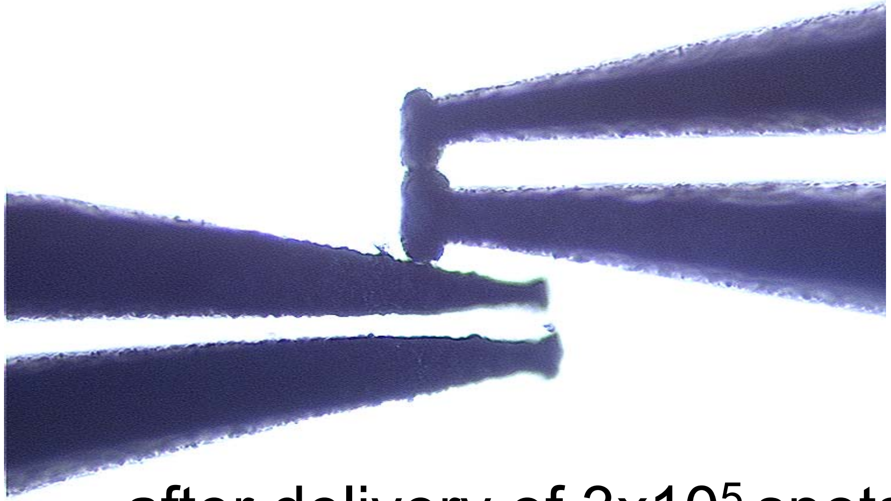
1 pin → 1 block

Spotting Pin Quality Decline





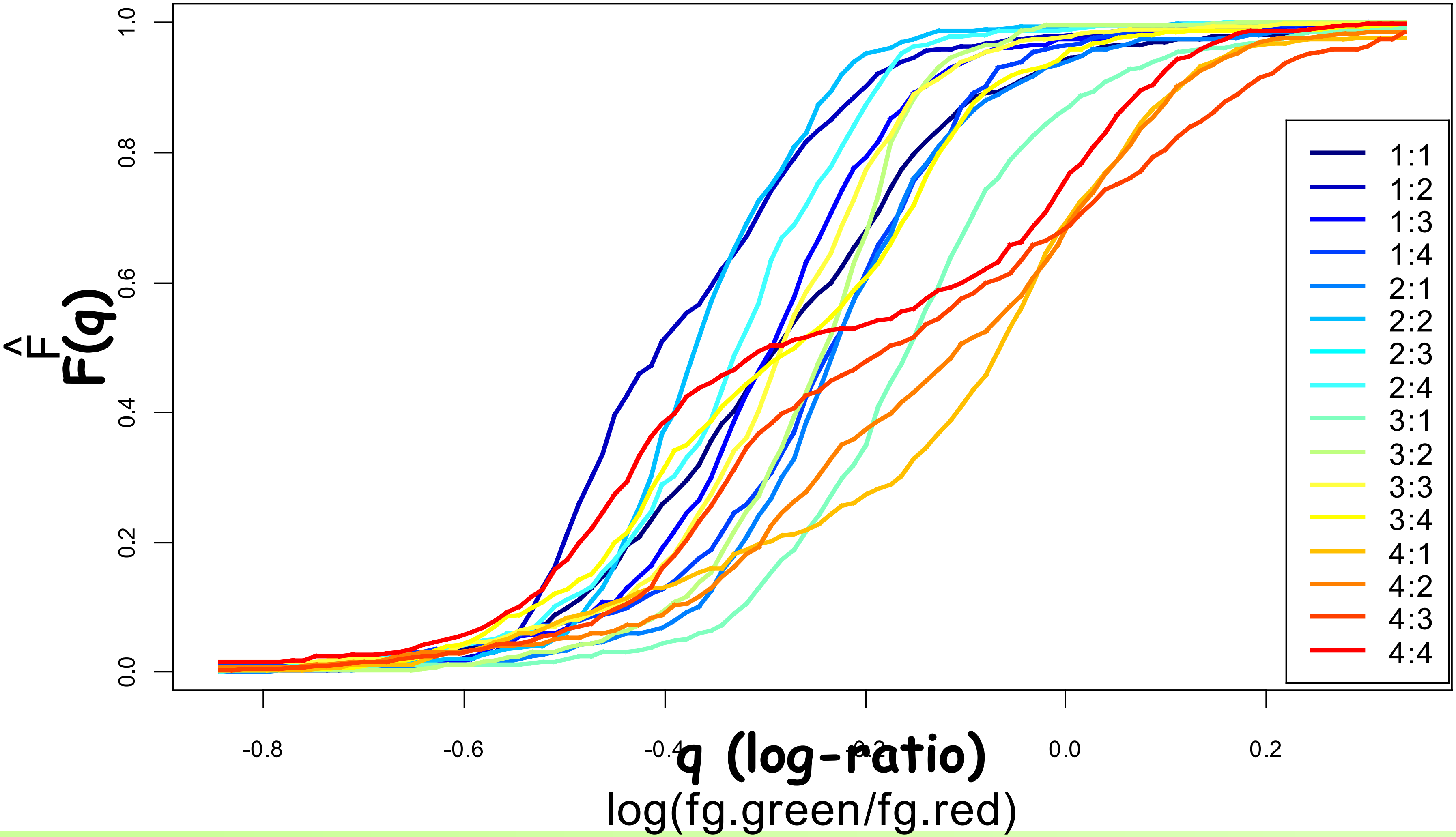
after delivery of 5x10^5 spots



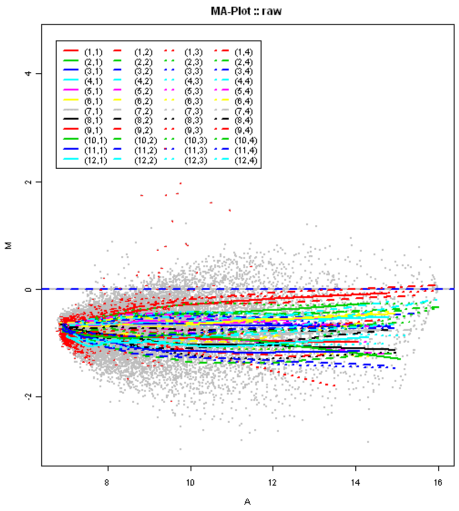
after delivery of 3x10^5 spots

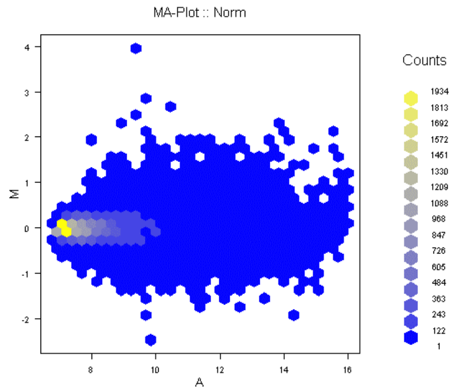
H. Sueltmann DKFZ/MGA

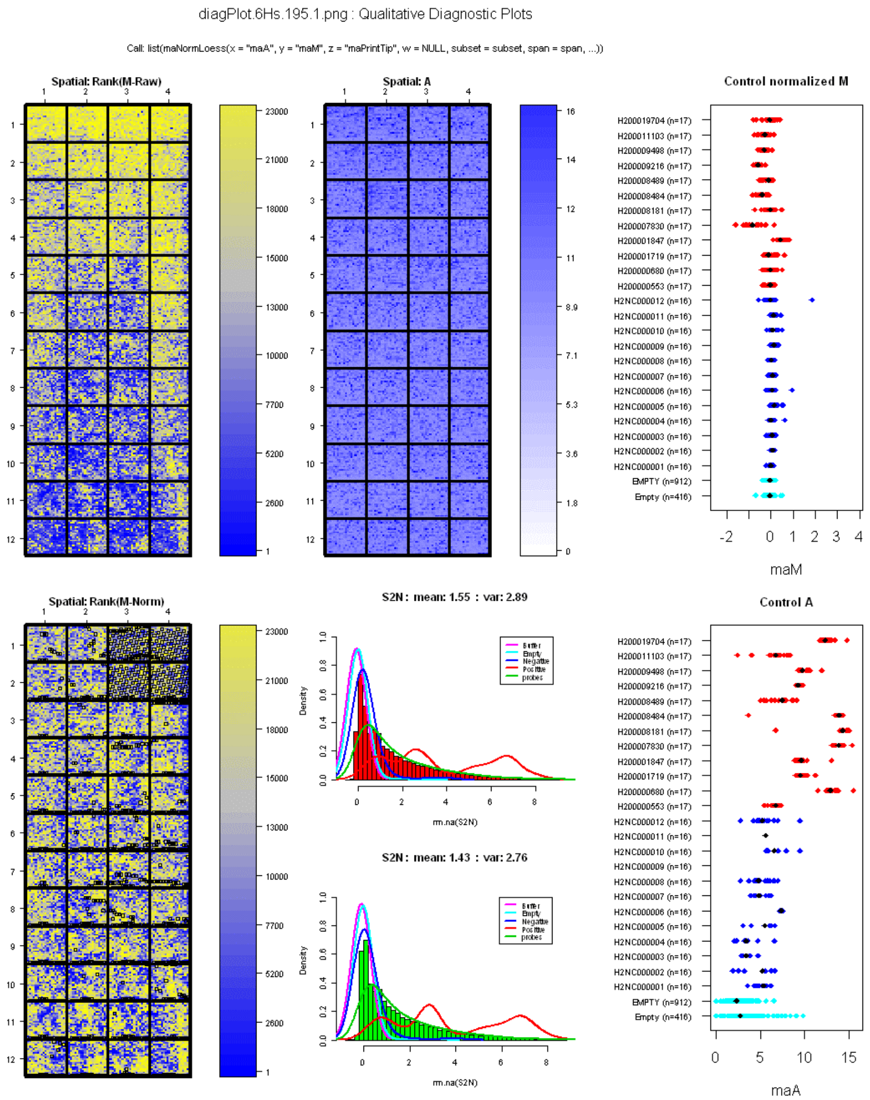
Print-tip Effects – ECDF plot



Diagnostic plot with *arrayQuality*



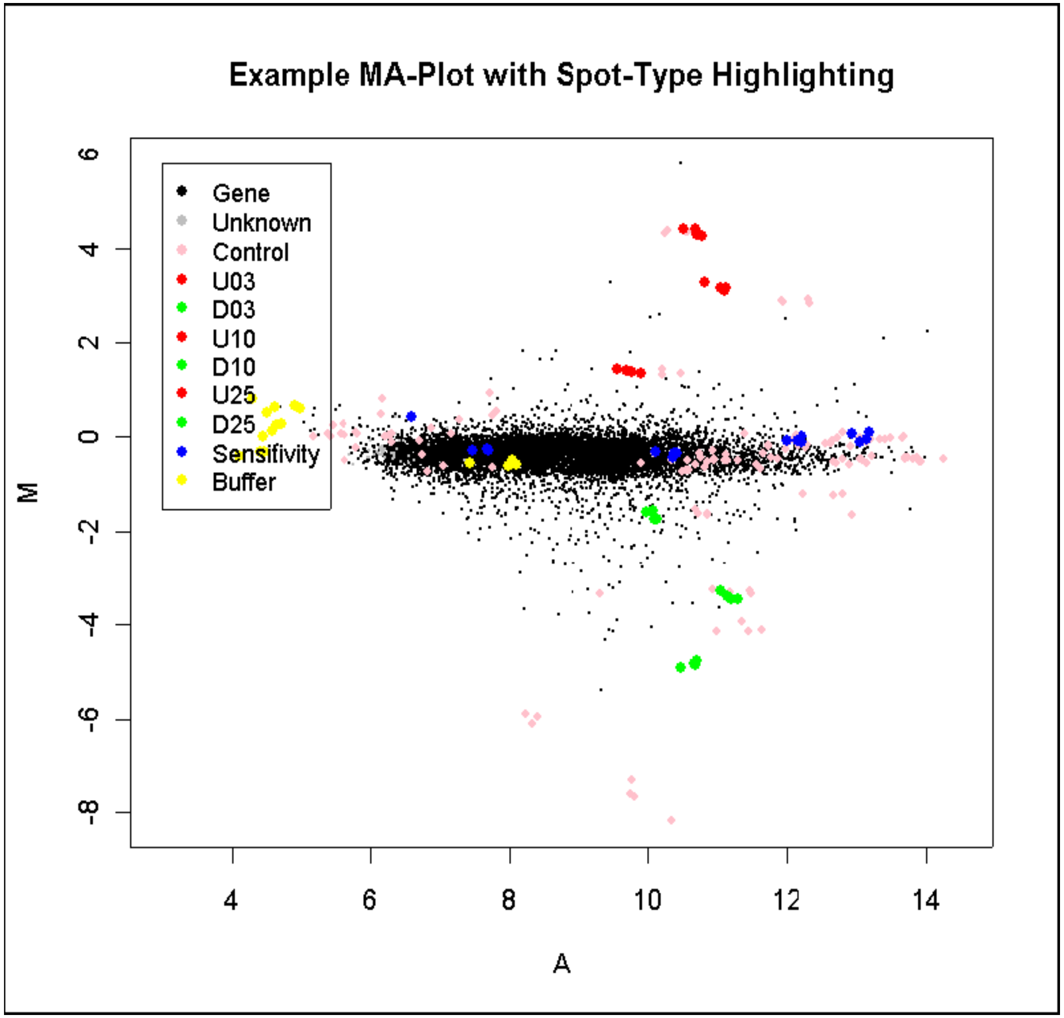




diagPlot.6Hs.195.1.png : Qualitative Diagnostic Plots

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 array:

- Diagnostics plots
- Stratify

BioC packages:

- *arrayQuality*
- *arrayMagic*
- *…*

Outline

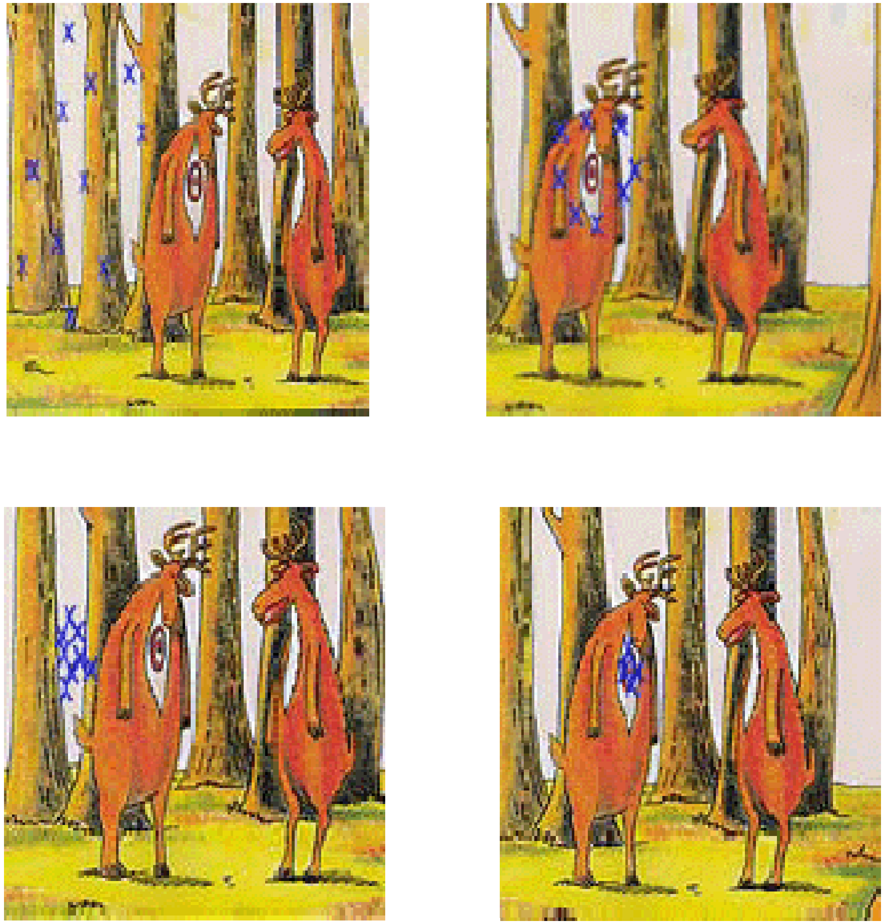
- Data acquisition & Pre-processing (chap. 4)
  - Image analysis
  - Quality assessment
  - **Pre-processing**
- Differential expression (chap. 14, 15 & 23)

- Lab : case studies (chap 4 & 23)
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)
  - limma (G.K Symth)

Variance-Bias trade off

<- Bias Accuray ->

<- Precision Variance ->



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

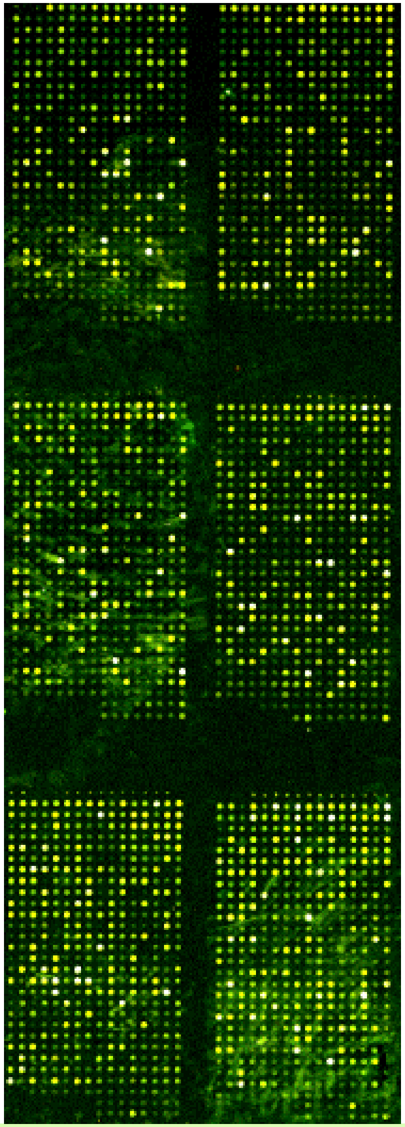
Stochastic

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

Error Model

- …

Background Correction

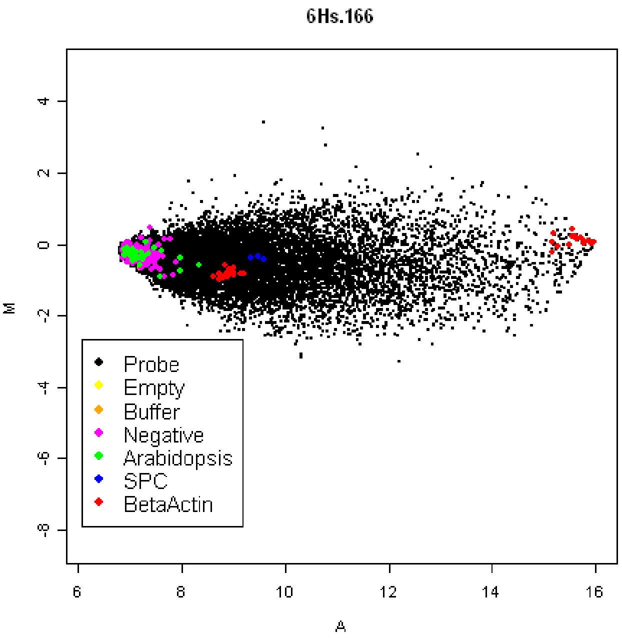


- none
- subtraction, movingmin
- *Minimun,edwards, normexp,…*

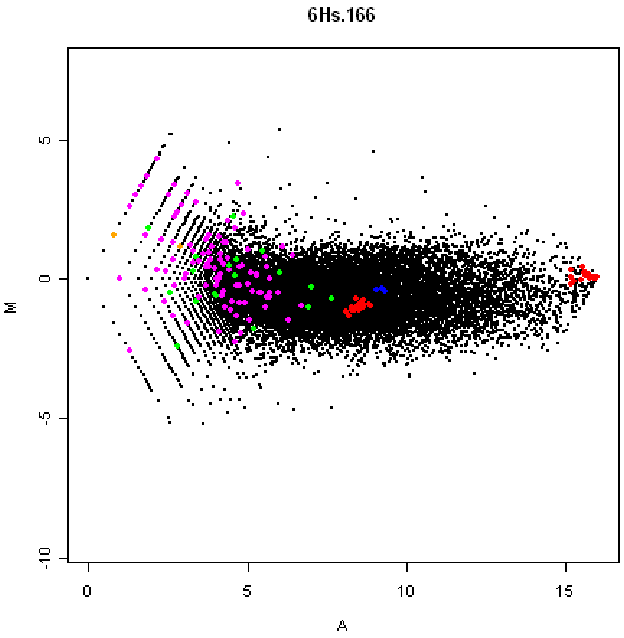
- More details … *limma*

  >?backgroundCorrect

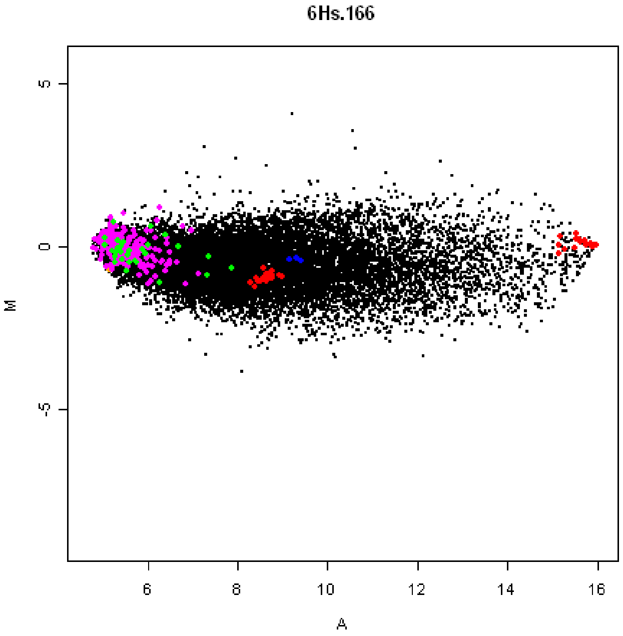
Background Correction



none



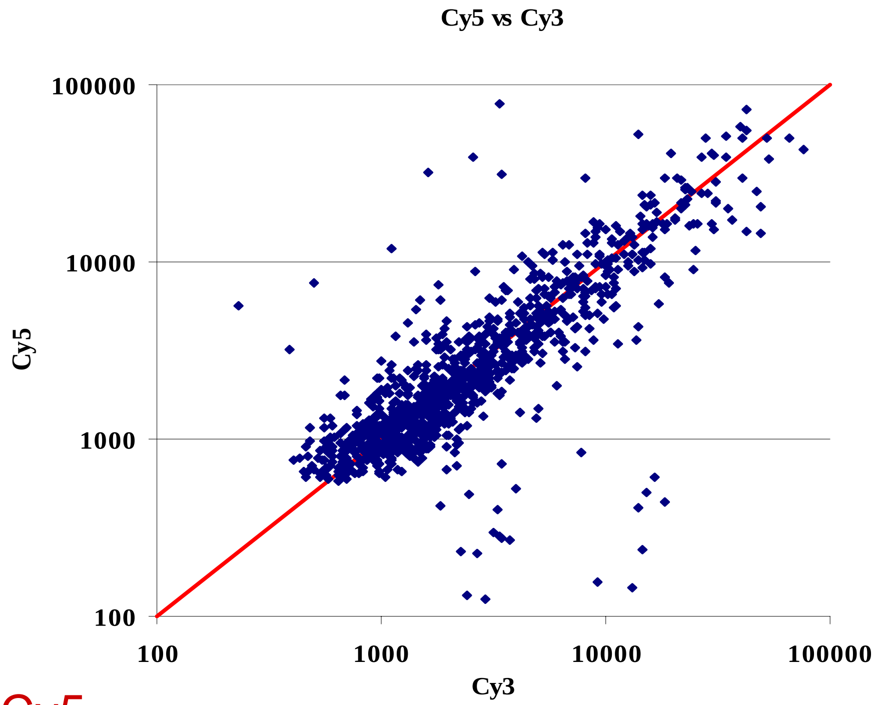
substraction



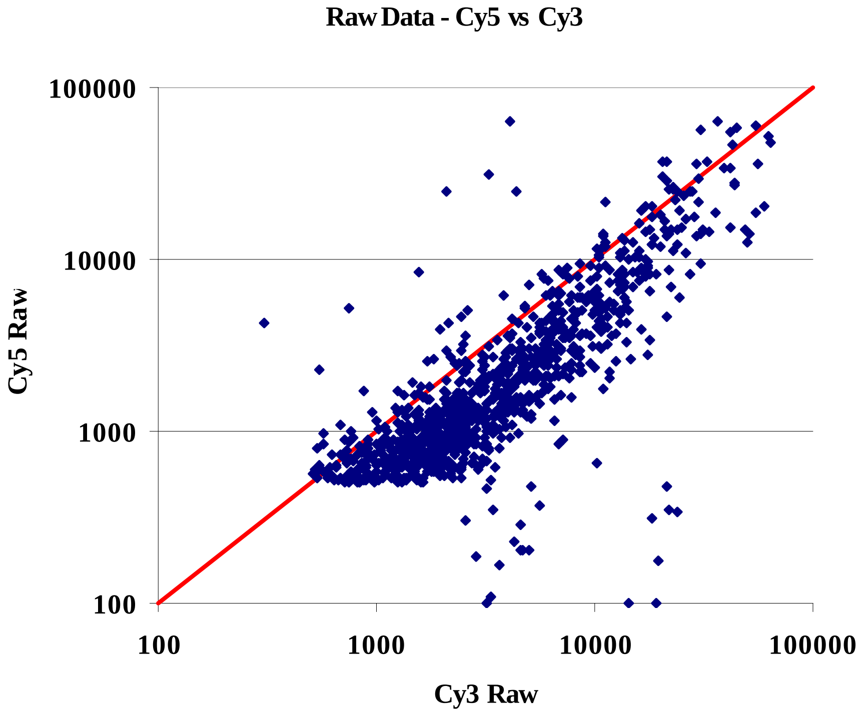
normexp

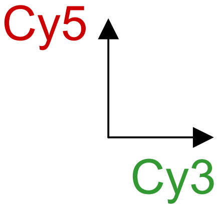
Why Normalize?

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

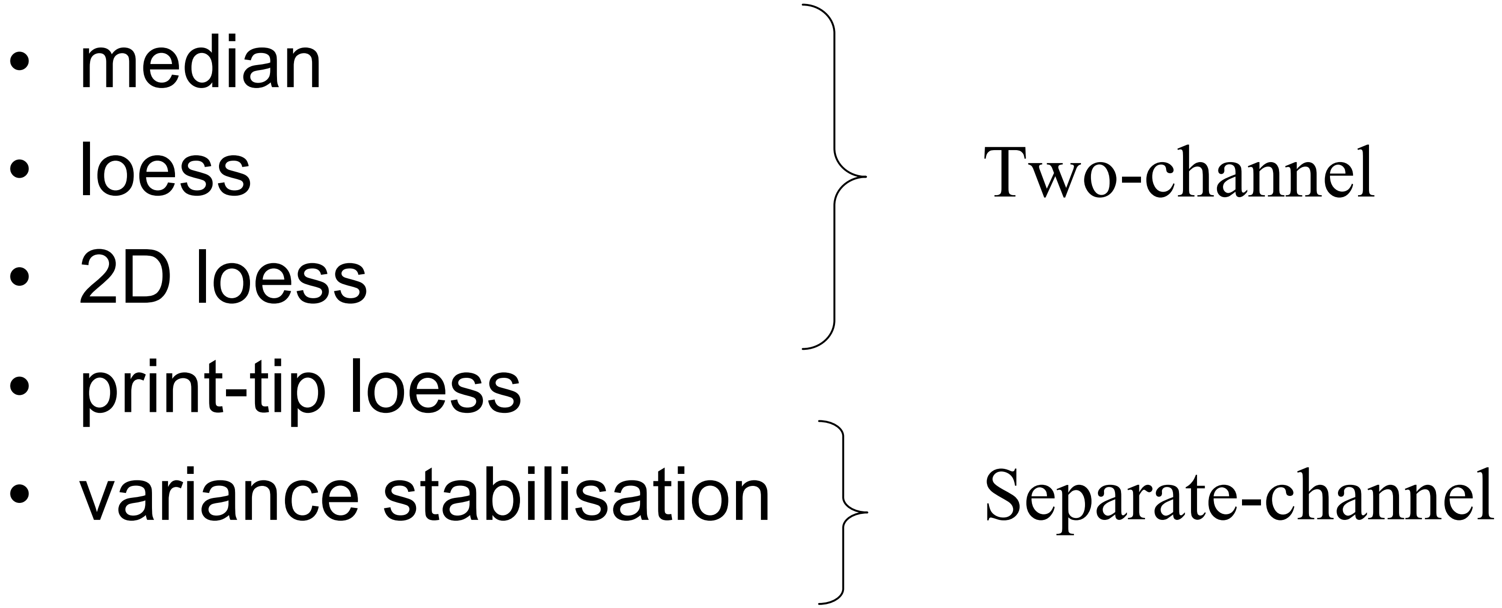
measured intensity = offset + gain × true abundance

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

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

Example: log transformation

Normalization methods



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

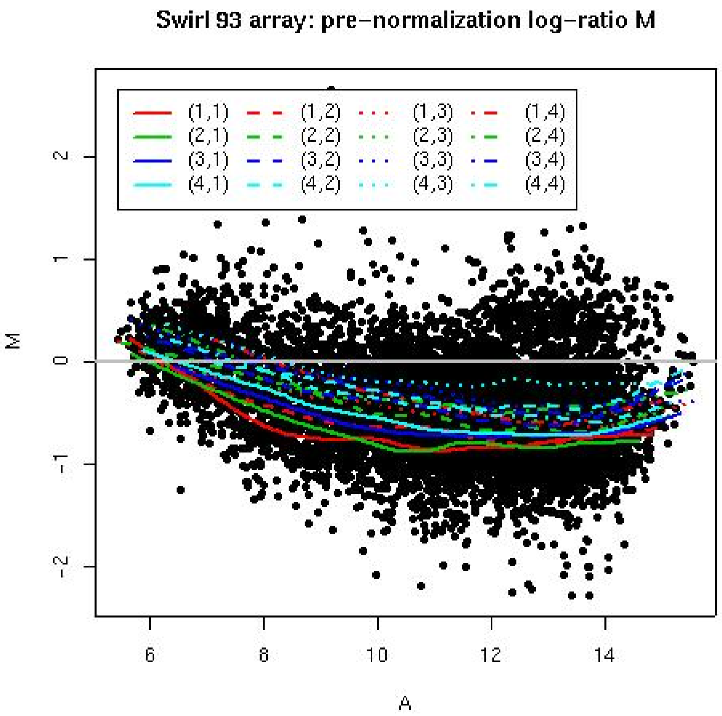
Two-channel

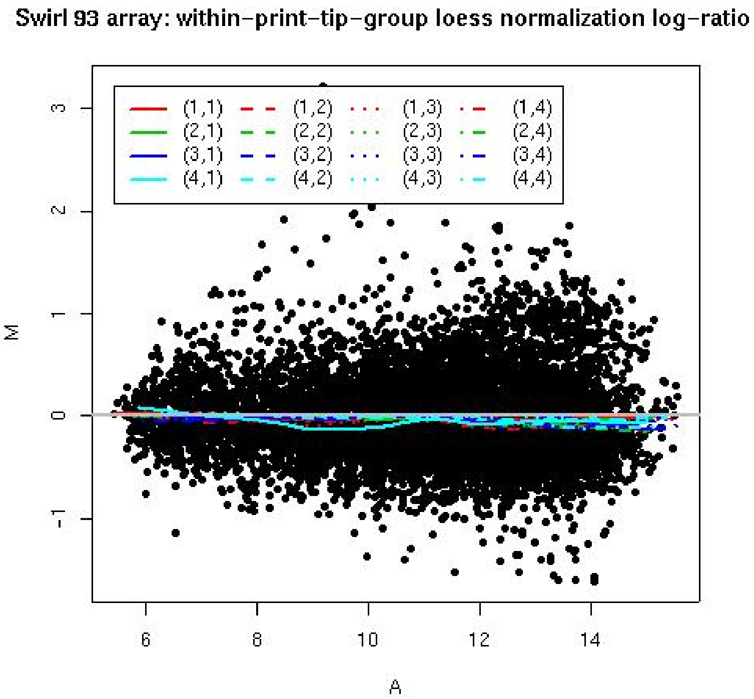
Separate-channel

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

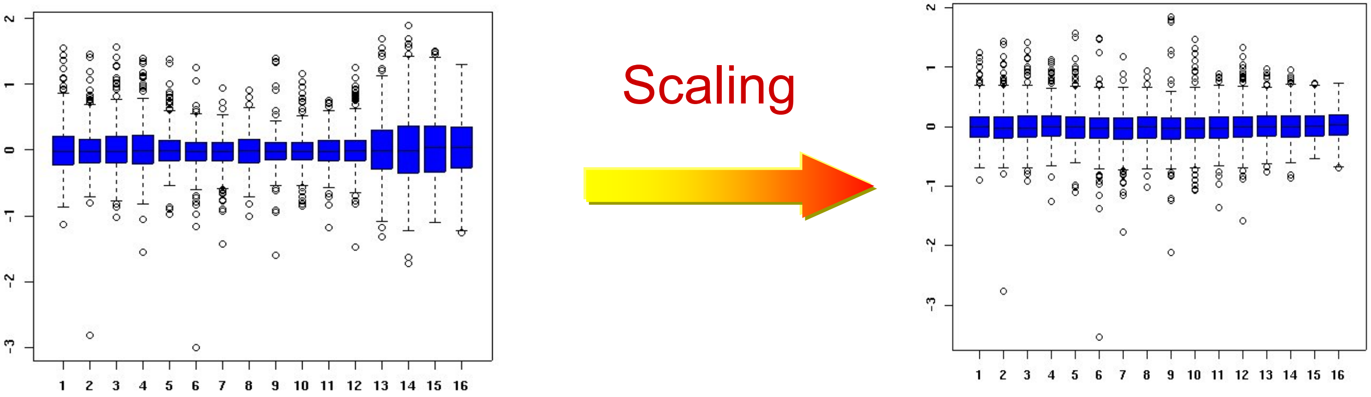

Swirl 93 array: pre−normalization log−ratio M


Swirl 93 array: within−print−tip−group loess normalization log−ratio

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

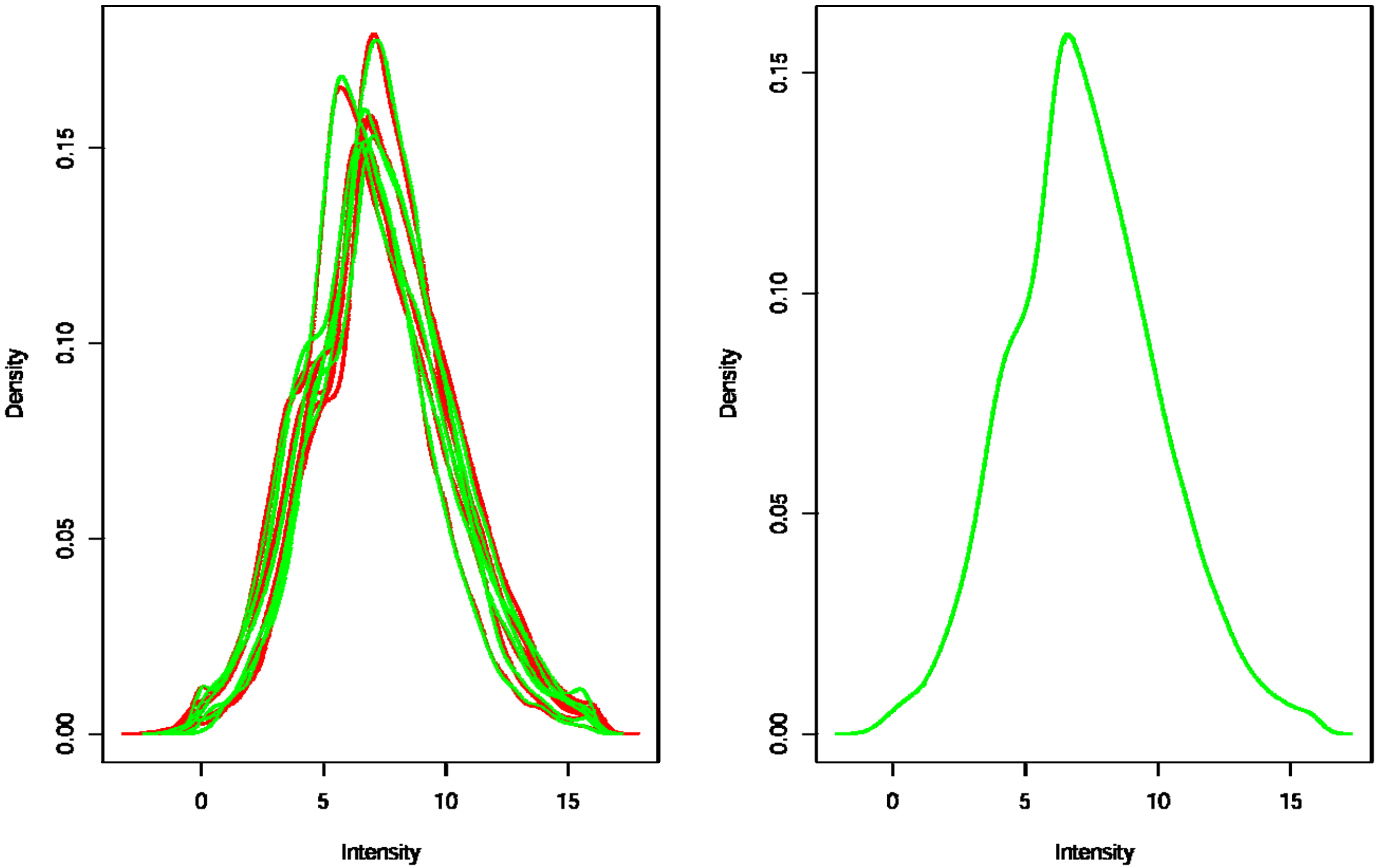
- As technology improves the spot-to-spot varation is reduced

- Development of normalization techniques that work on the absolute intensities

Ex: quantile normalization (*limma*)

variance stabilization (*vsn*)

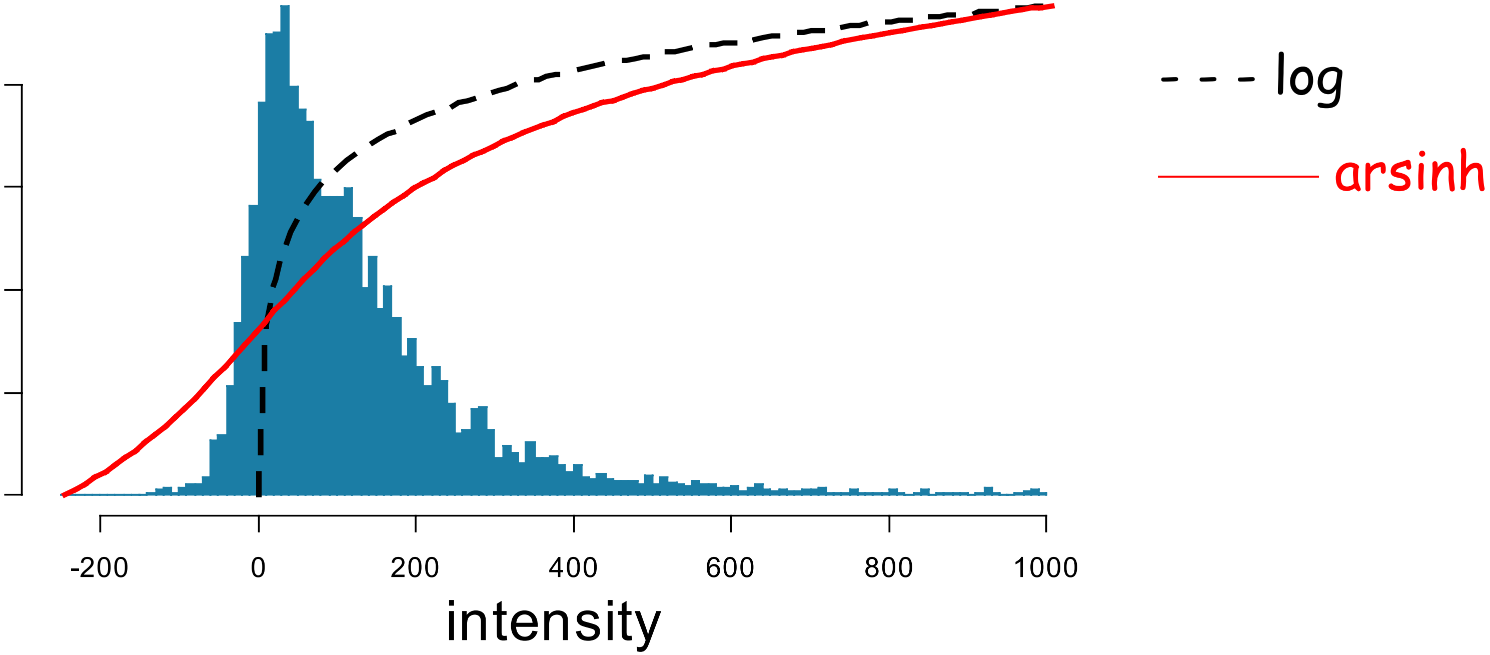
Quantile Normalization


Before

After

Bolstand *et al*.(2003)

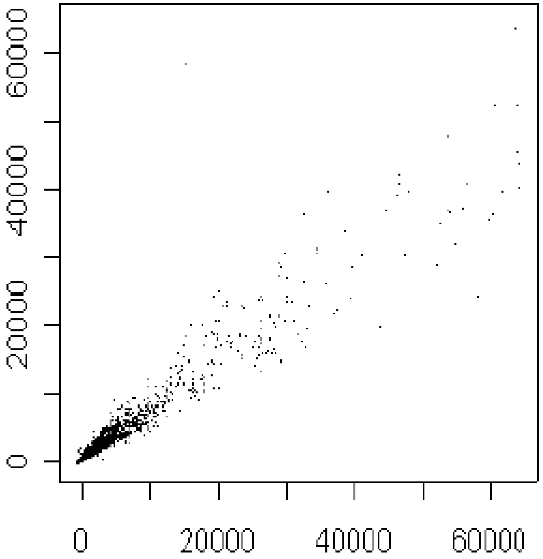
Variance Statibilizing Transformation



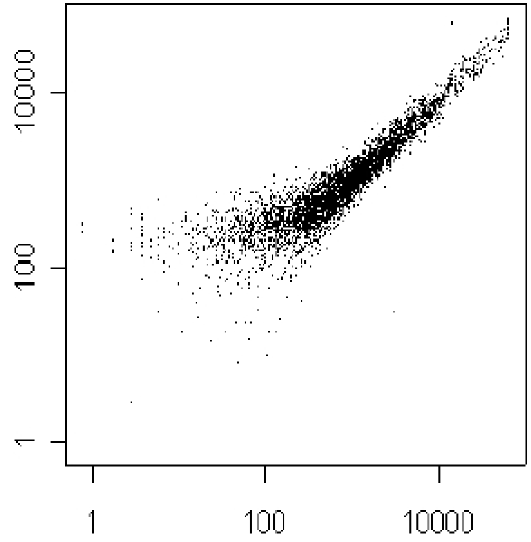
- Meaningful around 0
- Original intensities may be negatives

(Huber *et al.* 2004)

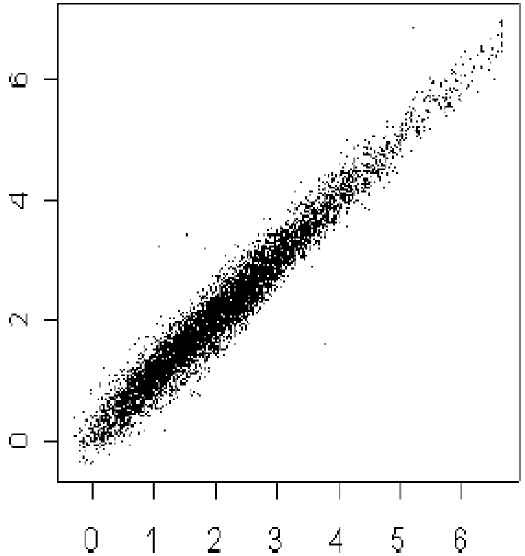
Variance stabilization (*vsn*)



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

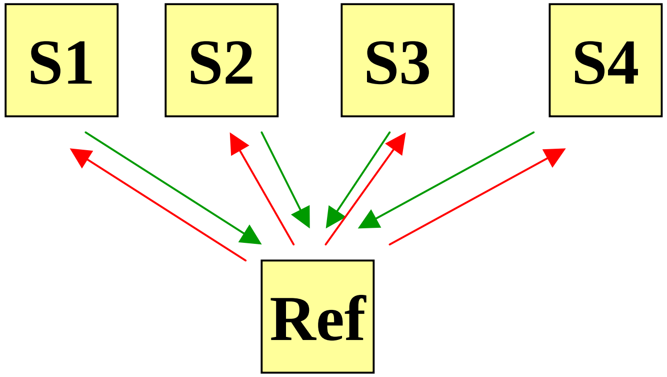
BioC pacakges:

- *marray*
- *limma*
- …

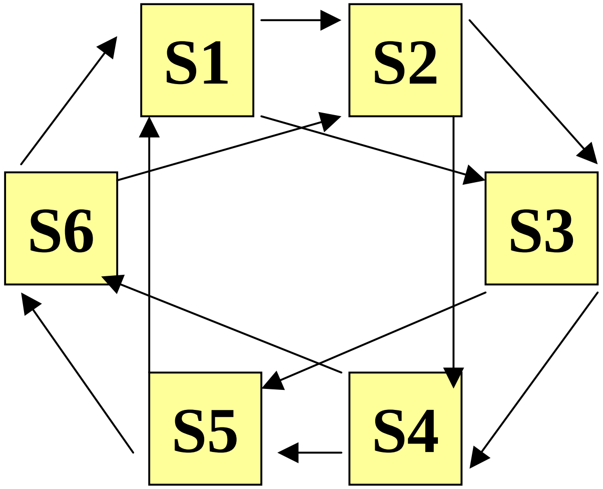
Outline

- Data acquisition & Pre-processing (chap. 4)
  - Image analysis
  - Quality assessment
  - Pre-processing
- **Differential Expression (chap. 14, 15 & 23)**
- Lab : case studies (chap 4 & 23)
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)
  - limma (G.K Symth)

Experimental Designs



Reference design with dye swap



Loop

Avoid Confounding effect

Yang, Y. H. et Speed, T. (2002). Design issues for cDNA microarray experiments. *Nat.Rev.Genet.*, **3**: 579-588.

Experimental Designs

- Simple comparisons
- Technical replicates
- Dye swap
- Within array replicate spots
- Two groups
- Several groups
- Direct two color designs
- Factorial design
- Time Course
- …

Case Studies Chap. 23

Differentially Express Genes

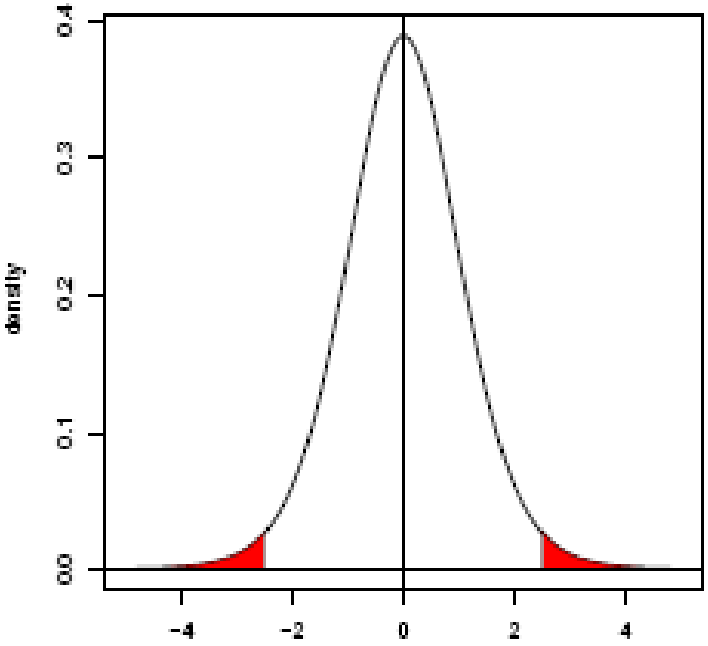
- Fold change

But no assessment of statistical significance

Differentially Express Genes

Example: The two–sample t–statistic is used to test equality of the group means μ_1, μ_2.

The *p–value* p_g is the probability under the null hypothesis (here: $\mu_1 = \mu_2$) that the test statistic is at least as extreme as the observed value T_g. Under the null hypothesis, $\Pr(p_g < \alpha) = \alpha$.



Differentially Express Genes

- Fold change
- Parametric test
  - standard *t*-test
  - Welch *t*-test
- Non parametric
  - Wilcoxon test
  - Mann-Whitney
- Permutation test

Multiple testing

Number of genes	Gene significance level			
	P-values < 0.01	0.05	0.1	0.15
10	< 1	< 1	1	1.5
20	< 1	1	2	3
1000	10	50	100	150
5000	50	250	500	750
10000	100	500	1000	1500

Test of Thousands of hypotheses simultaneously!

➢ Increased chance of false positives

Drăghici (Chapman &Hall 2003)

Individual p-values of 0.01 no longer correspond to significant findings.

-> Adjust for multiple testing

Nonspecific filtering

- Remove genes :
  - Low intensities
  - Do not show sufficient variation across all samples
- Select genes :
  - Known to interact in a specific biological process, e.g. GO (Chap 14.)

Type of Error

	*H*o is true	*H*o is false
*H*o not rejected	True negatives 1- α	False negatives (Type II error) β
*H*o rejected	False positives (type I error) α	True positives (Power) 1-β

Control of Error

- Type II error or Minimizing False negatives

  ->power of tests, sample size

- Type I error

  -> Control false positive rate (FWER,FDR) *or p-value*

  – **F**amily **W**ise **E**rror **R**ate

    control probability of false positive on entire set of genes

  – **F**alse **D**iscovery **R**ate

    control false discovery rate on set of identified genes

Control of Type Error I

Control	Method	Pros/Cons
FWER	Bonferroni	Very conservative
	Šidák	Very conservative
	Holm	Assumption free, conservative
	Hochberg	Independent variables
	Modified Westfall & Young	Exploit *joint* distribution of test statistics, need replicates
FDR	Benjamini & Hochberg	Independent variables
	Benjamini & Yekutieli	conservatives
	Tusher	Sensitive to the number of replicates

Ge, Y & Dudoit, S. (2003) Technical report #633

FWER vs FDR

- **FWER** if high confidence in **all** selected genes is desired. Loss of power due to large number of tests: many differentially expressed genes may not appear significant.

- If a certain proportion of **false positives** is **tolerable**: Procedures based on **FDR** are more flexible; the researcher can decide how many genes to select, based on practical considerations

Moderated t-statistics

- **t–test estimate** the variance of each gene individually.
  - > Ok if we have enough replicates,
  - but with few replicates (say 2–5 per group), these variance estimates are highly variable.
- **moderated t–statistic**, the estimated gene–specific variance s_g^2 is replaced by a weighted average of s_g^2 and s_0^2 , which is a global variance estimator obtained from pooling all genes.

This gives an interpolation between the t–test and a fold–change criterion.

Examples: packages *limma, siggenes*

limma moderated t-statistic

- complex experiments: linear models, contrasts
- empirical Bayes methods for differential expression: t-tests, F-tests, posterior odds
- inference methods for duplicate spots, technical replication
- control of FDR across genes and contrasts

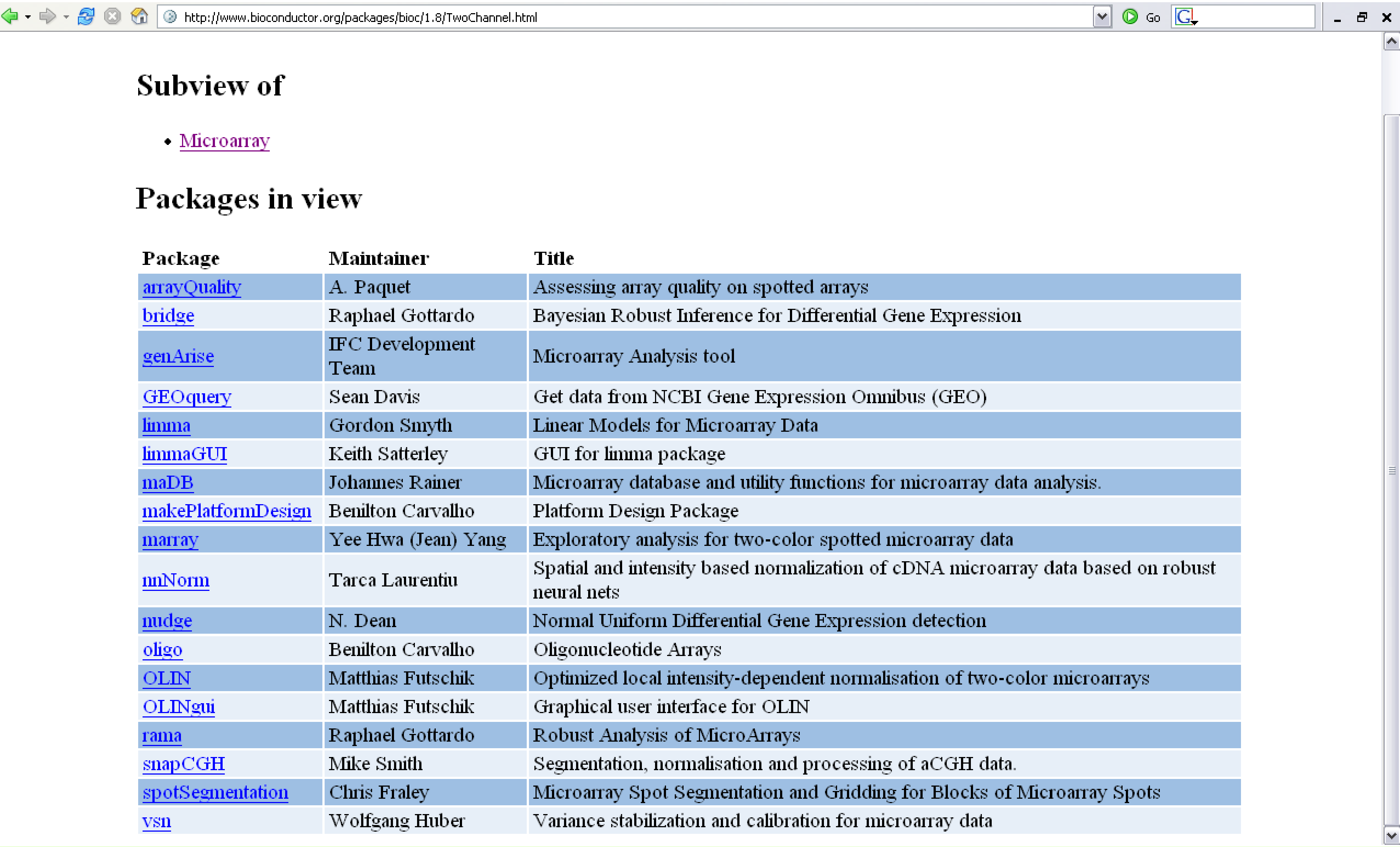
Differential Expr. : Summary

- Permutation tests
- Multiple testing
- Pre-filtering or subsetting
- Rank genes

BioC pacakges:

- *limma*
- *multtest*
- …

BioC Task View: TwoChannel



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

Outline

- Data acquisition & Pre-processing (chap. 4)
  - Image analysis
  - Quality assessment
  - **Pre-processing**
- Differential expression (chap. 14, 15 & 23)

- **Lab : case studies (chap 4 & 23)**
  - marray & arrayQuality (Y.H Yang & A.C. Paquet)
  - limma (G.K Symth)

Getting started

	Preprocessing			
	***limma* package**		***marray* package**	
Action	**Function**	**Class - Object**	**marray**	**Class - Object**
read target file	readTargets	*dataframe*	read.marrayInfo	marrayInfo
read image file	read.maimages	RGList	read.marrayRaw, read.GenePix, read.Spot, read.SMD, read.Agilent	marrayRaw
read gene list	readGAL	RGList$genes	read.Galfile	marrayInfo, marrayLayout
read spot type	readSpotTypes, controlStatus	RGList$genes$status		
array layout	getLayout	RGList$printer	read.marrayLayout, Layout	marrayLayout
background correction	backgroundCorrect			
one array normalization	normalizeWithinArrays, MA.RG	MAList	maNormMain	marrayNorm
normalization between arrays	normalizeBetweenArrays	MAList		