

## cDNA Microarray Quality Assessment and Quality Control with BioConductor packages

Nolwenn Le Meur

January 2007 Copyright 2007



### Outline

### Data acquisition & QA/QC Pre-processing

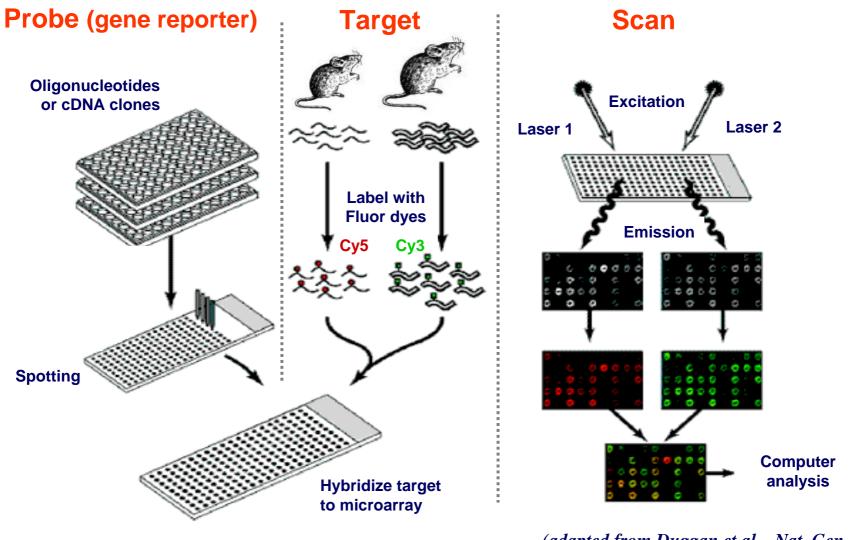
- Image analysis
- Quality Assessment
- Pre-processing and Quality Control







### **Two-color Microarray**





(adapted from Duggan et al., Nat. Gen., 1999) FRED HUTCHINSON CANCER RESEARCH CENTER A LIFE OF SCIENCE

## 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.

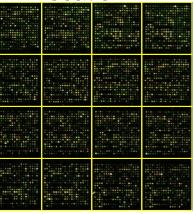
- **G**, **Gb**: Cy3 signal and background intensities
- **R**, **Rb**: Cy5 signal and background intensities
- $\mathbf{M} = \log 2(\mathbf{R}) \log 2(\mathbf{G})$
- A = 1/2(log2(R) + log2(G))





### **Image Analysis**





A (A) (A) 🕲 🎕 🌒 🧶 😑 🚳 🚳 🖲

0 0 💿 💿 🖕 0 💿 O O O O 💿 🕺 O • 0 • 0

0 0 0 0 0

0 0

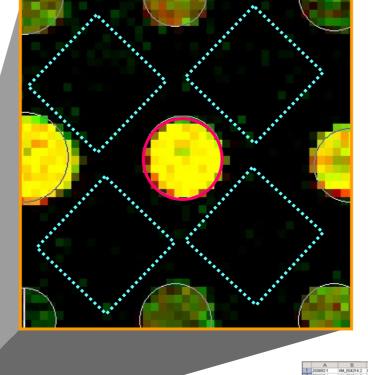
0 () 0 0 0 0 0

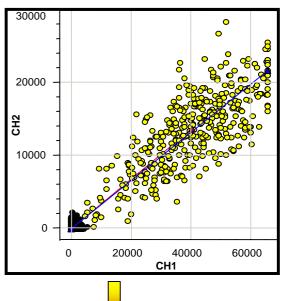
0 ΟΦΦΦΦ • • • • 🕘 💿 💿 💿 💿 💿 💿 💿 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 💩 💩 🖕 🔘 🔘 🔘 🚳 🍓

#### 2. Segmentation

#### 3. Quantification

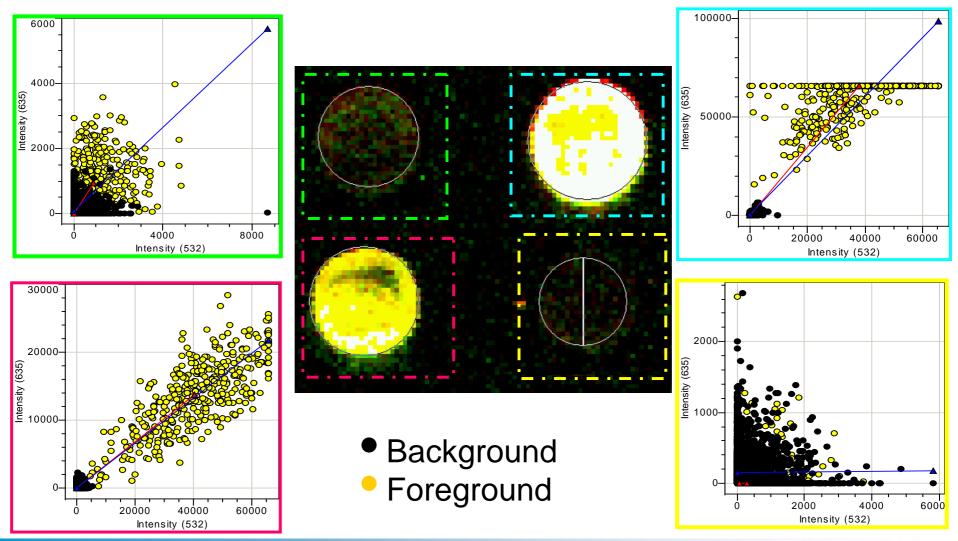




|   | A           | B             | C            | D           | E         | F                  | G           | H           | 1            | J             | K           | L           | M        | N          |
|---|-------------|---------------|--------------|-------------|-----------|--------------------|-------------|-------------|--------------|---------------|-------------|-------------|----------|------------|
|   | 209092.1    | XM_054214.2   | XM_041018.1  | XM_030011.2 | X63432    | \$42658.1          | NM_006471.1 | NM_005159.1 | [NM_003090.1 | NM_001825.1   | NM_001101.2 | NM_000258.1 | M14603.1 | 1,301415.1 |
|   | Z69092.1    | XM_054214.2   | XM_040948.2  | XM_029192.1 | X63432    | NM_032169.1        | NM_006471.1 | NM_005159.1 | NM_002007.1  | NM_001025.1   | NM_001101.2 | NM_000257.1 | M11146.1 | 1 J01415.1 |
| l | Z24725.1    | XM_053038.1   | XM_040948.2  | XM_029192.1 | X80819.1  | NM_032169.1        | NM_006471.1 | NM_005159.1 | 3NM_002807.1 | NM_001825.1   | NM_001100.2 | NM_000257.1 | L39210.1 | J01415.1   |
|   | Z24725.1    | XM_053038.1   | XM_039448.1  | XM_028372.1 | X50819.1  | 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   |
| ī | 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   |
| ı | Z15030.1    | XM_052918.1   | 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 | .301415.1  |
|   | XM_058357.1 | XM_052627.2   | XM_030278.3  | XM_016697.2 | X16009.1  | NM_016160.1        | NM_000111.1 | NM_005061.1 | NM_002799.1  | NM_001689.1   | NM_001098.1 | NM_000256.1 | L32035.1 | .301415.1  |
| ī | XM_058357.1 | XM 052627.2   | XM 038027.1  | XM 016290.2 | X15869.1  | NM_014819.1        | NM 006044.1 | NM 005005.1 | NM 002715.1  | NM_001689.1   | NM 001035.1 | NM 000255.1 | L32835.1 | .J01415.1  |
| ŀ | 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   |
| ŀ | 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_001008.1 | NM_000237.1 | L07782.1 | J01415.1   |
|   |             | XM_051945.1   | XM_037923.1  | XM_016190.2 | X16009.1  | NM_014713.1        | NM_006007.1 | NM_004768.1 | NM_002623.2  | NM_001686.1   | NM_001006.1 | NM_000126.1 | L05007.1 | 301415.1   |
|   |             |               | XM_037923.1  | XM_000989.1 | X16869.1  | NM_014391.1        | NM_006003.1 | NM_004548.1 | NM_002623.2  | NM_001681.1   | NM_000992.1 | NM_000125.1 | L05087.1 | J01415.1   |
|   | XM_057346.1 | XM_051885.3   | 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   |
|   |             |               | XM_037797.21 | XM_007127.2 | X14891.1  | NM_014391.1        | NM_005917.1 | NM_004415.1 | NM_002612.1  | NM_001681.1   | NM_000988.1 | NM_000065.1 | K02043.1 | .301415.1  |
|   | XM_057063.1 | XM_050614.1   | XM_036050.1  | XM_007031.4 | X14091.1  | NM_014391.1        | NM_005917.1 | NM_004415.1 | NM_002521.1  | NM_001620.1   | NM_000905.1 | NM_000019.1 | H02043.1 | 301415.1   |
|   | XM_057063.1 | XM_050614.1   | XM_036858.1  | XM_007031.4 | U9        |                    |             |             |              | 1628.1        | NM_000972.1 | NM_000019.1 | K02043.1 | D79994.1   |
|   | XM_056761.1 | XM_049679.1   | XM_035796.1  | XM_006238.4 | 09        |                    |             |             |              | 1613.1        | NM_000972.1 | NM_000018.1 | K02043.1 | 079994.1   |
|   | XM_058761.1 | XM_049679.1   | XM_035796.1  | XM_005848.2 | -129      |                    |             |             |              | 1613.1        | NM_000970.2 | NM_000018.1 | K02043.1 | 050683.1   |
|   | XM_055059.1 | XM_049575.2   | XM_034179.1  | XM_005848.2 | -U9       |                    |             |             | $\mathbf{x}$ | H613.1        | NM_000970.2 | NM_000016.1 | H02043.1 | 050683.1   |
|   | XM_055859.1 | XM_049131.2   | XM_034179.1  | XM_005417.4 | U9        |                    |             | da          | 110          | 11553.1       | NM_000919.1 | NM_000016.1 | 303620.1 | D30648.1   |
|   | XM_055793.1 | XM_049131.2   | XM_034146.2  | XM_005417.4 | 09        |                    |             | MU          |              | )1553.1       | NM_000919.1 | NC_001807.4 | J03620.1 | D28908.1   |
|   | M_055793.1  | XM_046843.1   | XM_034146.2  | XM_004377.3 | UBS       |                    |             |             |              | 11450.1       | NM_000587.1 | NC_001807.3 | 303015.1 | D28908.1   |
|   | XM_055602.1 | XM_046043.1   | XM_034036.1  | XM_003317.4 | Ulli one  | and some states of |             |             |              | ren_sa)1450.1 | NM_000507.1 | NC_001807.3 | 303015.1 | 020900.1   |
|   | XM_055682.1 | 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 | 301415.1 | D17409.1   |
|   | XM_055602.1 | XM_046056.2   | XM_033374.1] | XM_003317.4 | U62138.2  | NM_007361.1        | NM_005530.1 | NM_003319.1 | NM_002300.1  | NM_001450.1   | NM_000543.1 | M94859.1    | 301415.1 | D17409.1   |
|   | 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_000368.1 | M94859.1    | 301415.1 | 010040.1   |
|   | XM_055545.1 | XM_045954.1   | XM_032396.1  | XM_002862.4 | U49020.1  | NM_007159.1        | NM_005360.1 | NM_003319.1 | NM_002156.1  | NM_001450.1   | NM_000366.1 | M64247.1    | 301415.1 | 010040.1   |
|   | XM_055545.1 | XM_044022.1   | XM_032004.1  | XM_002852.4 | U40490.1  | NM_007107.1        | NM_005368.1 | NM_003319.1 | NM_002156.1  | NM_001402.1   | NM_000366.1 | M64247.1    | 301415.1 | D00943.1   |
|   | XM_055358.1 | XM_044022.1   | XM_032004.1  | XM_002659.3 | U40490.1  | NM_007107.1        | NM_005368.1 | NM_003319.1 | NM_002138.1  | NM_001402.1   | NM_000366.1 | M31776.1    | 301415.1 | 000943.1   |
|   | XM_055358.1 | XM_043689.1   | XM_031823.1  | XM_002659.3 | \$72481.1 | NM_007079.1        | NM_005368.1 | NM_003197.2 | NM_002138.1  | NM_001402.1   | NM_000368.1 | M31776.1    | 301415.1 | 000943.1   |
|   | XM_055266.1 | XM_043669.1   | XM_031023.1  | XM_002601.3 | \$72401.1 | NM_007079.1        | NM_005162.2 | NM_000130.1 | NM_002107.1  | NM_001402.1   | NM_000294.1 | M31776.1    | 301415.1 | 00050053   |
|   | KM_055266.1 | XM_043419.2   | XM_031736.2  | XM_002601.3 | \$69022.1 | NM_006076.1        | NM_005162.2 | NM_000130.1 | NM_002107.1  | NM_001402.1   | NM_000294.1 | M27024.1    | J01415.1 |            |
|   | XM_055102.1 | XM_041875.1   | XM_031736.2  | XM_002556.5 | 589022.1  | NM_006876.1        | NM_005159.2 | NM_003130.1 | NM_002079.1  | NM_001402.1   | NM_000289.1 | M27024.1    | 301415.1 | BC017495   |
|   | M_055102.1  | XM_041869.2 * | XM_031661.1  | XM_002558.5 | \$89022.1 | NM_006793.1        | NM_005159.2 | NM_003130.1 | NM_002079.1  | NM_001232.1   | NM_000289.1 | M26700.1    | 301415.1 | 80017495   |
|   | XM_054049.1 | XM_041069.2+  | XM_031661.1  | X91647.1    | \$69022.1 | NM_006793.1        | NM_005159.2 | NM_000130.1 | NM_001909.1  | NM_001232.1   | NM_000209.1 | M26700.1    | 301415.1 | 00017105   |
|   | XM_054049.1 | XM_041393.1   | XM_030102.1  | X91647.1    | \$69022.1 | NM_006513.1        | NM_005159.1 | NM_003094.1 | NM_001969.1  | NM_001103.1   | NM_000209.1 | M26576.1    | J01415.1 | DC017108   |
|   | XM_054461.1 | XM_041393.1   | XM_030182.1  | X66699.1    | 589022.1  | NM_006513.1        | NM_005159.1 | NM_003094.1 | NM_001885.1  | NM_001103.1   | NM_000258.1 | M26576.1    | 301415.1 | BC017185   |
| 1 | M 054461.1  | XM_041018.1   | XM 030011.2  | X66699.1    | \$42658.1 | NM 006471.1        | NM 005159.1 | NM 003090.1 | NM 001885.1  | NM 001103.1   | NM 000258.1 | M14603.1    | 301415.1 | BC017080   |



۵ ن





A LIFE OF SCIENCE

**HUTCHINSON** 

CENTER

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)





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





For each array:

#### Visual inspection

 hairs, dust, scratches, air bubbles, dark regions, regions with haze

### Diagnostics plots of spot statistics

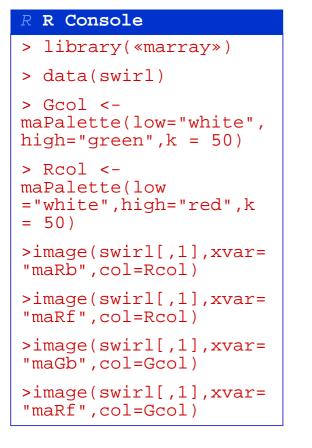
- e.g. R and G log-intensities, M, A, spot area.
  - 2D spatial images;
  - ECDF plots;
  - Boxplots;
  - Scatter-plots;
  - Density plots.
- Stratify plots according to layout parameters, *e.g.* print-tipgroup, plate.

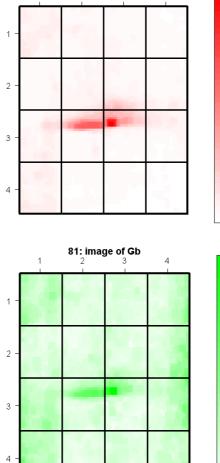




### **Image Plots**

4





81: image of Rb

1



- 280

250

- 230

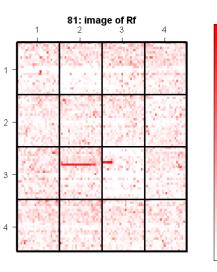
· 200

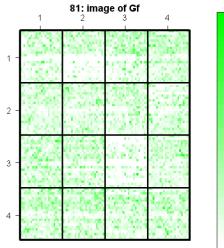
- 180

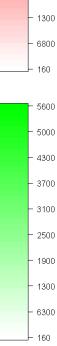
- 150

- 130

· 100 · 76







- 6000

- 5300

- 4700

- 4000

- 3300

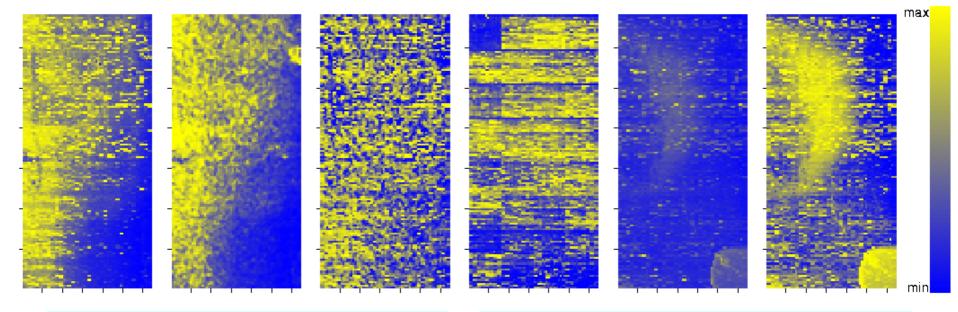
- 2700

- 2000





### **Spatial Effects – Image Plots**



## R Rb R-Rb color scale by rank

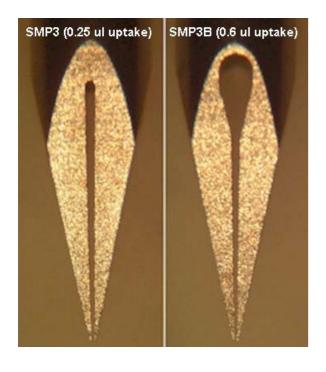
#### Print-tip

#### Washing





### **Spotting Pin Quality Decline**



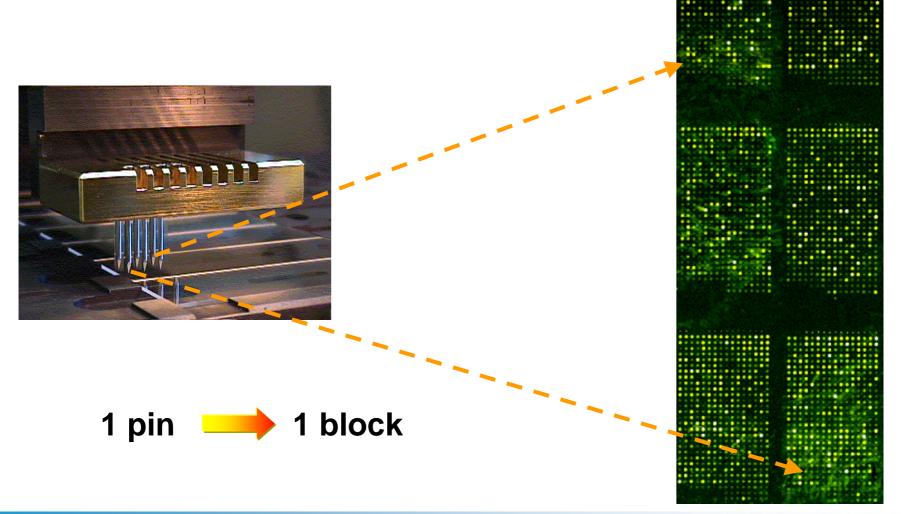
#### after delivery of 5x10<sup>5</sup> spots

### after delivery of 3x10<sup>5</sup> spots





### **Spatial Effects**

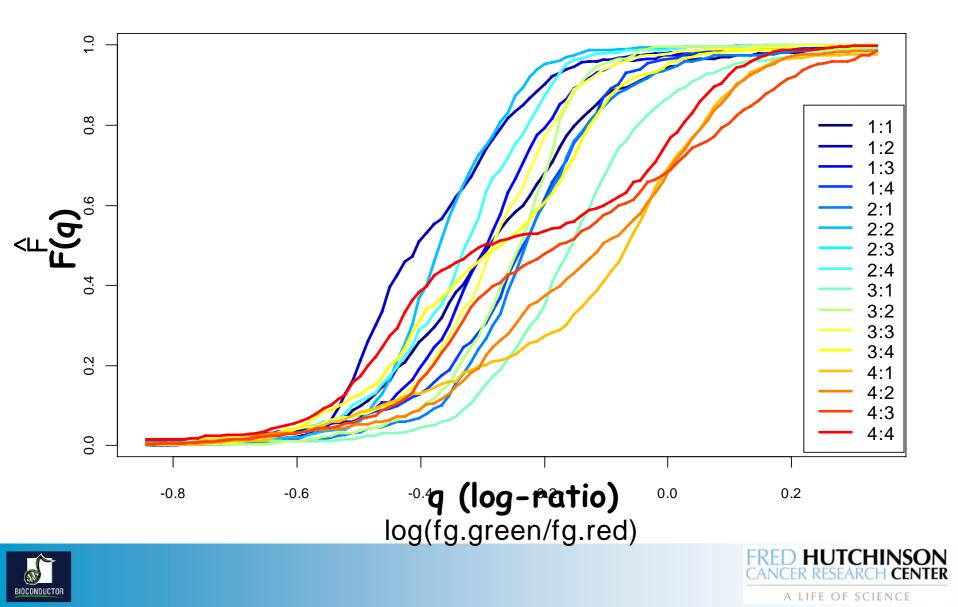




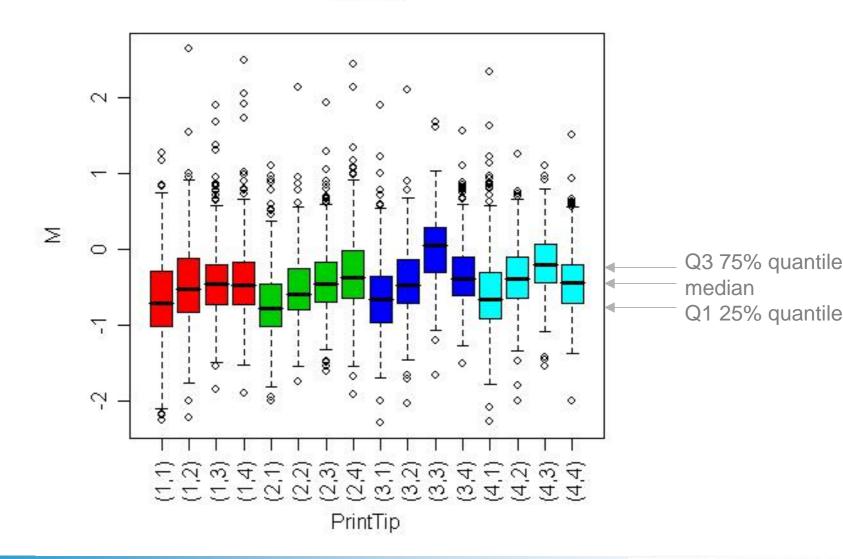


A LIFE OF SCIENCE

### **Print-tip Effects – ECDF plot**



# Print-tip Effects - Boxplots

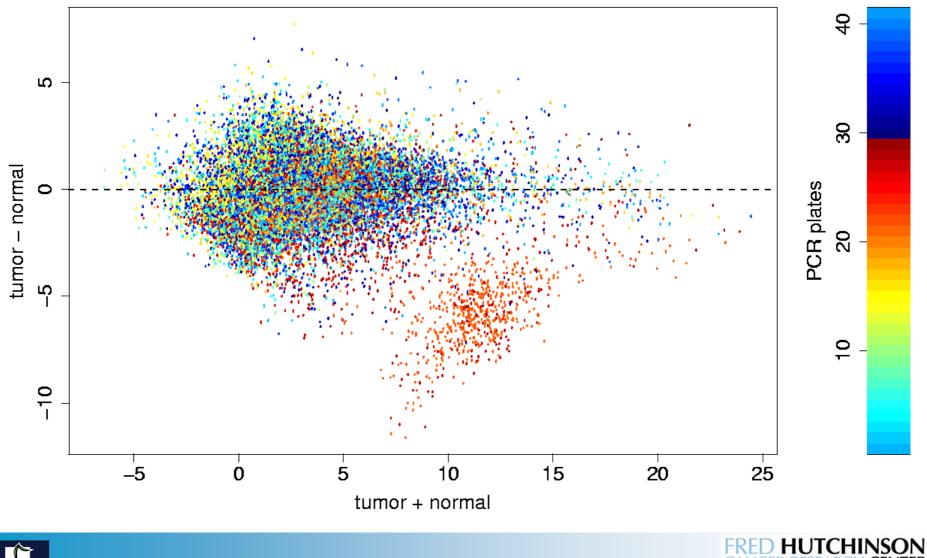


CENTER

A LIFE OF SCIENCE



### **PCR plates**

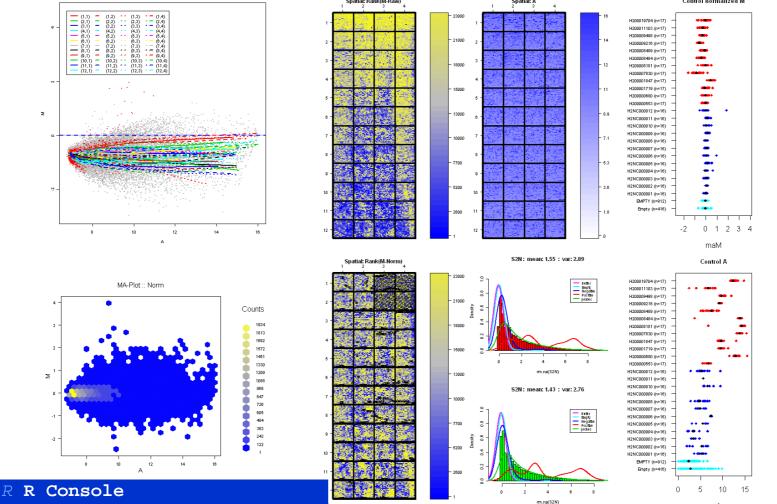




A LIFE OF SCIENCE

CENTER

### Diagnostic plot with the fing Quality Could and the state of the state



library(«arrayQuality»)

?maQualityPlots





A LIFE OF SCIENCE

maA

| FileName                 | DL32          | DL31          | DL30        | HuPr_4005_E |
|--------------------------|---------------|---------------|-------------|-------------|
| СуЗ                      | 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_E |
| 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 R Console

library(«arrayMagic»)

?qualityParameters





#### R Console R distributionOfRawDataSlideWise library(«arrayMagic») quantiles: lower:0.25; middle:0.5; upper:0.75 ?qualityDiagnostics 0 -





### **Quality Assessment: Summary**

#### For each spot:

weight

#### For each array:

- Diagnostics plots
- Stratify
- Controls

#### **BioC packages:**

- arrayQuality
- arrayMagic





### Outline

### Data acquisition & QA/QC Pre-processing

- Image analysis
- Quality Assessment
- Pre-processing and Quality Control







### **Sources of Variation**

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

PCR

- DNA quality
- Spotting efficiency
- cross-hybridization

### Systematic

similar effect on many measurements
corrections can be estimated from data

### Calibration

### Stochastic

- too random to be explicitely accounted for
- "noise"

#### **Error Model**





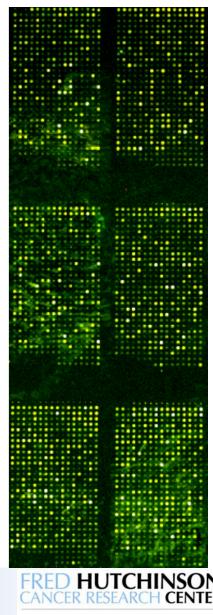
### **Background Correction**



subtraction, movingmin

Minimun,edwards, normexp,...

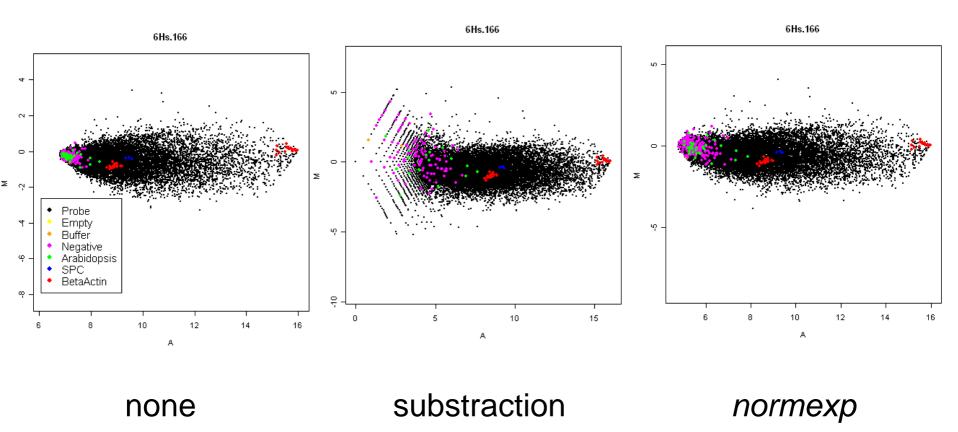
More details ... *limma* >?backgroundCorrect



A LIFE OF SCIENCE



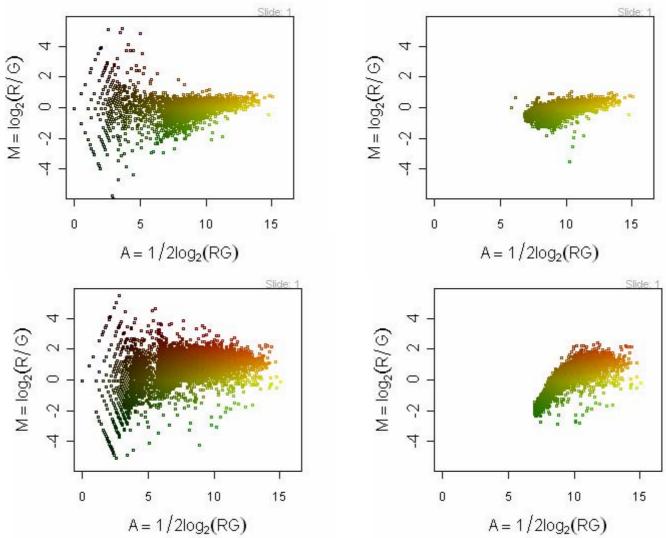
### **Background Correction**







### **Background Correction**





FRED HUTCHINSON CANCER RESEARCH CENTER A LIFE OF SCIENCE

### Normalization

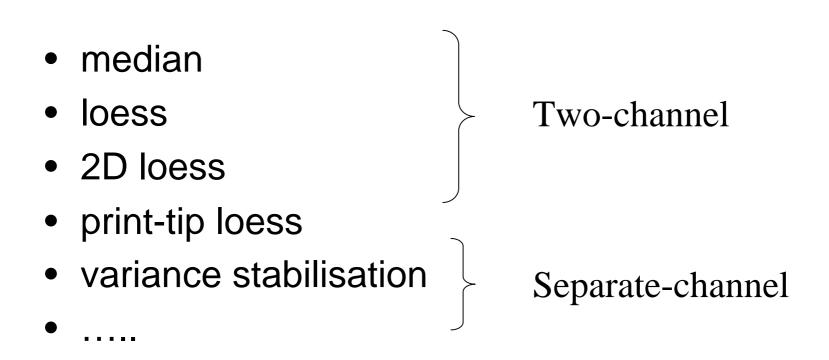
Identify and remove the effects of systematic variation

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





### **Normalization methods**



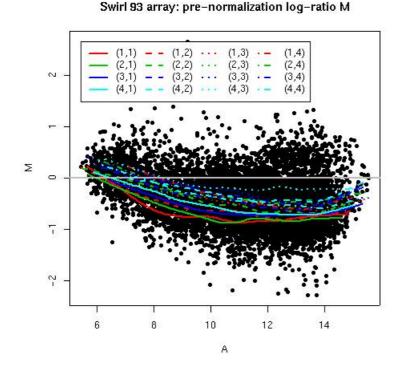
Smyth, G. K., and Speed, T. P. (2003). In: *METHODS: Selecting Candidate* Genes from DNA Array Screens: Application to Neuroscience



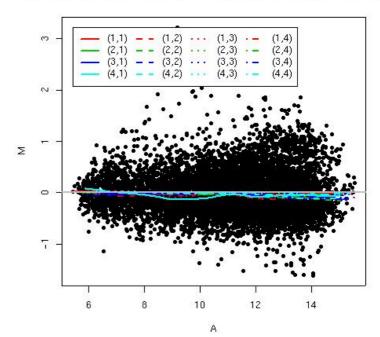


### **Two channel normalization**

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



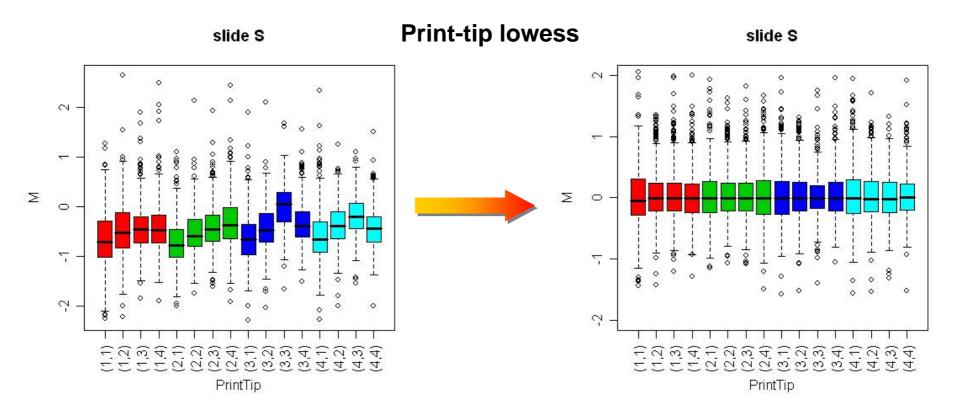
Swirl 93 array: within-print-tip-group loess normalization log-ratio







### **Two channels normalization**



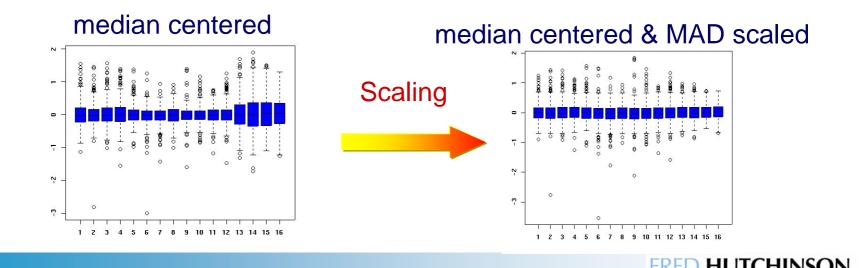




### **Two channels normalization**

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

Scale: adjust for different in scale between multiple arrays



A LIFE OF SCIENCE



### **One channel normalization**

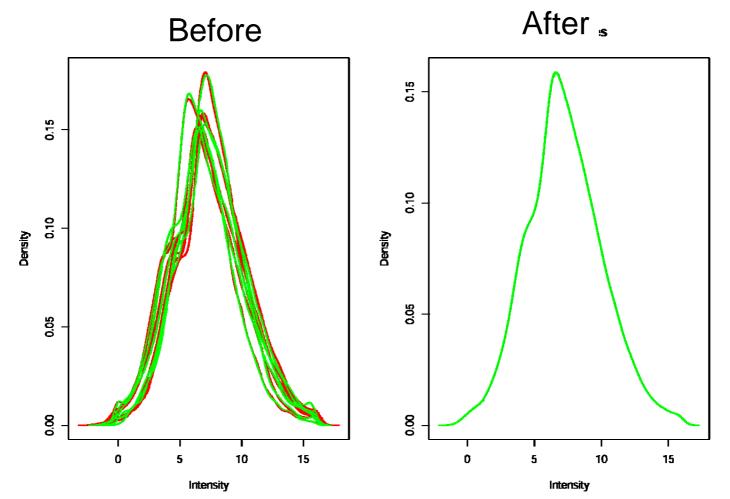
- As technology improves the spot-to-spot varation is reduced
- Development of normalization techniques that work on the absolute intensities

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





### **Quantile Normalization**



Bolstand et al.(2003)

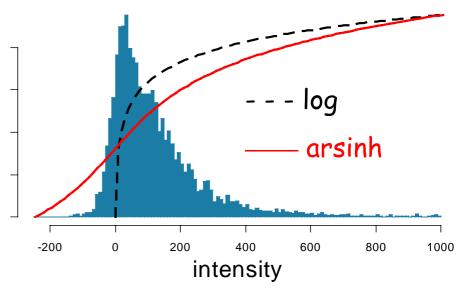


**FRED HUTCHINSON** CANCER RESEARCH CENTER

### Variance Stabilization Transformation

Iog-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

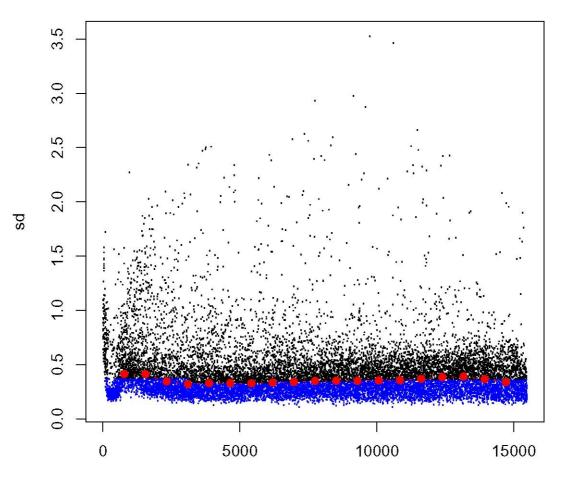


(Huber et al. 2004)

A LIFE OF SCIENCE



### **Variance Stabilization**



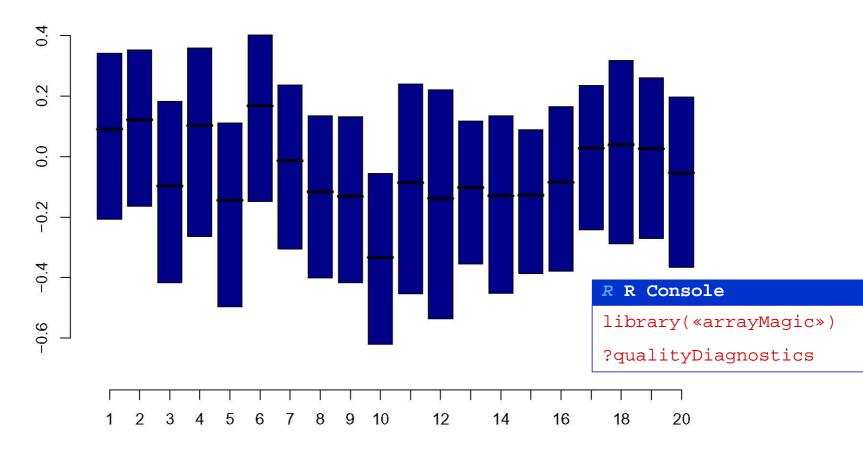
rank(mean)





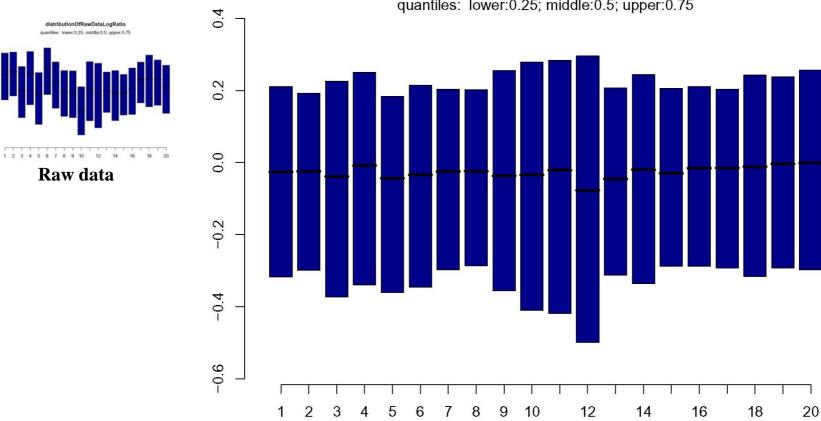
#### distributionOfRawDataLogRatio

quantiles: lower:0.25; middle:0.5; upper:0.75









distributionOfNormalisedDataLogRatio

quantiles: lower:0.25; middle:0.5; upper:0.75

vsn normalized data

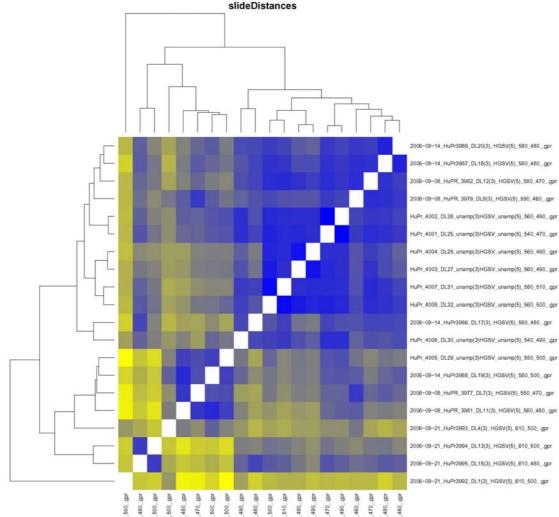




slideDistances 20 19 18 0.40 17 -16 15 0.35 14 13 -12 0.30 11 10 9 25 8 7 6 0.20 5 4 3 -0.15 2 -1 -







92\_DL1(3)\_

DL15(3)\_\_\_\_\_

DL11(3)

DL4(3).

77\_DL7(3)\_

DL19(3)

DL17(3)\_

GSV

DL12(3).

GSV

FRED HUTCHINSON CANCER RESEARCH CENTER



### **Preprocessing : Summary**

#### For each array:

- Background correction or not
- Normalization
- Diagnostic plots QA/QC

#### BioC pacakges:

- arrayQuality
- arrayMagic
- vsn
- limma

• • • •





### **BioC Task View: TwoChannel**

🖌 🔘 Go 🚺

24 packages (18 Bioc1.8)

- 8 ×

A LIFE OF SCIENCE

~

🕶 🗣 🚽 🔁 💿 🏠 💿 http://bioconductor.org/packages/1.9/TwoChannel.html

#### Subview of

• <u>Microarray</u>

#### Packages in view

| Package         Maintainer         Title           aroma_light         Henrik Bengtsson         Light-weight methods for normalization and visualization of microarray data using only basic           arrayMagic         Andreas Buness         two-colour cDNA array quality control and preprocessing           arrayQuality         A. Paquet         Assessing array quality on spotted arrays | R data types |
|---|--------------|
| arrayMagic     Andreas Buness     two-colour cDNA array quality control and preprocessing       arrayQuality     A. Paquet     Assessing array quality on spotted arrays  | R data types |
| arrayQuality A. Paquet Assessing array quality on spotted arrays  |              |
|   |              |
| the Assess CND Les Contine Neuralisation of Assessmine CND has domained   |              |
| <u>beadarraySNP</u> 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 n   | iets         |
| 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   |              |





