Microarray Data Analysis (V)

Preprocessing (i): two-color spotted arrays

Preprocessing

- Probe-level data: the intensities read for each of the components.
- Genomic-level data: the measures being used in real research.

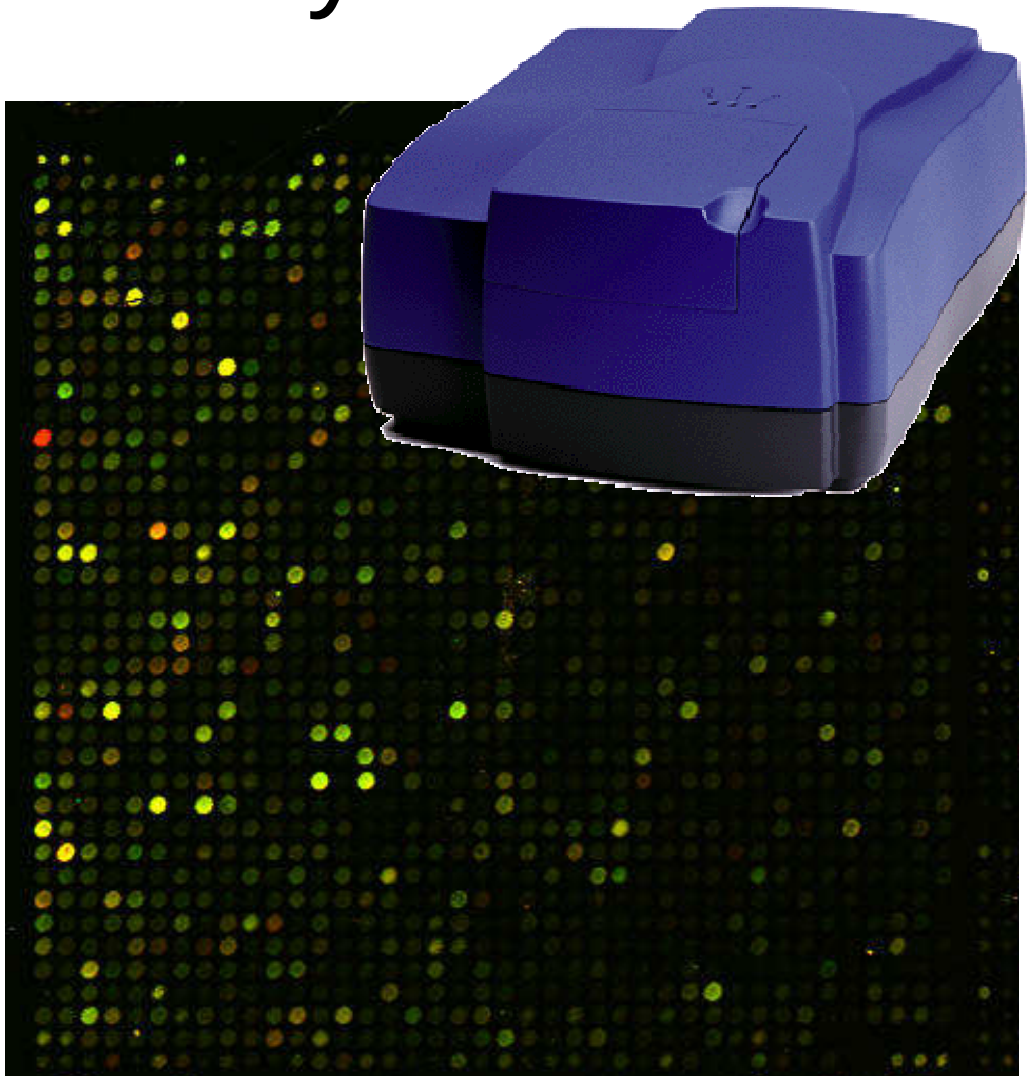
Preprocessing Data

- Preprocessing: A procedure that extracts meaningful data characteristics and eliminates unrelated system biases.
  - Image analysis
  - Data import
  - Normalization
  - Summarization
  - Quality assessment

Preprocessing Two-Color Spotted Arrays

Image Analysis

- The raw data from a microarray experiment consist of pairs of image files, 16-bit TIFFs, one for each of the dyes.
- Image analysis is required to extract measures of the red and green fluorescence intensities for each spot on the array.

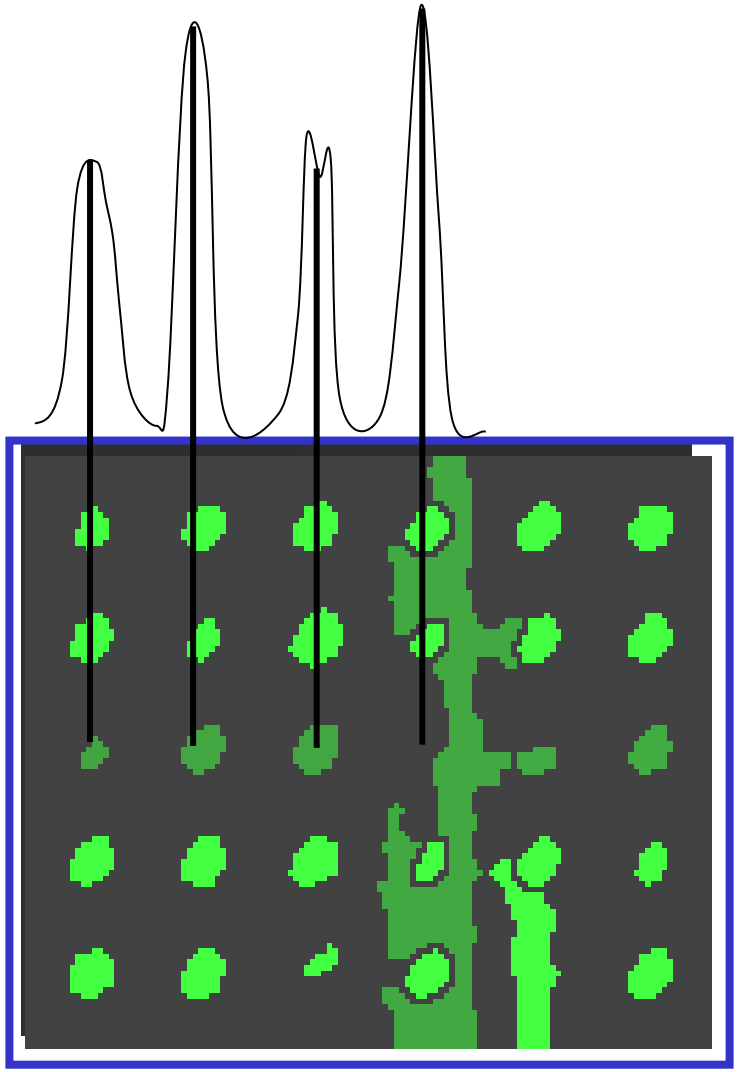


Steps in Image Analysis

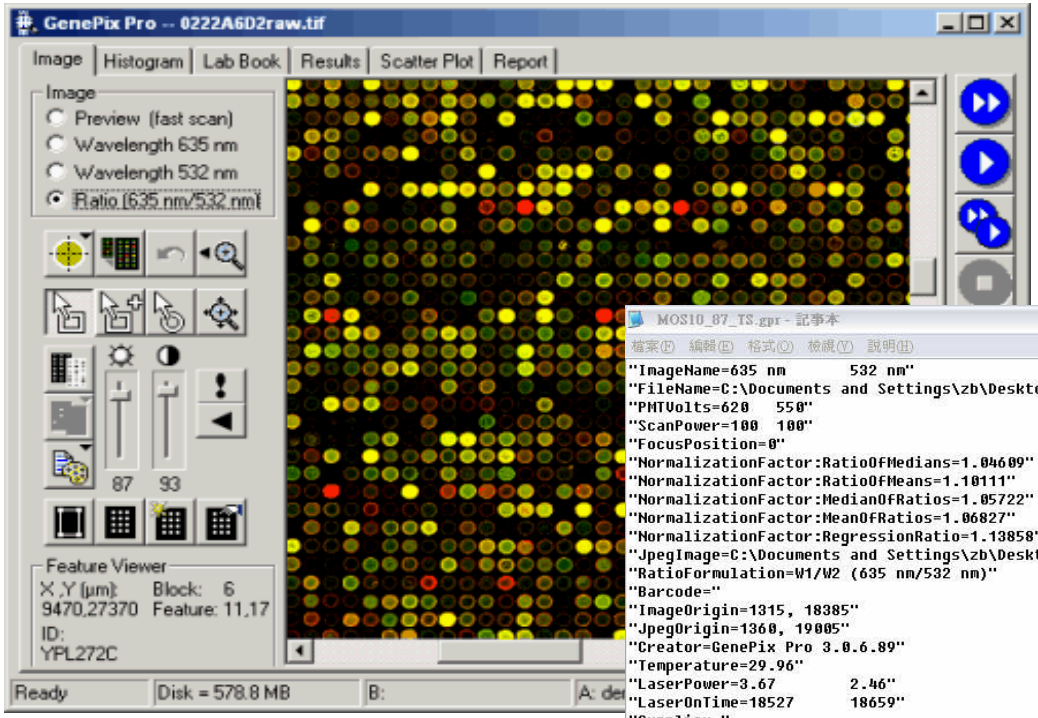
1. Addressing. Estimate location of spot centers.

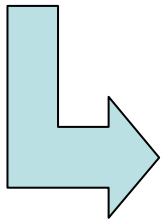
2. Segmentation. Classify pixels as foreground (signal) or background.

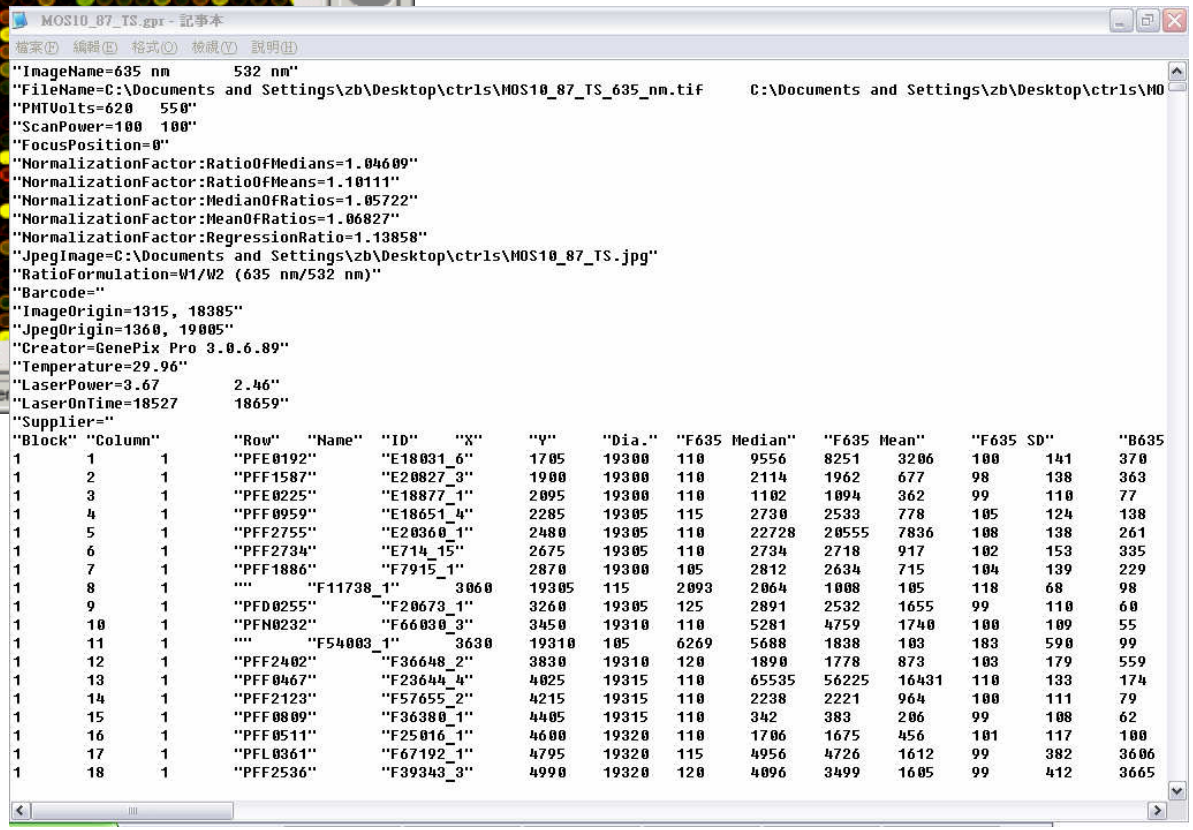
3. Information extraction. For each spot on the array and each dye
    - signal intensities;
    - background intensities;
    - quality measures.



- Bioconductor does NOT provide image analysis utilities and relies on other software.



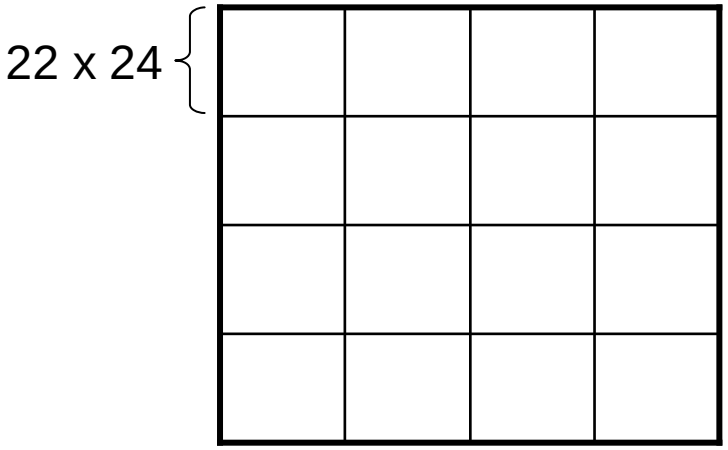




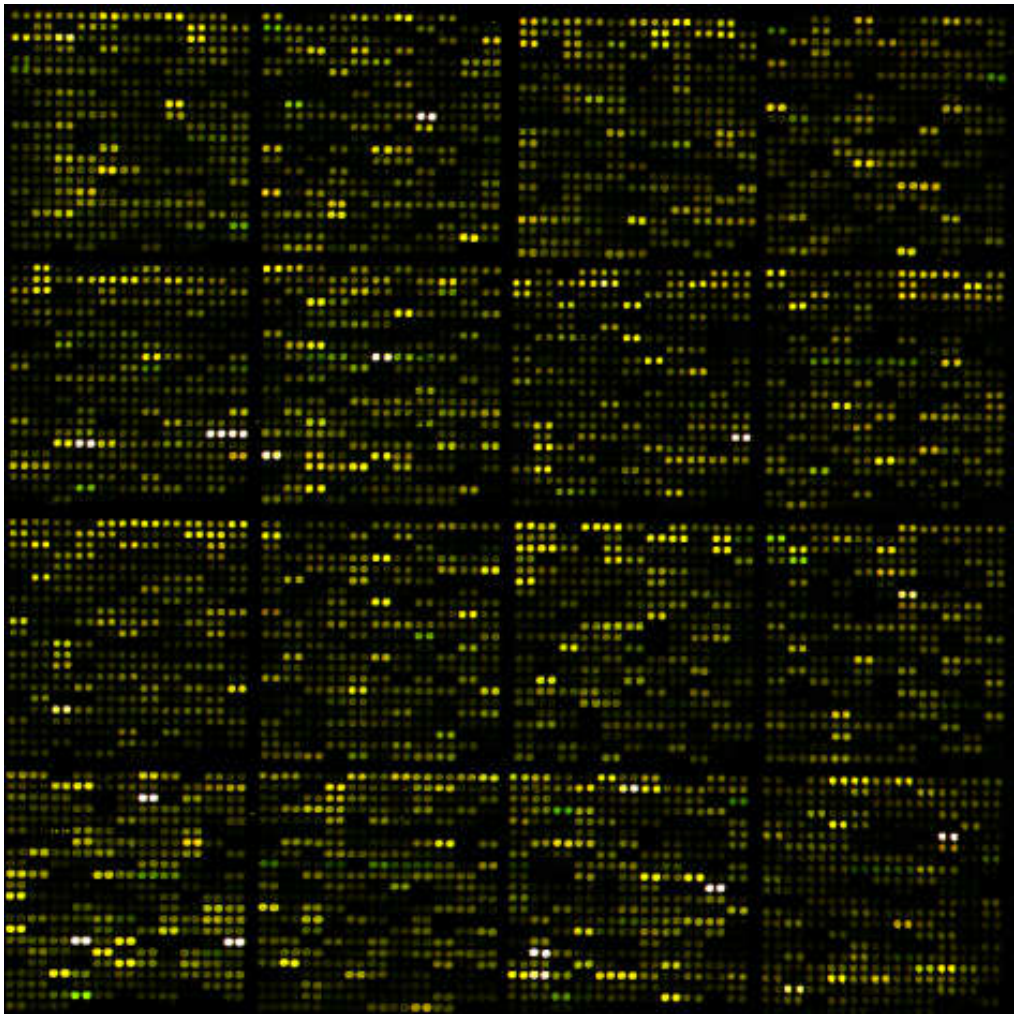
- The marray package helps to import the resulting data from different software:
  - GenePix: .gpr
  - Spot: .spot
  - SMD: .xls
  - Agilent: .txt

Data Import

- Illustrative data: *Swirl Zebrafish*
  - *Swirl* is a point mutant in the BMP2 gene of Zebrafish that affects the dorsal/ventral body axis.
  - Objective: to identify genes differentially expressed in the Swirl mutant compared to wild-type zebrafish
  - Microarray layout:
    - 8448 probes.
    - 4 x 4 print-tips; each grid consisted of a 22 x 24 spots matrix



22 x 24 x 4 x 4 = 8448



Data Import

- Information needed for effective statistical analysis:

  1) the sample target information

  2) the probe information

  3) the probe spot and background intensities

Data Import

1) the sample target information describes which RNA samples were hybridized to each array; typically a tab-delimited text file (.txt)

1	Names	slide number	experiment Cy3	experiment Cy5	date	comments
2	swirl.1.spot	81	swirl	wild type	2001/9/20	
3	swirl.2.spot	82	wild type	swirl	2001/9/20	
4	swirl.3.spot	93	swirl	wild type	2001/11/8	
5	swirl.4.spot	94	wild type	swirl	2001/11/8	

C:\Program Files\R\R-2.5.1\library\marray\swirldata\SwirlSample.txt

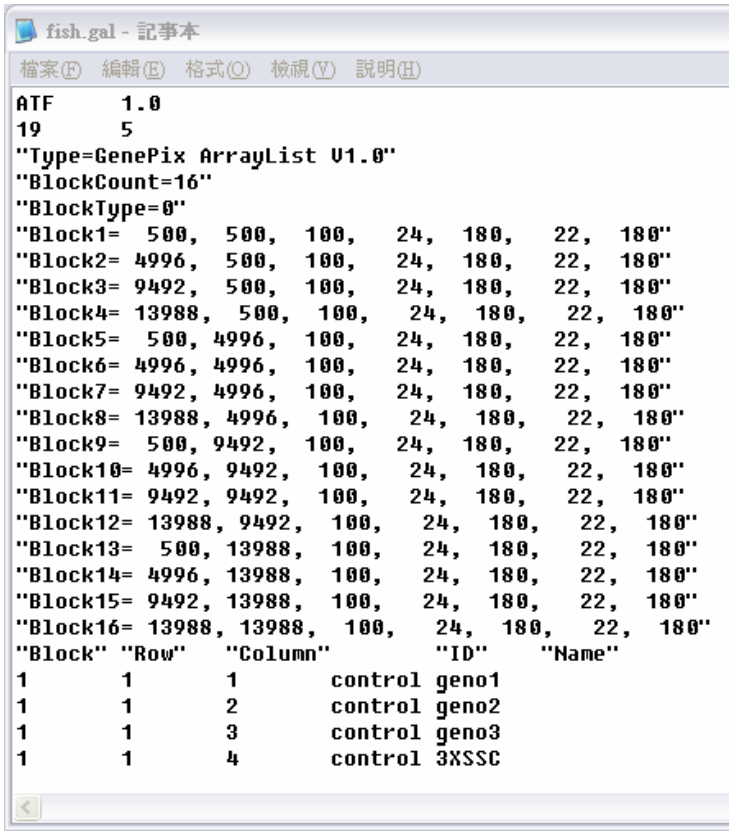
1	Names	slide number	experiment Cy3	experiment Cy5	date	comments
2	swirl.1.spot	81	swirl	wild type	2001/9/20	
3	swirl.2.spot	82	wild type	swirl	2001/9/20	
4	swirl.3.spot	93	swirl	wild type	2001/11/8	
5	swirl.4.spot	94	wild type	swirl	2001/11/8	

> read.marrayInfo(*fname*)

```
> swirl.samples = read.marrayInfo(file.path(datadir, "SwirlSample.txt"))
> swirl.samples
An object of class "marrayInfo"
@maLabels
[1] "swirl.1.spot" "swirl.2.spot" "swirl.3.spot" "swirl.4.spot"

@maInfo
         Names slide number experiment Cy3 experiment Cy5      date comments
1 swirl.1.spot           81          swirl      wild type 2001/9/20       NA
2 swirl.2.spot           82      wild type          swirl 2001/9/20       NA
3 swirl.3.spot           93          swirl      wild type 2001/11/8       NA
4 swirl.4.spot           94      wild type          swirl 2001/11/8       NA

@maNotes
[1] "C:/PROGRA~1/R/R-25~1.1/library/marray/swirldata/SwirlSample.txt"
```

2) the probe information describes the spotted probe sequences (gene names, annotations, etc.); if using GenePix or Spot, it is a tab-delimited text file (.gal) with rows corresponding to spotted probe sequences and columns containing various coordinates and annotations.

C:\Program Files\R\R-2.5.1
\library\marray\swirldata\fish.gal



> read.Galfile(*fname*)

3) the probe spot and background intensities:
microarray image processing results

> read.Spot(*fnames*) # Spot

> read.GenePix(*fnames*) # GenePix

> read.Agilent(*fnames*) # Agilent

The data is stored as *marrayRaw* class

```
> data = read.Spot(path=datadir, targets=swirl.targets)
Reading ...  C:/PROGRA~1/R/R-25~1.1/library/marray/swirldata/swirl.1.spot
Reading ...  C:/PROGRA~1/R/R-25~1.1/library/marray/swirldata/swirl.2.spot
Reading ...  C:/PROGRA~1/R/R-25~1.1/library/marray/swirldata/swirl.3.spot
Reading ...  C:/PROGRA~1/R/R-25~1.1/library/marray/swirldata/swirl.4.spot
> slotNames(data)
[1] "maRf"       "maGf"       "maRb"       "maGb"       "maW"        "maLayout"
[7] "maGnames"   "maTargets" "maNotes"
```

- For *marrayRaw* and *marrayNorm* objects, various methods are defined to extract stored information

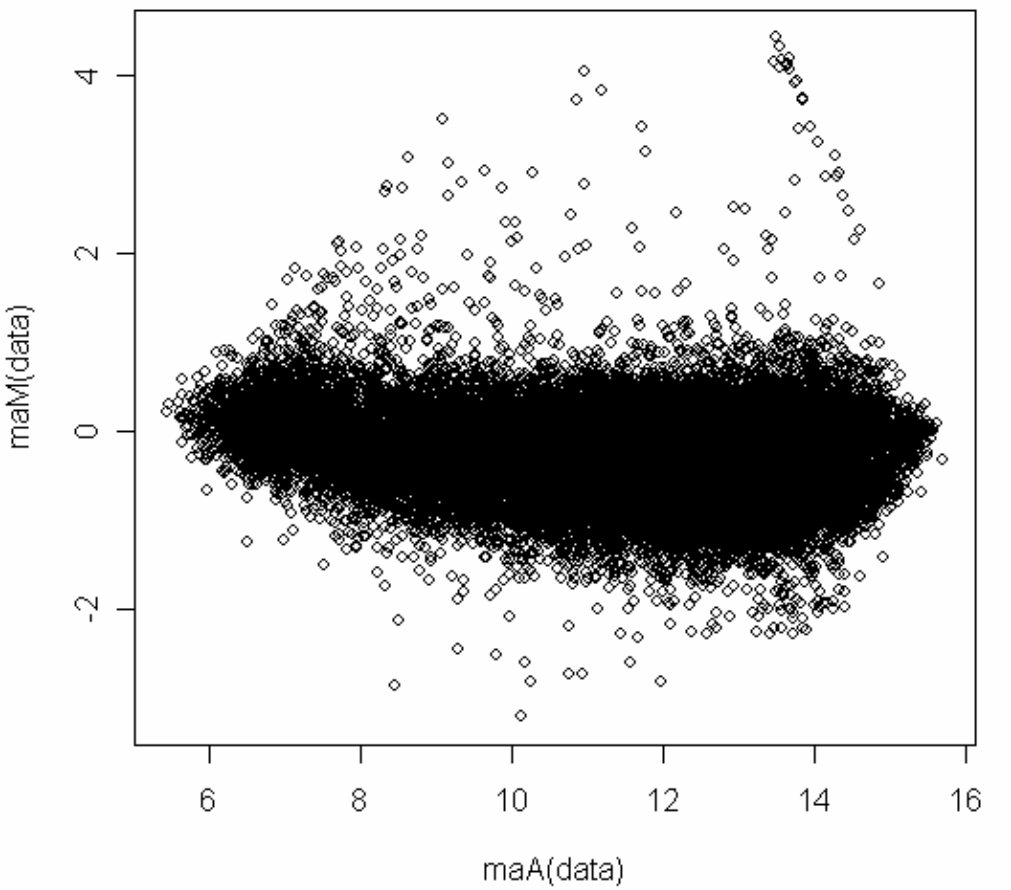
maM: $M = \log_2 \frac{Cy5}{Cy3} = \log_2(Cy5) - \log_2(Cy3)$

maA: $A = \log_2 \sqrt{Cy5 \cdot Cy3} = \frac{\log_2(Cy5) + \log_2(Cy3)}{2}$

maLG: the green log intensities

maLR: the red log intensities

maGeneTable: creates a *data.frame* of spot coordinates and gene names



> plot(maA(data),maM(data))

```
> maGeneTable(data)[1:4,1:5]
          Grid.R Grid.C Spot.R Spot.C       ID
control        1      1      1      1 control
control.1      1      1      1      2 control
control.2      1      1      1      3 control
control.3      1      1      1      4 control
```

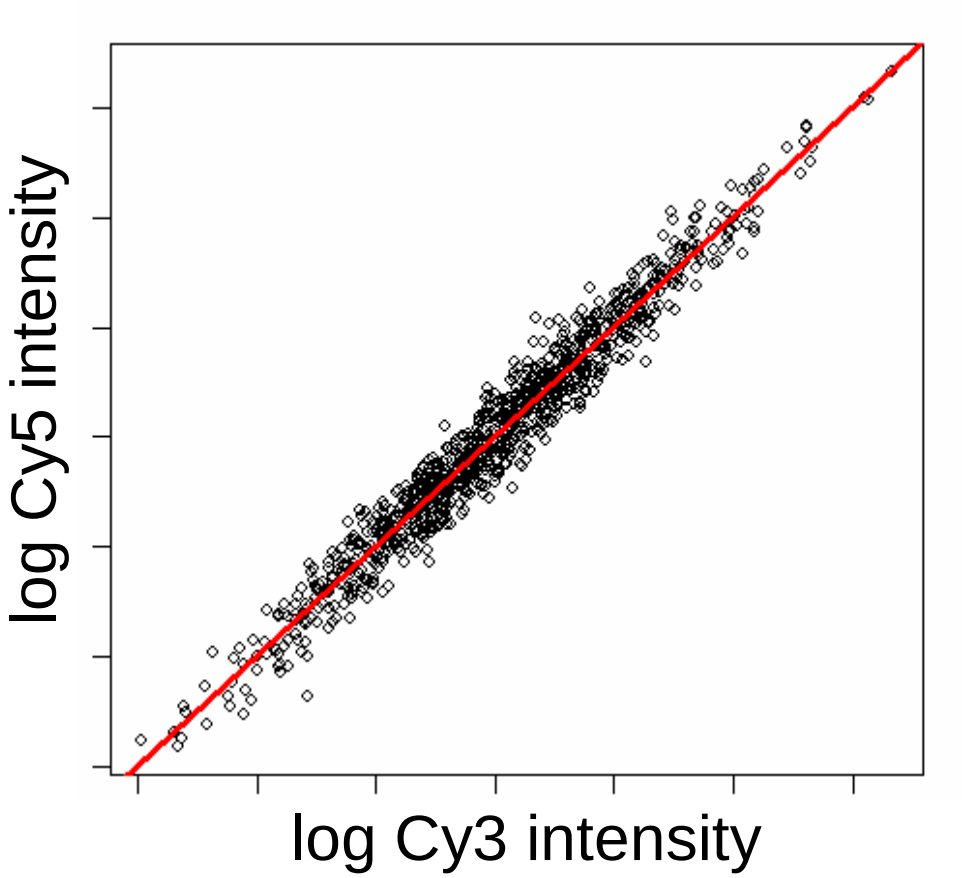
Normalization

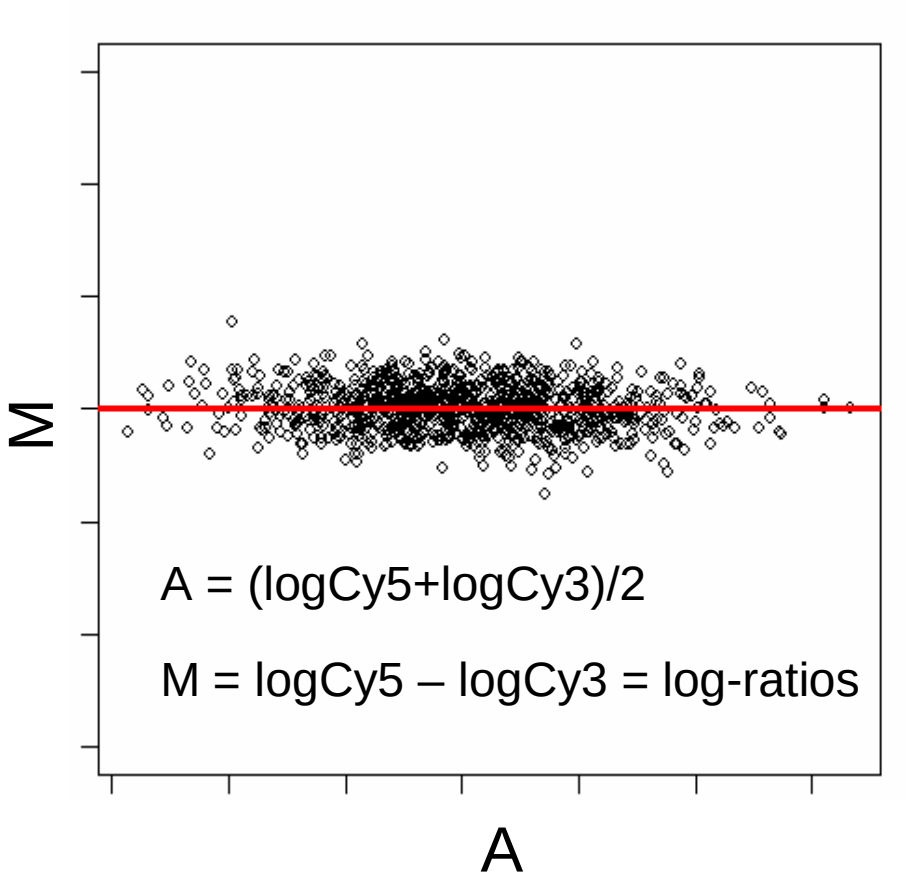
- Identify and remove systematic sources of variation in the measured fluorescence intensities, other than differential expression, for example
  - different labeling efficiencies of the dyes;
  - different amounts of Cy3- and Cy5- labeled mRNA;
  - different scanning parameters;
  - sector/print-tip, spatial, or plate effects, etc.
- Caution – Bias-variance trade-off: more complex normalization procedures tend to be able to remove more of the technical variation but they might also remove more of the biological signals.

Normalization

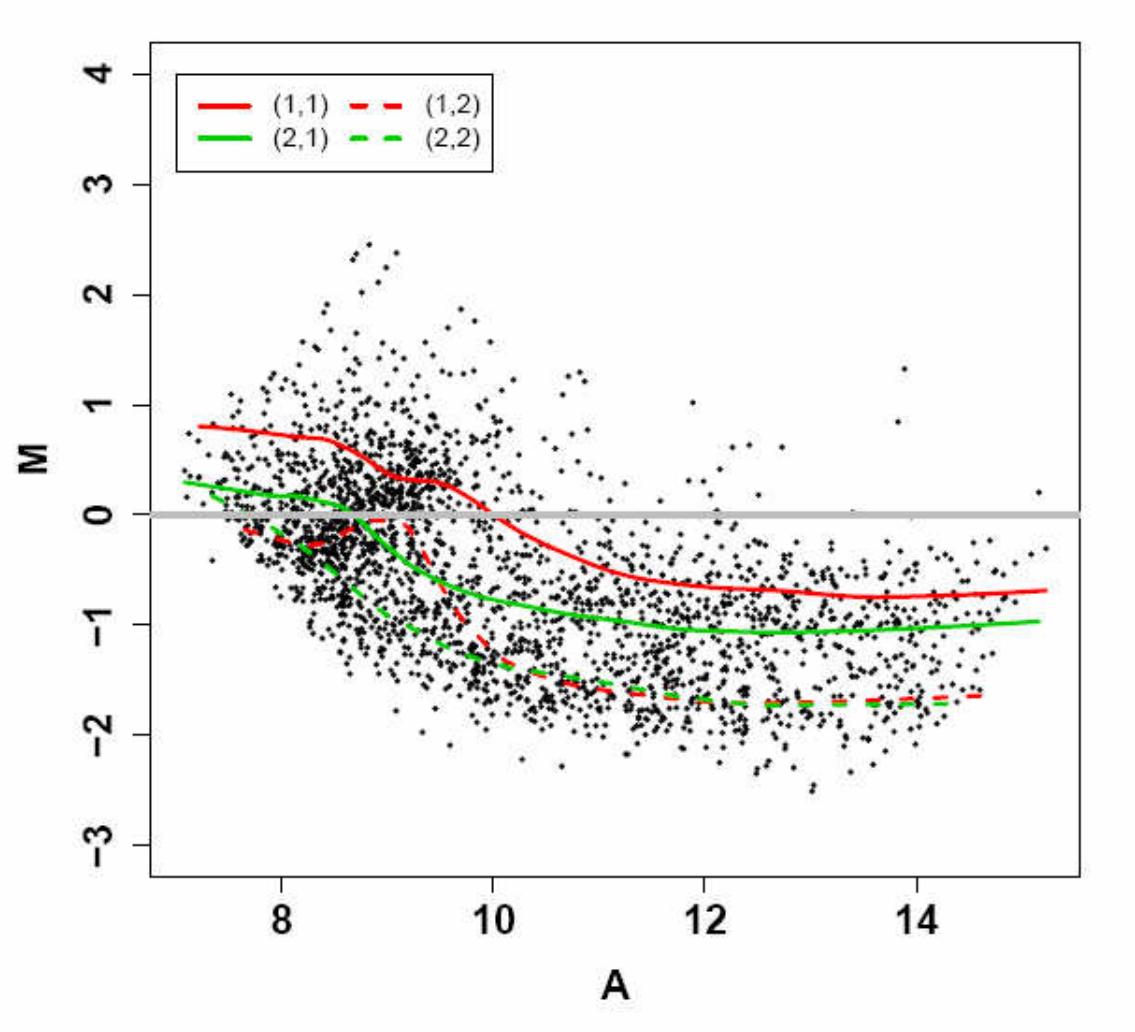
- The need for normalization can be seen most clearly in self-self hybridizations where the same mRNA sample is labeled with the Cy3 and Cy5 dyes.
- The imbalance in the red and green intensities is usually not constant across the spots within and between arrays, and can vary according to overall spot intensity, location, plate origin, etc.
- These factors should be considered in the normalization.

Self-Self Hybridization for Two-Channel Arrays (Ideal case)





Self-Self Hybridization for Two-Channel Arrays



Source: Yang and Throne (2003)

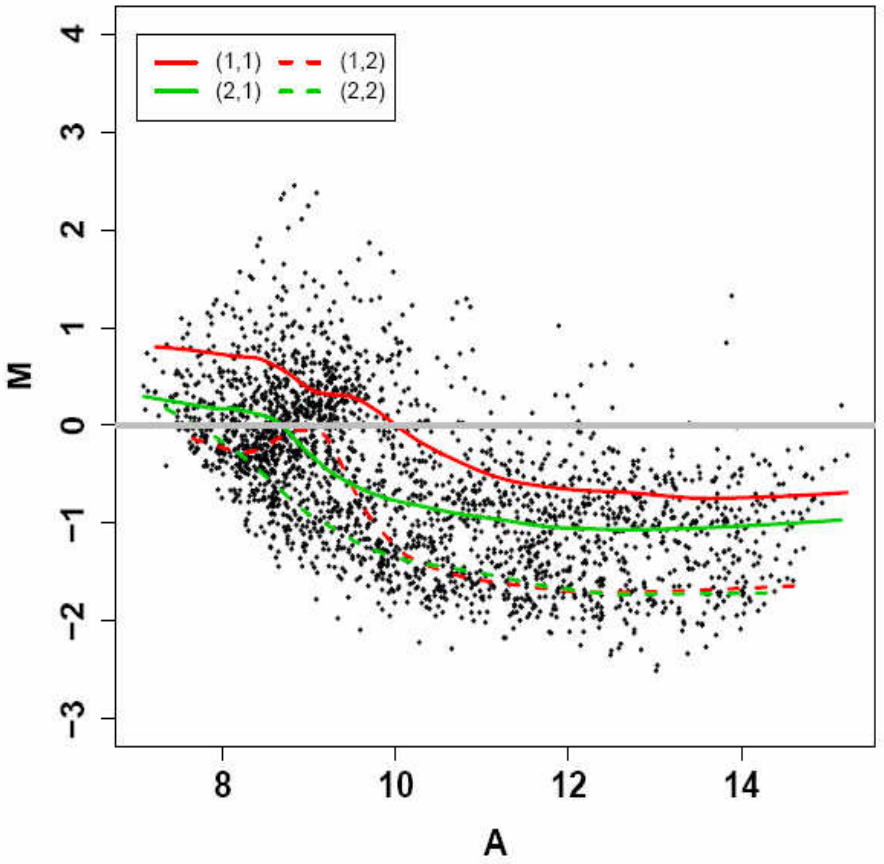
Within-Array Normalization

- Dealing with absolute log intensities; to eliminate spot-to-spot variation
  - Location normalization:

    $M_{norm} = M - l$
  - Scale normalization:

    $M_{norm} = M / s$
  - Location-scale normalization:

    $M_{norm} = (M - l) / s$



Location Normalization

- $M_{norm} = M - l$
  - It centers log-ratios around 0.
  - l can be identical for all values of M:
    - Global median normalization: l = median(M)
  - l can vary:
    - Global Lowess/Loess: l = the expected value after locally fitting a smooth regression curve on the global MA plot.
    - Print-tip Lowess/Loess: $l(i)$ = the loess/lowess fit to the *MA*-plot for the *i* th grid only.

LOWESS/LOESS

- The function fitted by LOWESS is a polynomial of the form:

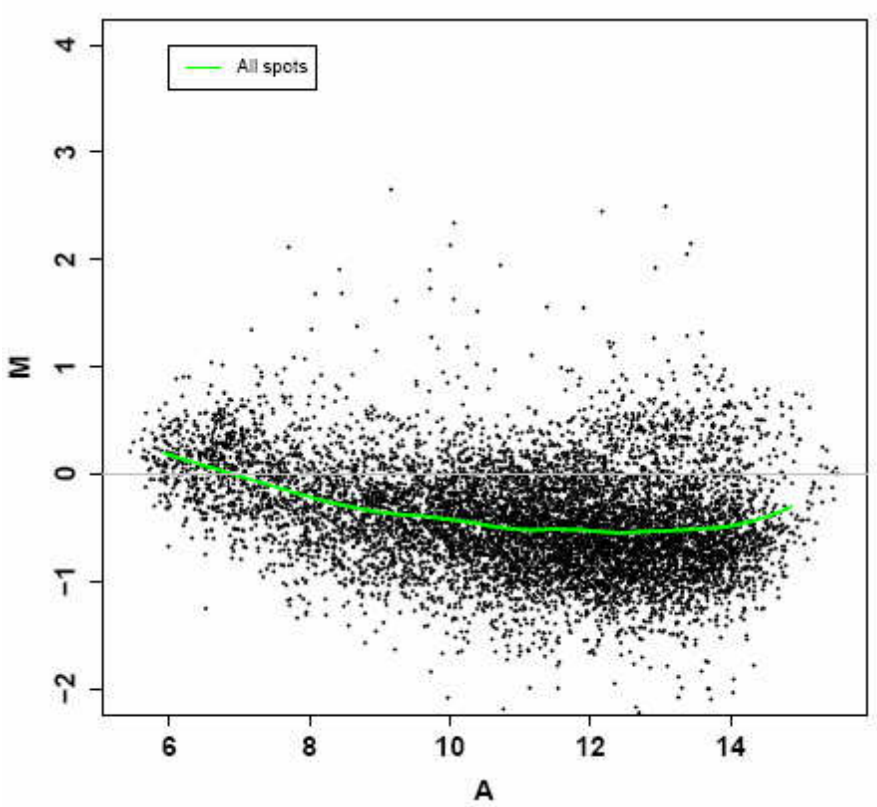
$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$$

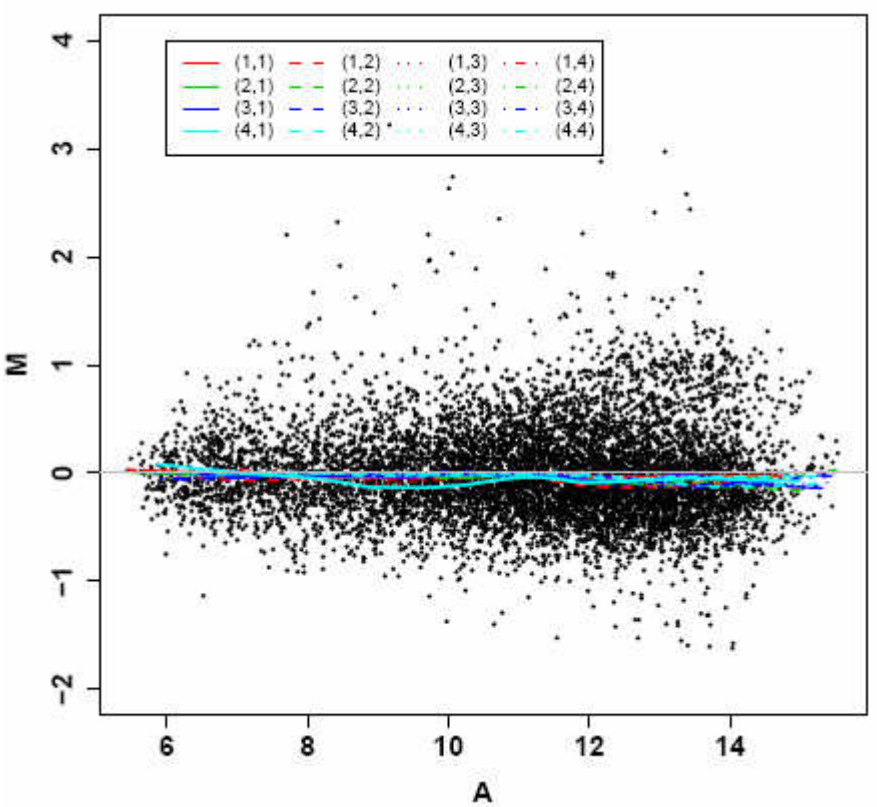
- The approach used by LOWESS/LOESS:

1. The degrees of the polynomials used are limited to 1 (LOWESS) or 2 (LOESS) avoid the over-fitting.
2. The data points that fall into this intervals will be used to fit the first polynomial in a weighted manner.

Loess Normalization Result

Global Loess





Scale Normalization

- Scale normalization can be applied if suspecting unequal variances.
- However, it is recommended to check the need of such normalization; there is a trade-off between the possible gain in bias achieved by scale normalization and the increase in variability introduced by this additional step.
- If the scale differences are relatively small, it is preferable to perform only a location normalization.

R: Within-Array Normalization

maNorm (limma): returns a *marrayNorm* object

Procedures	Description	Argument
None	No normalization.	n
Median	Global median location normalization.	m
Loess	Global *A*-dependent normalization using the scatter-plot smoother `loess`.	l
Print-tip *loess*	*A*-dependent normalization using the scatter-plot smoother `loess` within print-tip groups.	p
2D loess	2D–spatial normalization using the `loess` function.	twoD

maNormScale (limma)

Between-Array Normalization

- Dealing with log-ratios; to make array replicates comparable.
  - Quantile normalization method
  - VSN-method

Quantile Normalization

- **Assumption:** The measurements from different arrays share the same underlying distribution.
- It forces arrays have an identical distribution:

$$x_i' = F^{-1}(G_i(x_i))$$

G_i = empirical distribution functions for the *i* th array

F = the empirical distribution function of the means of the quantiles over all arrays

0.974[4]	0.341[3]	0.411[2]	-0.951[1]
1.857[5]	-0.036[2]	0.634[3]	-0.055[2]
-0.386[3]	-1.461[1]	0.885[4]	0.196[3]
-0.539[2]	0.368[4]	-0.530[1]	0.742[4]
-0.862[1]	2.634[5]	2.340[5]	2.277[5]

= (-0.862 – 1.461 – 0.951)/3

0.742[4]	0.196[3]	-0.055[2]
2.277[5]	-0.055[2]	0.196[3]
0.196[3]	-0.951[1]	0.742[4]
-0.055[2]	0.742[4]	-0.951[1]
-0.951[1]	2.277[5]	2.277[5]

Quantile Normalization

- Variants:
  - Gquantile: ensures that the green (first) channel has the same empirical distribution across arrays
  - Rquantile: ensures that the red (second) channel has the same empirical distribution across arrays
  - Aquantile: ensures that the A-values (average intensities) have the same empirical distribution across arrays
  - Tquantile: performs quantile normalization separately for the groups indicated by 'targets'

Variance Stabilization and Normalization (VSN) Method

- VSN assumes that less than half of the genes on the arrays is differentially transcribed across the experiment.
- The variance of microarray data may depend on the signal intensity; after normalization the variance is approximately constant
- x_{ki} = expression level of the k th probe in the i th array.

$$x'_{ki} = \mathrm{g}\log\left(\frac{x_{ki} - a_i}{b_i}\right)$$

a_i = background offset

b_i = scale parameter for array i

$\mathrm{glog}(.)$ = genealized logarithm,

R: Within-Array Normalization

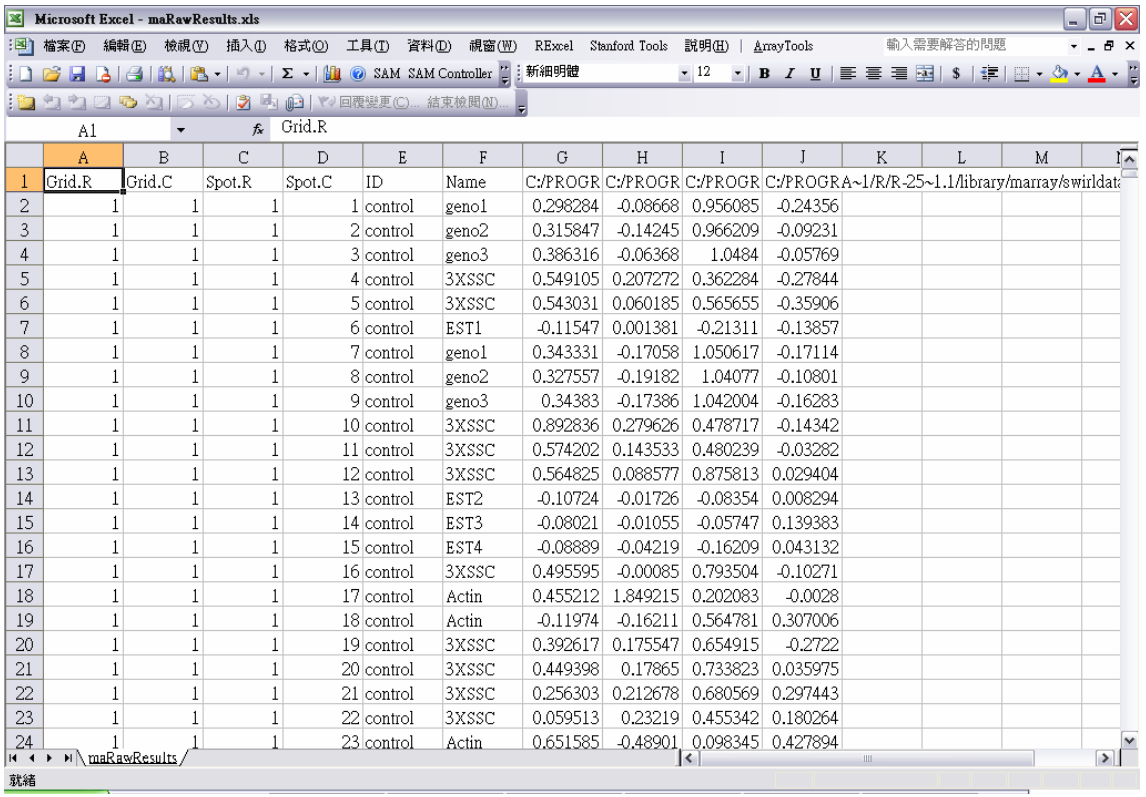
normalizeBetweenArrays (limma)

```
Usage:

     normalizeBetweenArrays(object, method="Aquantile", targets=NULL, ...)

Arguments:

  object: a 'matrix', 'RGList' or 'MAList' object containing expression
          ratios for a series of arrays

  method: character string specifying the normalization method to be
          used. Choices are '"none"', '"scale"', '"quantile"',
          '"Aquantile"', '"Gquantile"', '"Rquantile"', '"Tquantile"' or
          '"vsn"'. A partial string sufficient to uniquely identify the
          choice is permitted.
```

Data Output

- Export the normalized data into a file:

> write.marray(withinNormData)



Microsoft Excel - maRawResults.xls

A1: Grid.R

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Grid.R	Grid.C	Spot.R	Spot.C	ID	Name	C:/PROGR	C:/PROGR	C:/PROGR	C:/PROGRA~1/R/R-25~1.1/library/marray/swirldat			
2	1	1	1	1	control	geno1	0.298284	-0.08668	0.956085	-0.24356			
3	1	1	1	2	control	geno2	0.315847	-0.14245	0.966209	-0.09231			
4	1	1	1	3	control	geno3	0.386316	-0.06368	1.0484	-0.05769			
5	1	1	1	4	control	3XSSC	0.549105	0.207272	0.362284	-0.27844			
6	1	1	1	5	control	3XSSC	0.543031	0.060185	0.565655	-0.35906			
7	1	1	1	6	control	EST1	-0.11547	0.001381	-0.21311	-0.13857			
8	1	1	1	7	control	geno1	0.343331	-0.17058	1.050617	-0.17114			
9	1	1	1	8	control	geno2	0.327557	-0.19182	1.04077	-0.10801			
10	1	1	1	9	control	geno3	0.34383	-0.17386	1.042004	-0.16283			
11	1	1	1	10	control	3XSSC	0.892836	0.279626	0.478717	-0.14342			
12	1	1	1	11	control	3XSSC	0.574202	0.143533	0.480239	-0.03282			
13	1	1	1	12	control	3XSSC	0.564825	0.088577	0.875813	0.029404			
14	1	1	1	13	control	EST2	-0.10724	-0.01726	-0.08354	0.008294			
15	1	1	1	14	control	EST3	-0.08021	-0.01055	-0.05747	0.139383			
16	1	1	1	15	control	EST4	-0.08889	-0.04219	-0.16209	0.043132			
17	1	1	1	16	control	3XSSC	0.495595	-0.00085	0.793504	-0.10271			
18	1	1	1	17	control	Actin	0.455212	1.849215	0.202083	-0.0028			
19	1	1	1	18	control	Actin	-0.11974	-0.16211	0.564781	0.307006			
20	1	1	1	19	control	3XSSC	0.392617	0.175547	0.654915	-0.2722			
21	1	1	1	20	control	3XSSC	0.449398	0.17865	0.733823	0.035975			
22	1	1	1	21	control	3XSSC	0.256303	0.212678	0.680569	0.297443			
23	1	1	1	22	control	3XSSC	0.059513	0.23219	0.455342	0.180264			
24	1	1	1	23	control	Actin	0.651585	-0.48901	0.098345	0.427894			

maRawResults

就緒

Quality Assessment

- Before and after normalization, it is important to consider and unsure the quality of the data. $\Rightarrow$ Visualization tools.
- The package *arrayQuality* gives user a quick visual way to access the quality of individual arrays by providing per-slide diagnostic plots.

> maQualityPlots(data, dev=“jpeg”)

dev: A "character" string naming the graphics device. This will take arguments "png", "jpeg" and "ps" only. By default, dev is set to "png".

diagPlot.swirl.1.png : Qualitative Diagnostic Plots

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