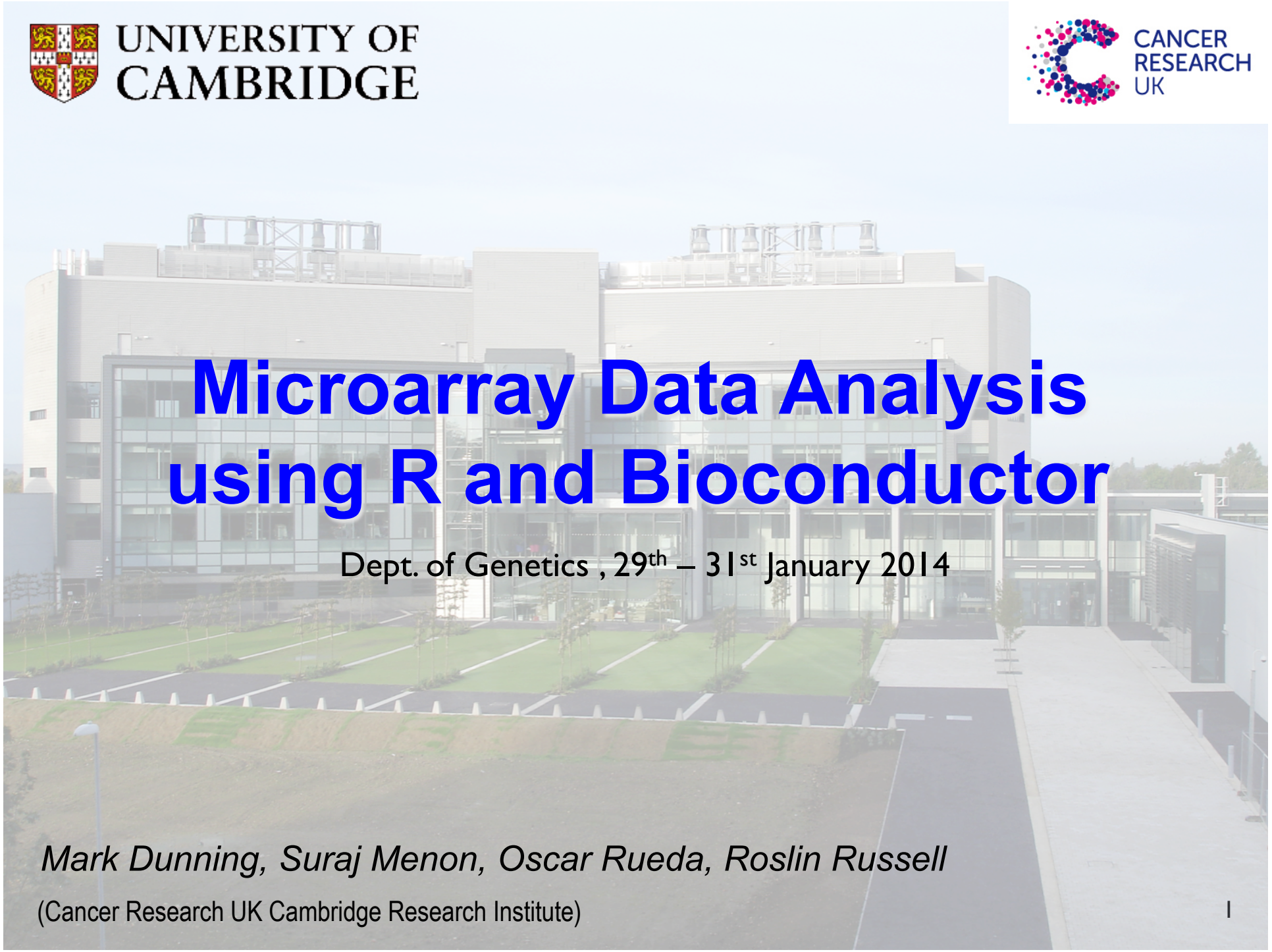


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Microarray Data Analysis using R and Bioconductor

Dept. of Genetics , 29th – 31st January 2014

Mark Dunning, Suraj Menon, Oscar Rueda, Roslin Russell

(Cancer Research UK Cambridge Research Institute)

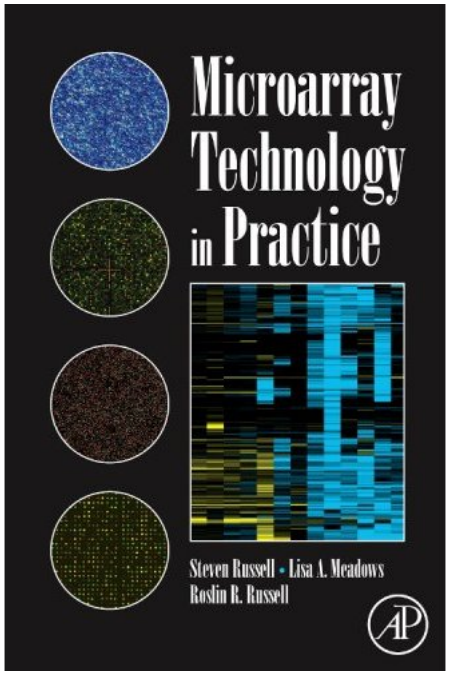
Outline

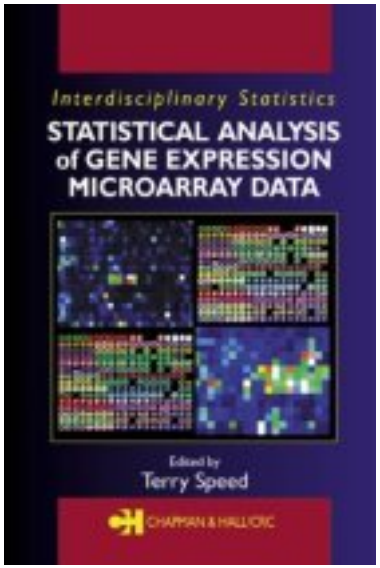
- *R* and *Bioconductor* for microarray data analysis [lecture/practical]
- Exploratory data analysis, quality assessment [lecture/practical]
- Experimental design [lecture]
- Linear models, design matrices [lecture/practical]
- Statistics for differential expression [lecture/practical]
- Processing Illumina BeadChips [lecture/practical]
- Processing Affymetrix Arrays [lecture/practical]
- Processing Copy Number Arrays [lecture/practical]
- Pathway Enrichment Analysis [lecture/practical]

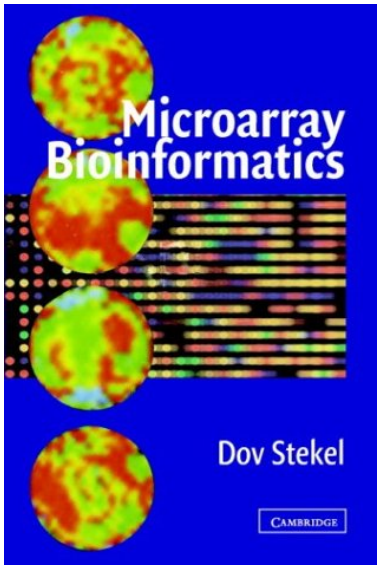
`http://www.compbio.group.cam.ac.uk/teaching.html`

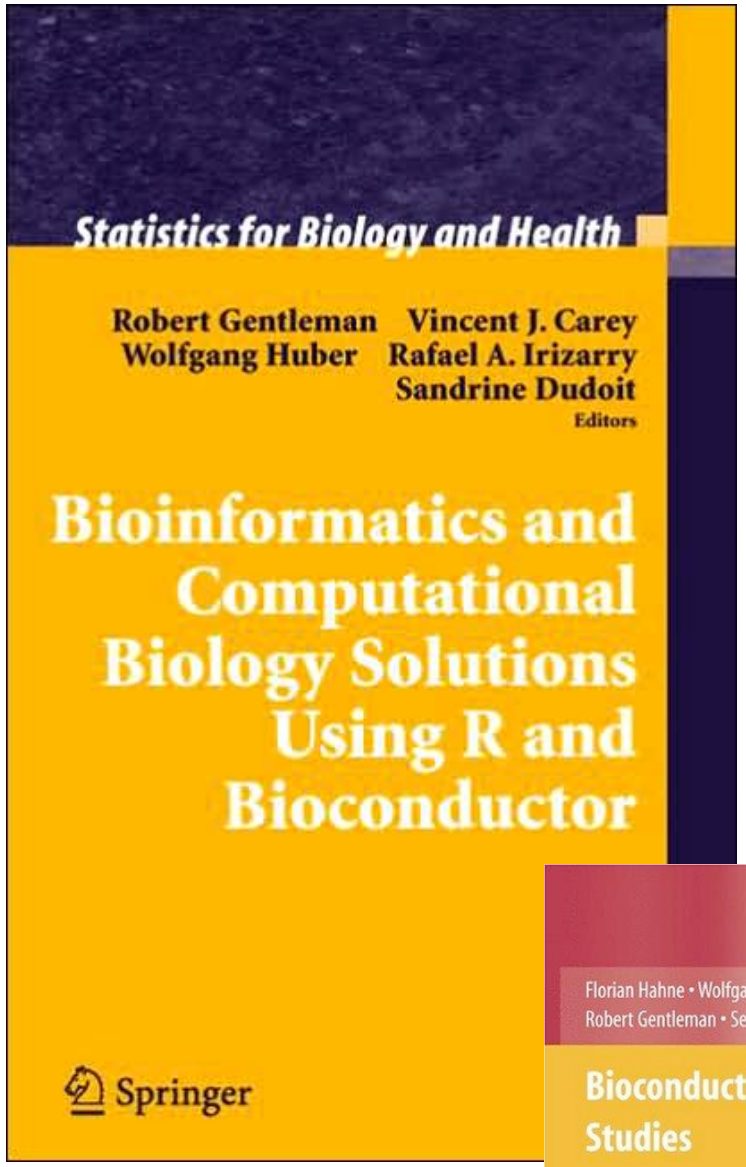
Mark Dunning, Suraj Menon, Oscar Rueda, Roslin Russell

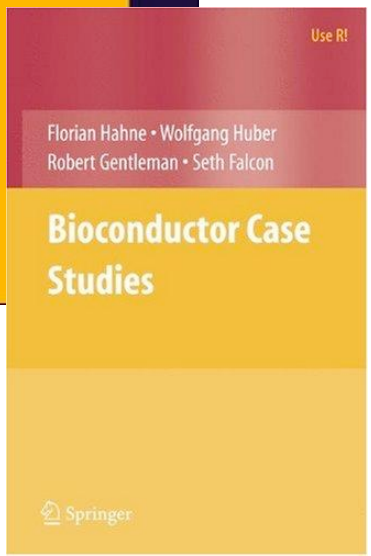
Textbooks


Microarray Technology in Practice
Steven Russell • Lisa A. Meadows
Roslin R. Russell
AP


Interdisciplinary Statistics
STATISTICAL ANALYSIS of GENE EXPRESSION MICROARRAY DATA
Edited by
Terry Speed
CHAPMAN & HALL/CRC


Microarray Bioinformatics
Dov Stekel
CAMBRIDGE


Statistics for Biology and Health
Robert Gentleman
Vincent J. Carey
Wolfgang Huber
Rafael A. Irizarry
Sandrine Dudoit
Editors
Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Springer


Use R!
Florian Hahne • Wolfgang Huber
Robert Gentleman • Seth Falcon
Bioconductor Case Studies
Springer

Issues in the analysis of two-colour microarray data

Natalie Thorne

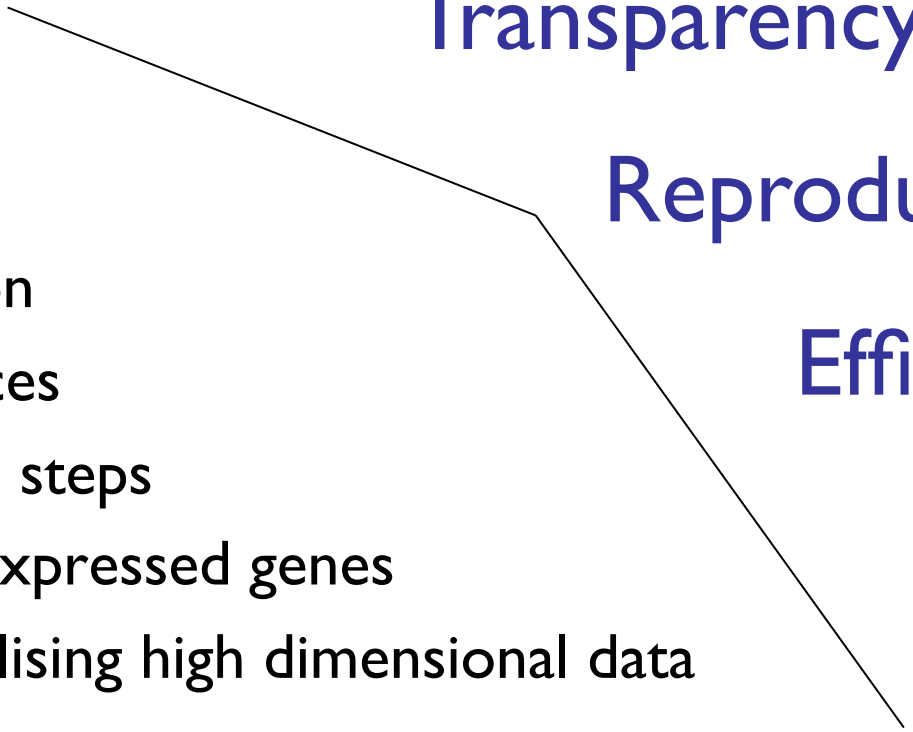
Many thanks to Terry Speed, Gordon Smyth, Jean (Yee Hwa) Yang, Ingrid Lonnstedt, Matthew Ritchie for sharing their teaching material with me.

Software needs (I)

- Software for computational biology and bioinformatics, especially the analysis of microarray data had to meet the needs of both:
  - **biologists**,
  - **Statisticians**
  - **computer scientists**.
- The analysis of microarray data was not simply the domain of statisticians, an entry point for biologist was required.
- New conceptual needs and challenges, requirement for biologists to access high quality statistical methods

Software needs (II)

Data acquisition
Data management
Data transformation
Background correction
Combining data sources
Various normalisation steps
Finding differentially expressed genes
Interpreting and visualising high dimensional data



Transparency

Reproducibility

Efficiency

R

- Robert Gentleman & Ross Ihaka, 1996
- Free software environment for statistical computing and graphics
- High level interpreted language
- Packaging protocols
- Interfaces with Perl, Python, Java, C, XML
- Active user community

Get the latest R version
(precompiled binaries for Windows and Mac OS X)
http://cran.r-project.org/

Literature on 

Books

- R programming for Bioinformatics - Gentleman (2005)
- An Introduction to R - Venables WN et al (2005)
- R Graphics - Murrell (2009)

Online Guides

- Kickstarting R:

  http://cran.r-project.org/doc/contrib/Lemon-kickstart/index.html

- R & Bioconductor Manual:

  http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual

CambR Meeting

Cambridge R user group
Organised by Mark Dunning & Laurent Gatto
https://groups.google.com/forum/?fromgroups=#!forum/cambridge-r-user-group

Where: The Fountain Inn [2], 12 Regent Street, Cambridge, CB2 1DB

Oct 2012
Talk given by **Martin Morgan**, who leads the famous Bioconductor project.
http://blog.revolutionanalytics.com/2011/05/the-r-files-martin-morgan.html

Sweave

What is it?

Sweave is a tool that allows to embed the R code for complete data analyses in latex documents.

Create dynamic reports, which can be updated automatically if data or analysis change.

Where can I get it?

The Sweave software itself is part of every R installation, see

```
help("Sweave", package="utils")
```

to get started.

http://www.stat.uni-muenchen.de/~leisch/Sweave/

Sweave

Why Sweave?

Reproducibility

Make your research more reproducible.

Efficiency

Statistical output is automatically incorporated into your report.

Reliability

The integration of analyses with the report reduces the chance of errors entering in through copying and pasting of statistical output into documents.

Education & Communication

By providing data analysis code for a report, this teaches others how to do similar analyses.

http://www.r-bloggers.com/getting-started-with-sweave-r-latex-eclipse-statet-texlipse/

Bioconductor

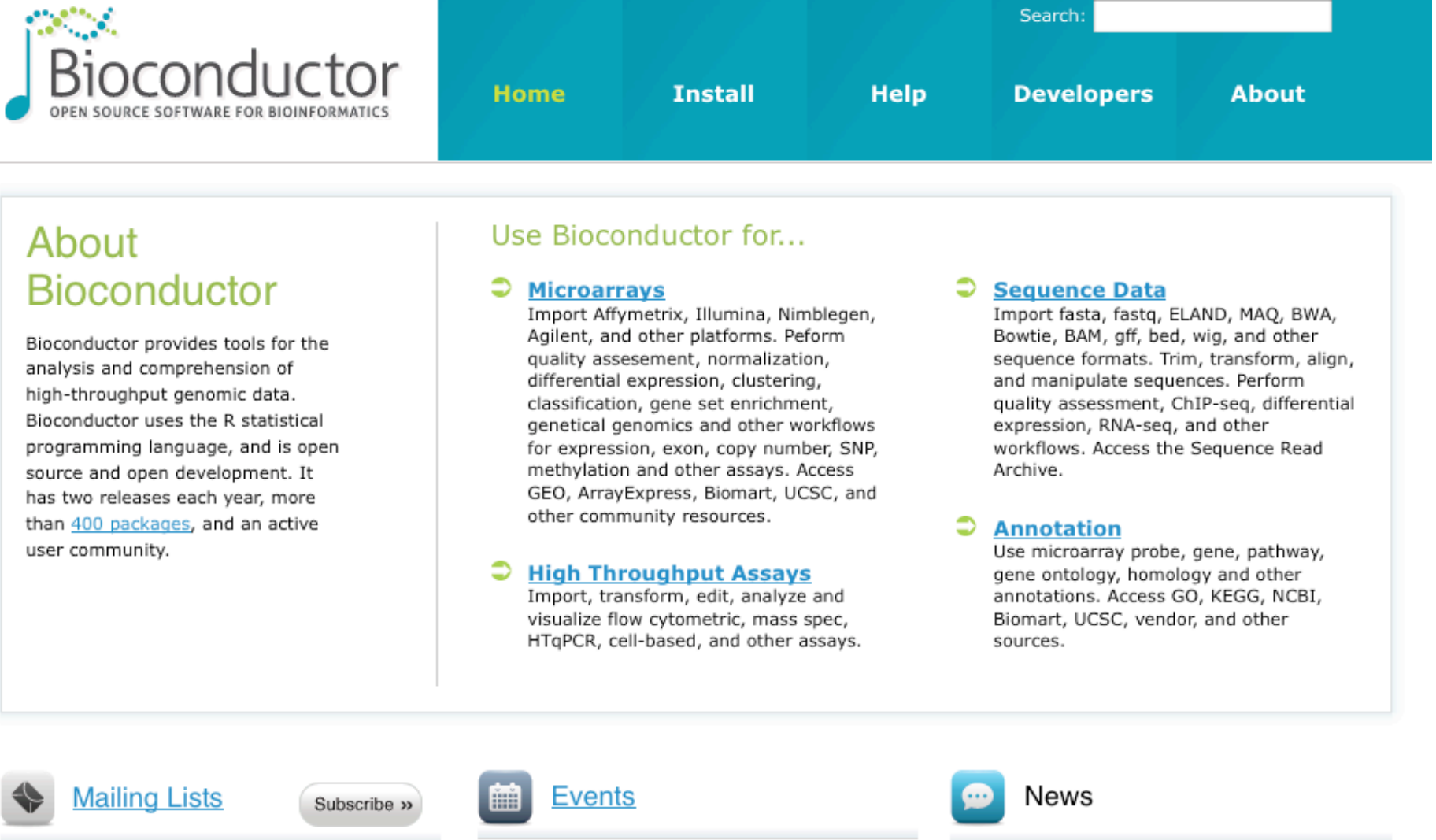
- Open source software for bioinformatics
- Bioconductor is a series of R packages
- Core emphasis on reproducible research, good documentation and training, re-usable data structures, designed to work with different variations of data
- Questions about the analysis of array data using Bioconductor can be posted on their mailing list. This is a very informative mailing list for the analysis of data from a wide variety of high throughput genomic technologies.

  http://www.bioconductor.org/docs/mailList.html

  http://www.bioconductor.org/

For a quick install of a subset of the most common packages:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite()
```


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
About Bioconductor
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, more than 400 packages, and an active user community.
Use Bioconductor for...
Microarrays
Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Peform quality assesement, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.
High Throughput Assays
Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.
Sequence Data
Import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.
Annotation
Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.
Mailing Lists
Subscribe »
Events
News

Limma

- Especially for the application of **linear models** for analysing designed experiments and the assessment of differential expression
- Includes processing capabilities for two-colour spotted arrays and affymetrix chip data. The differential expression methods treat two-colour, affymetrix and single-colour experiments in a **unified** way.

http://bioinf.wehi.edu.au/limma (home page)

http://www.bioconductor.org/packages/2.5/bioc/html/limma.html

(link to userguide)

Smyth, G. K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397- 420

References

Good collection of references

http://www.statsci.org/micrarra/refs/index.html

This is a good overview (old, but good!)

Smyth, G. K., Yang, Y.-H., Speed, T. P. (2003). Statistical issues in microarray data analysis. In: *Functional Genomics: Methods and Protocols*, M. J. Brownstein and A. B. Khodursky (eds.), Methods in Molecular Biology Volume 224, Humana Press, Totowa, NJ, pages 111-136. (PDF)

Bioconductor Software packages list

http://www.bioconductor.org/packages/2.5/Software.html

links to packages according to whether they are for: Microarray, Annotation, Visualisation, Statistics, GraphsAndNetworks, Technology, Infrastructure

There are also Annotation packages and ExperimentData packages

http://www.bioconductor.org/packages/2.5/BiocViews.html

Other key packages

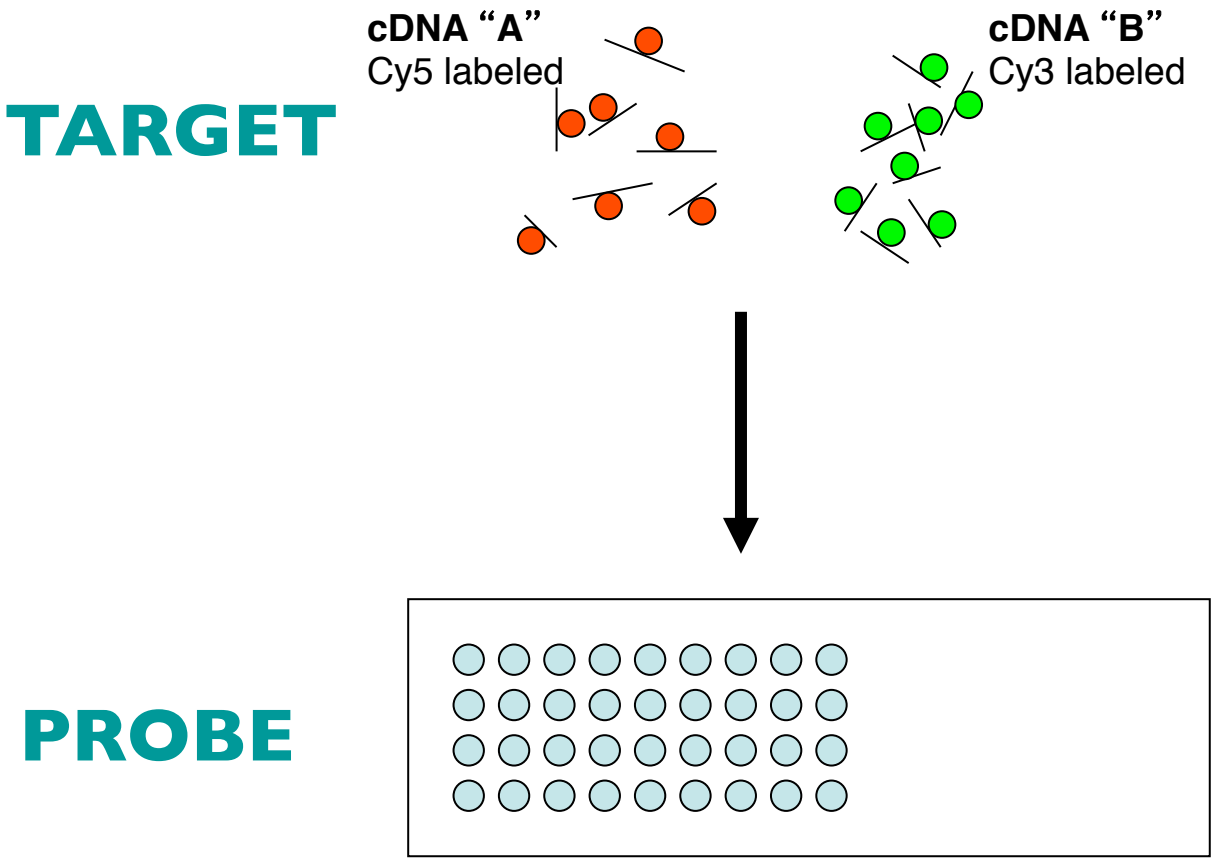
- Microarray data
  - limma, marray, aroma.light….
- Affymetrix data
  - affy, oligo, crlmm, simpleaffy, affyPLM, gcrma, affyio….
- aCGH data
  - snapCGH, aCGH, DNAcopy, GLAD….
- High density arrays (Nimblegen)
  - tilingArray, oligo,…
- Illumina beadarray data
  - beadarray, crlmm, lumi, beadarraySNP, BeadExplorer…

snapCGH

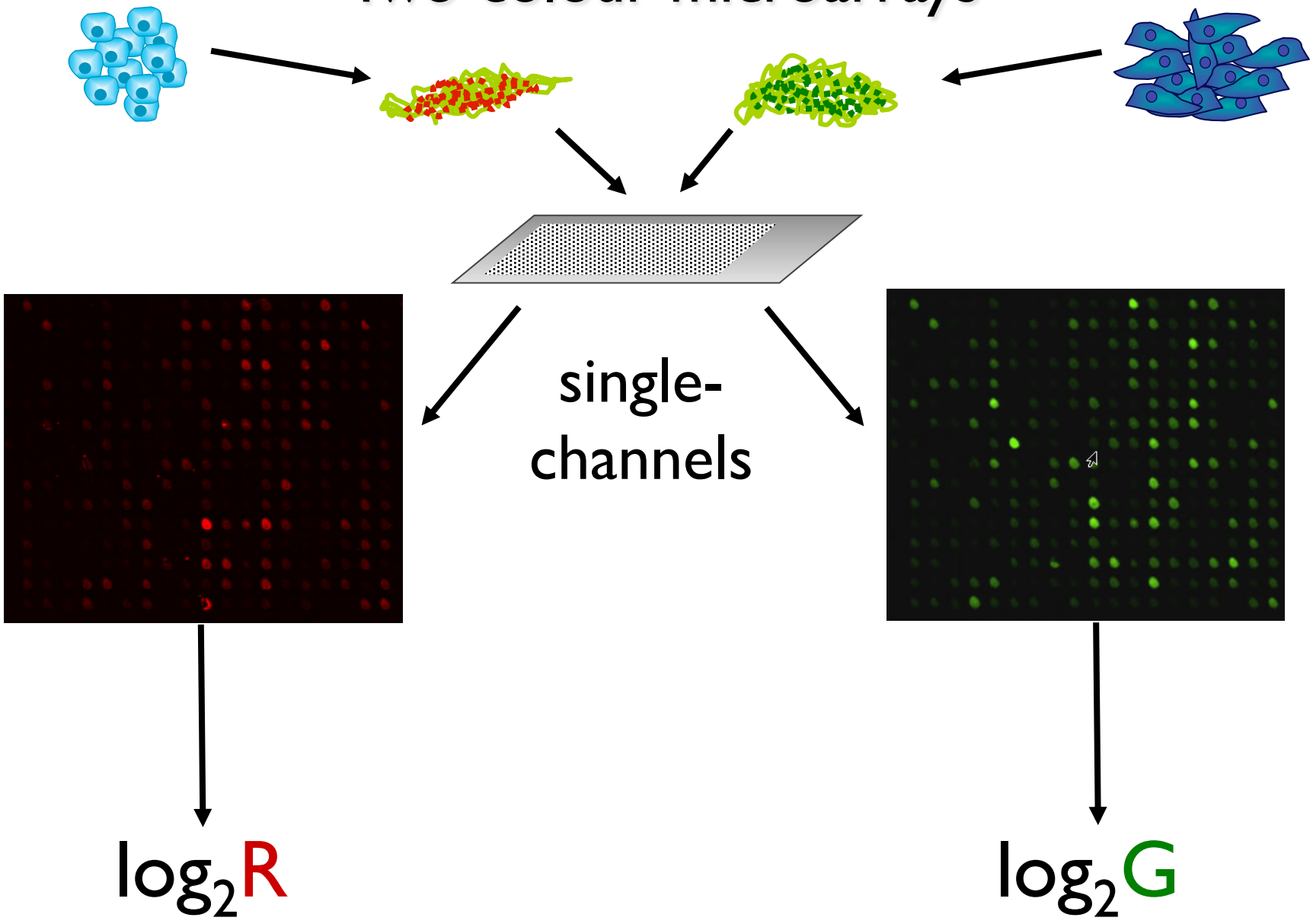
- Designed to be compatible with the popular R package *limma*
- This allows smooth progression between pre-processing, normalization and the segmentation steps
- Potentially reduces learning curve for anyone already familiar with *limma* e.g. biologists

Exploratory data analysis and normalisation of one- and two-colour microarray data

Definition of probe and target



Two-colour microarrays



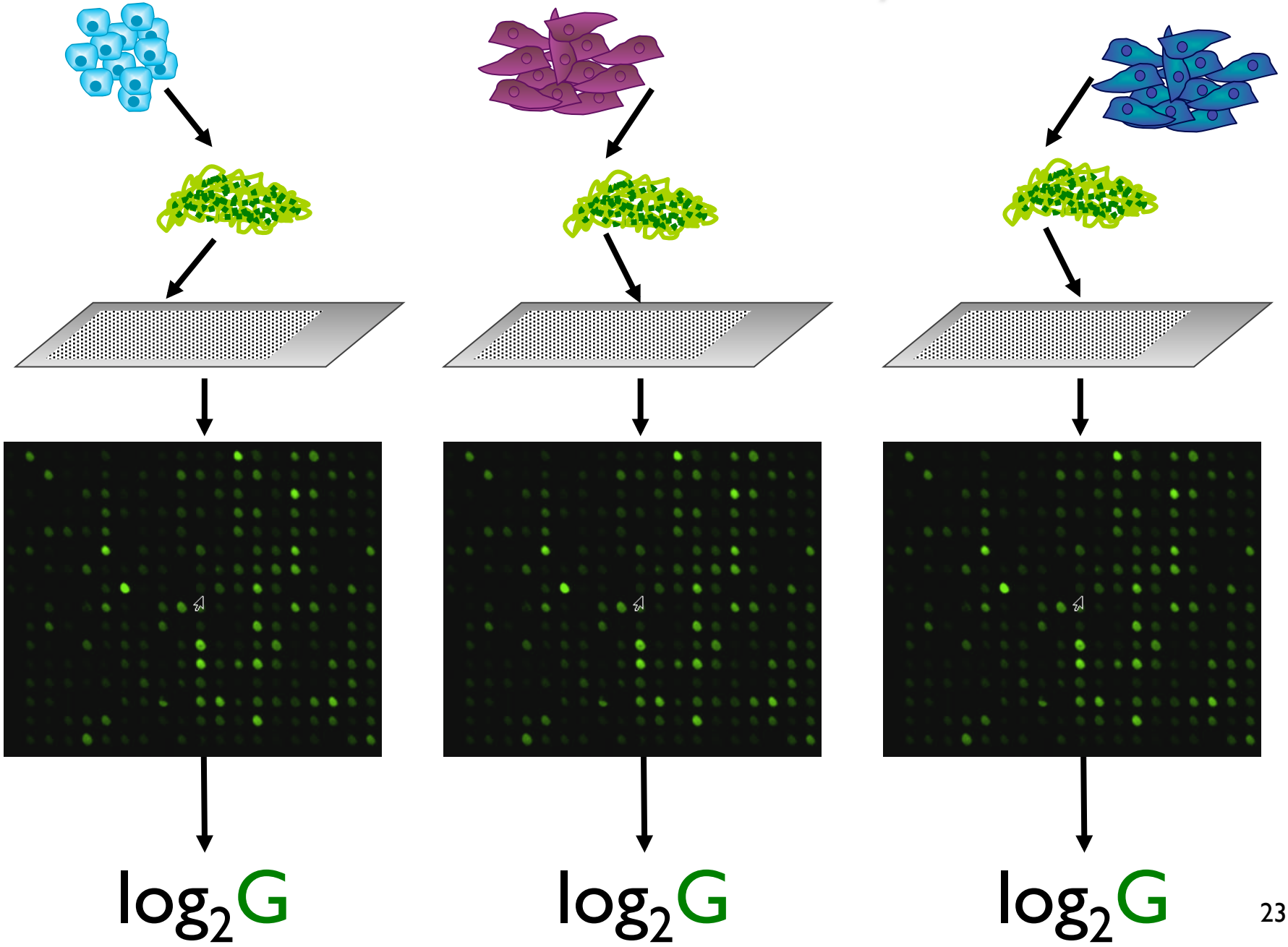
Two-colour microarray statistics

$\log_2 R$ $\log_2 G$

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

$$A = {}^1/_2 (\log_2 R + \log_2 G)$$

One-colour microarrays



Overview of exploratory analysis

Tools for exploratory analysis

Exploratory analysis in Limma

Normalisation in Limma

Further exploratory analysis

Quality assessment

Tools for exploratory analysis

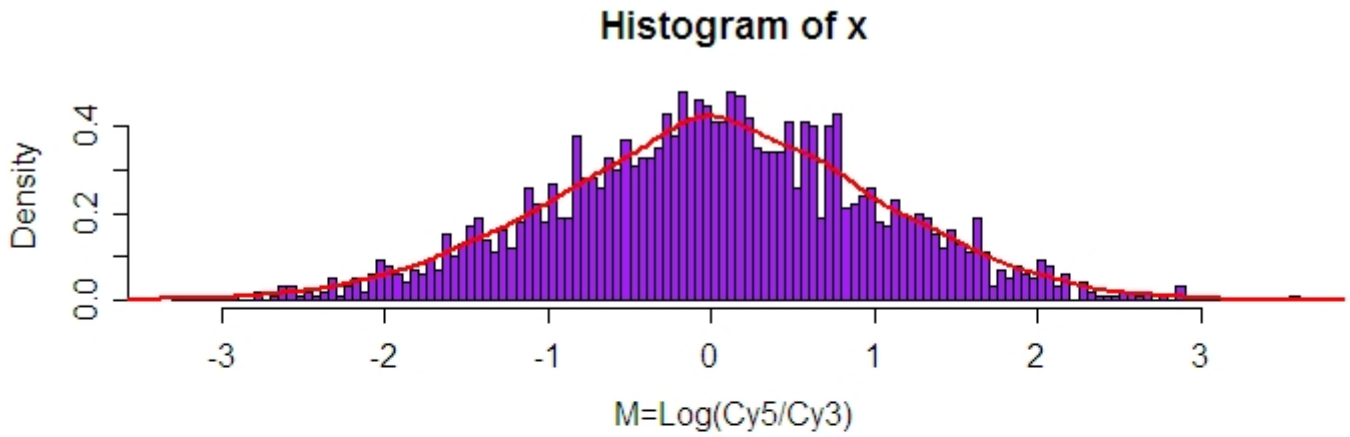
Histogram / Density plots

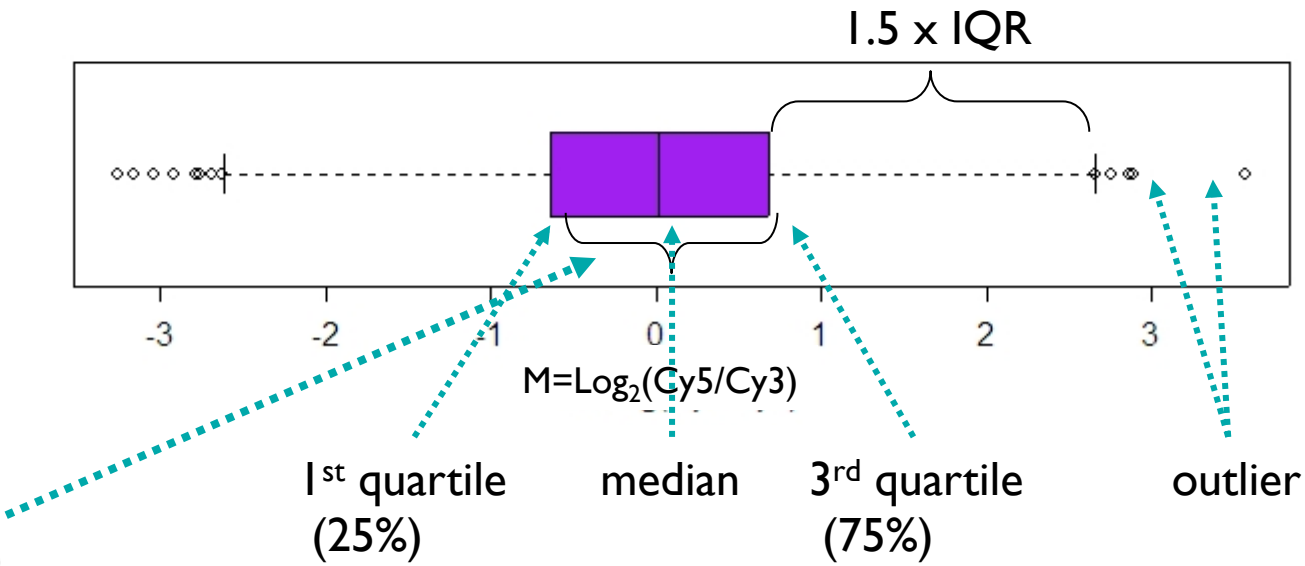
Box plot

Scatter plot or MA-plot

Spatial plot

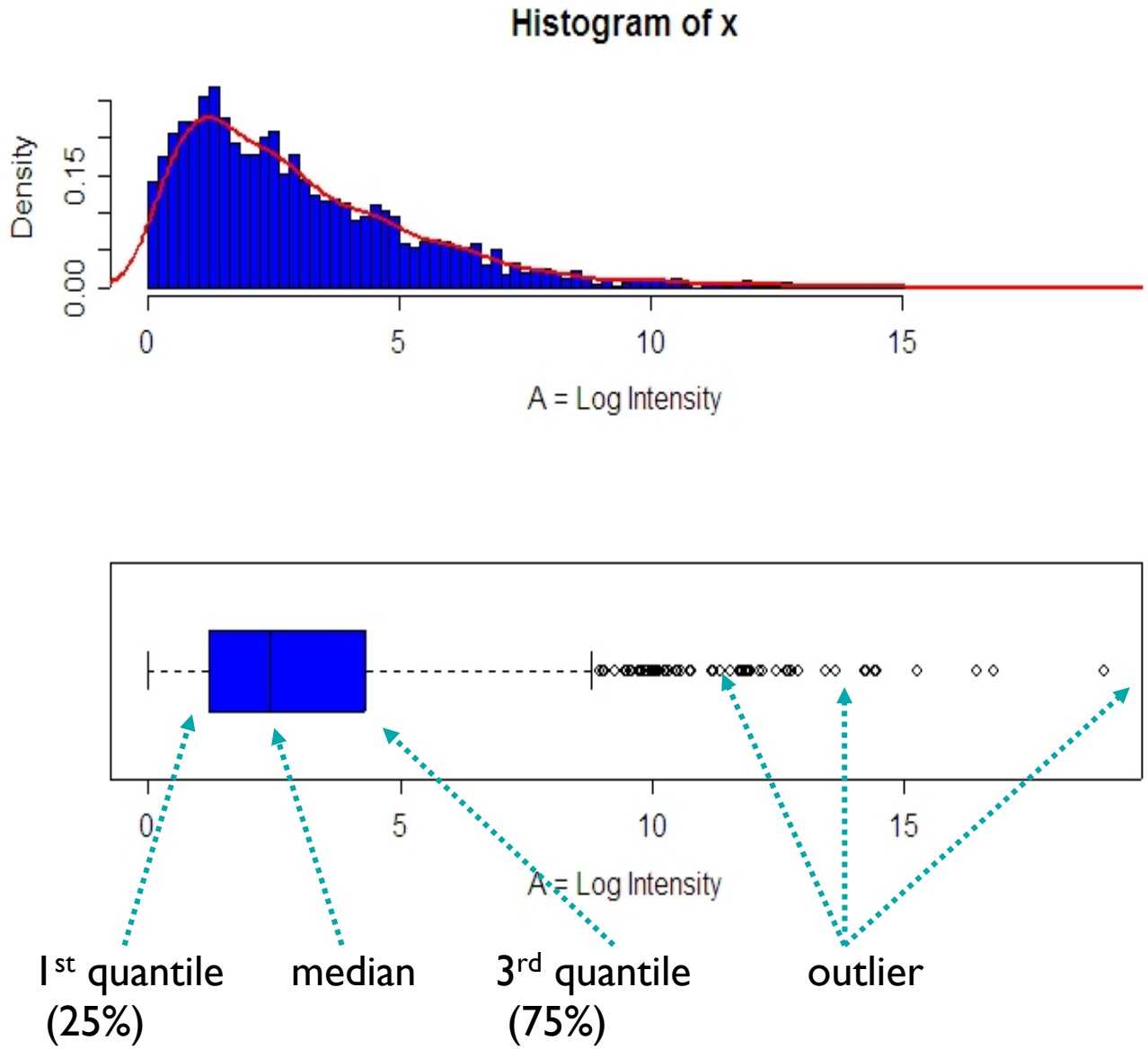
Example 1
- Histogram
- Density line
- Boxplot



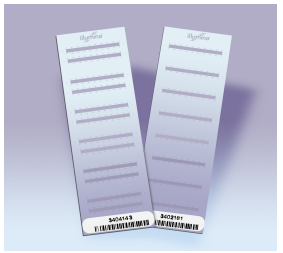


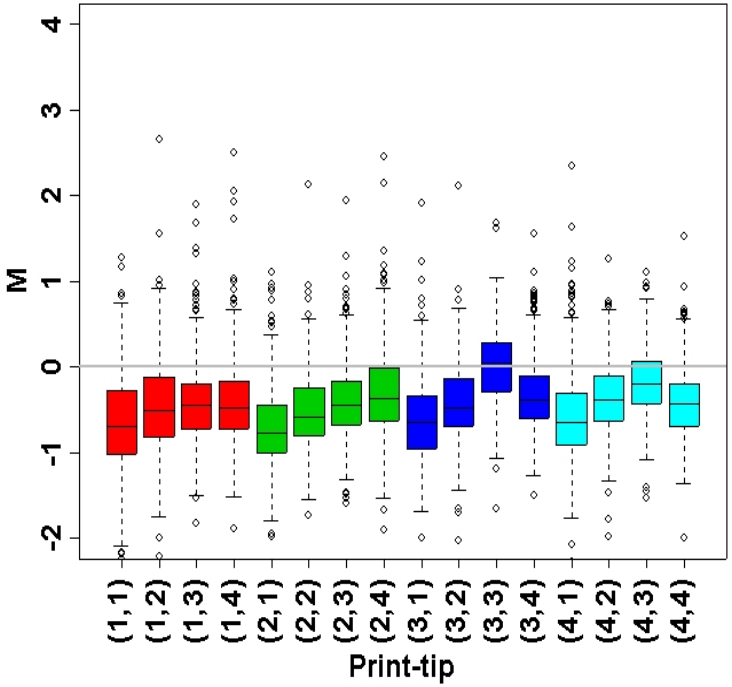
Inter-quartile range (IQR)
a robust measure of the variability

Example 2:
Asymmetric data

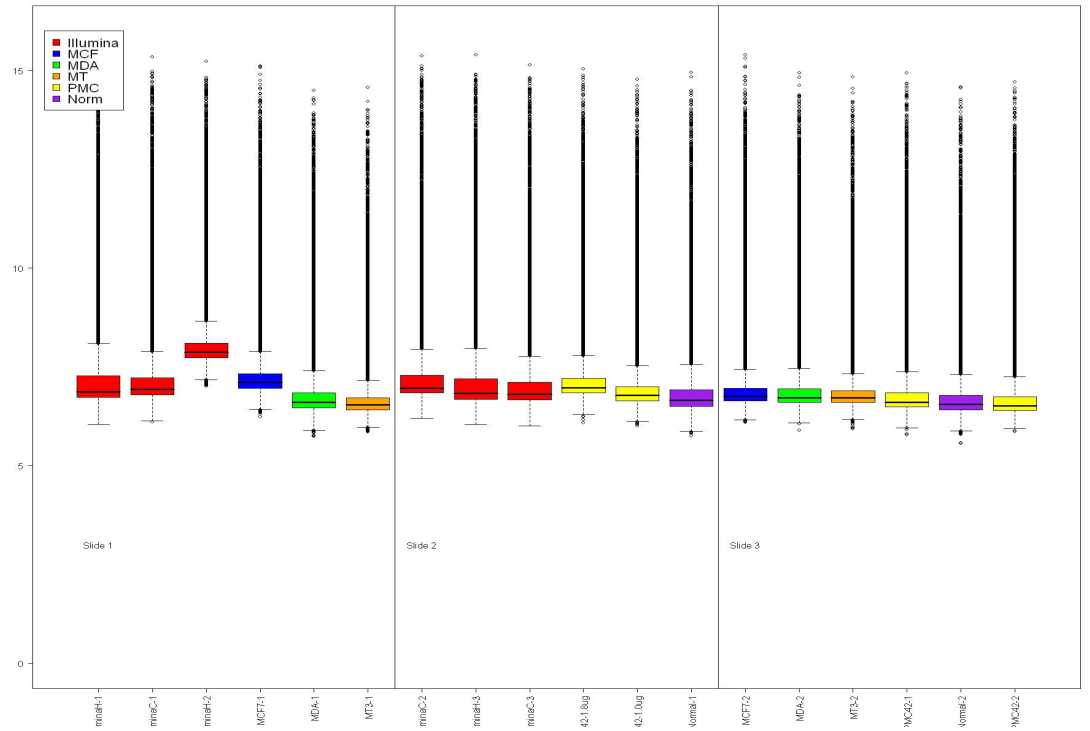


Boxplots



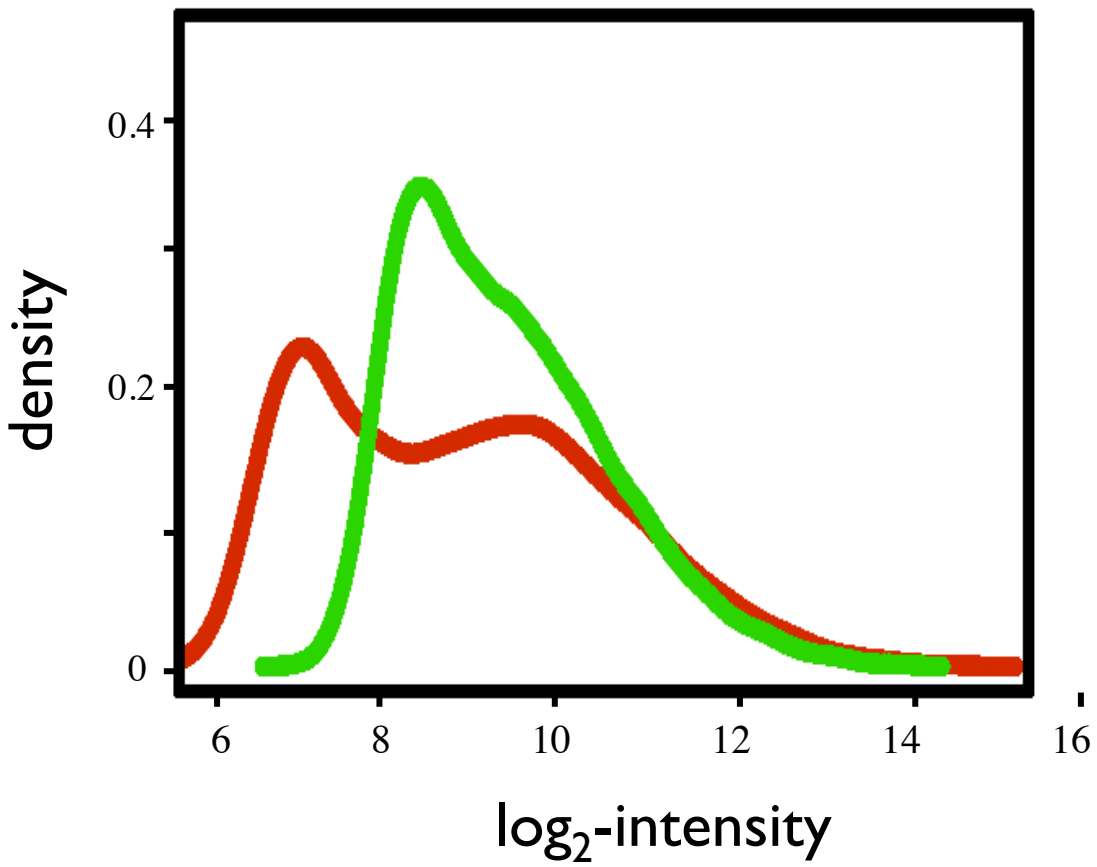


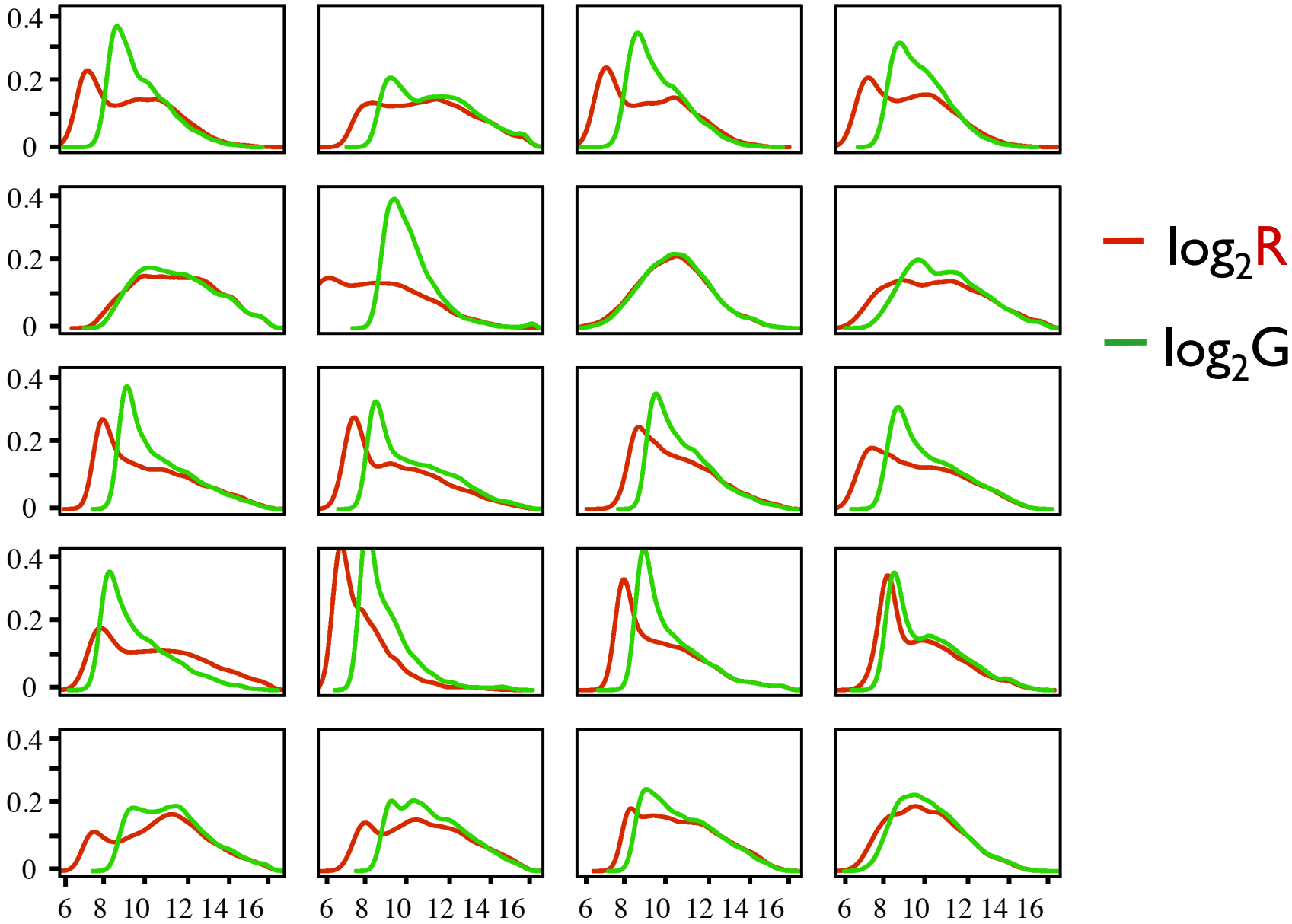
Print-tip / spatial Effects



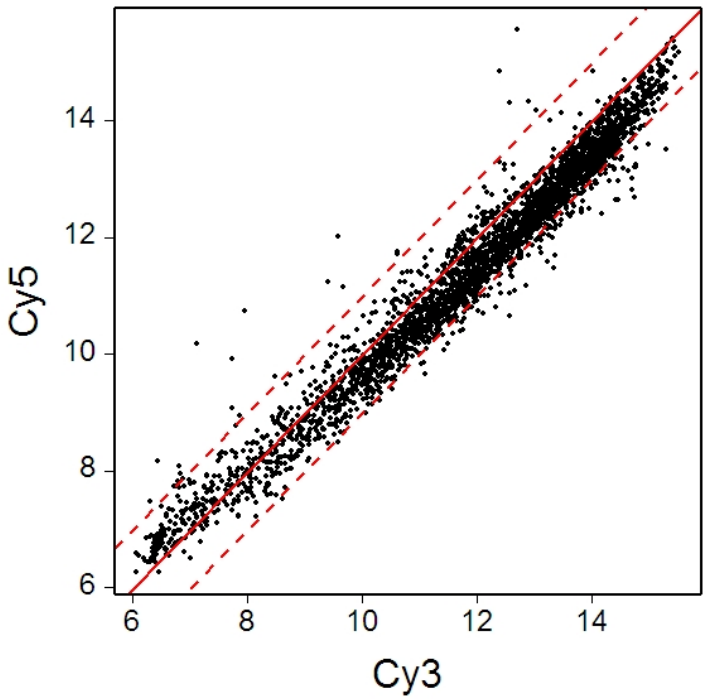
3 Illumina bead chips

Density plots

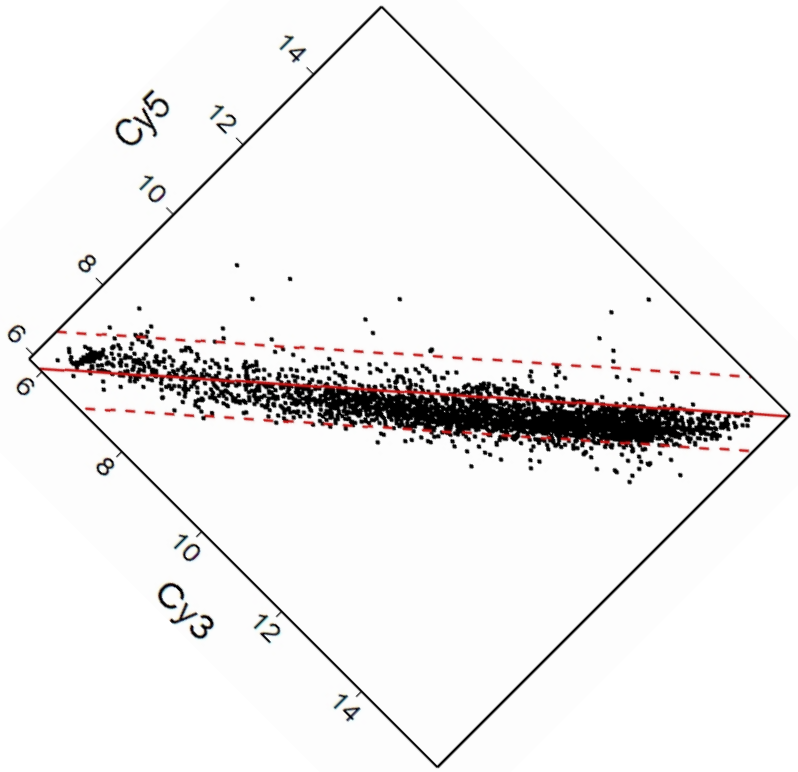



0.4
0.2
0
6 8 10 12 14 16
log2R
log2G

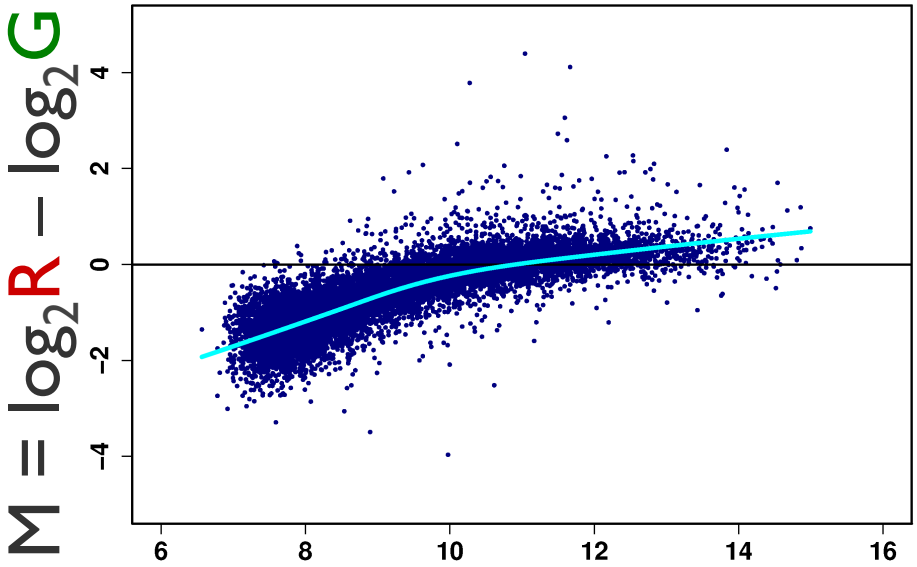
Scatter plots

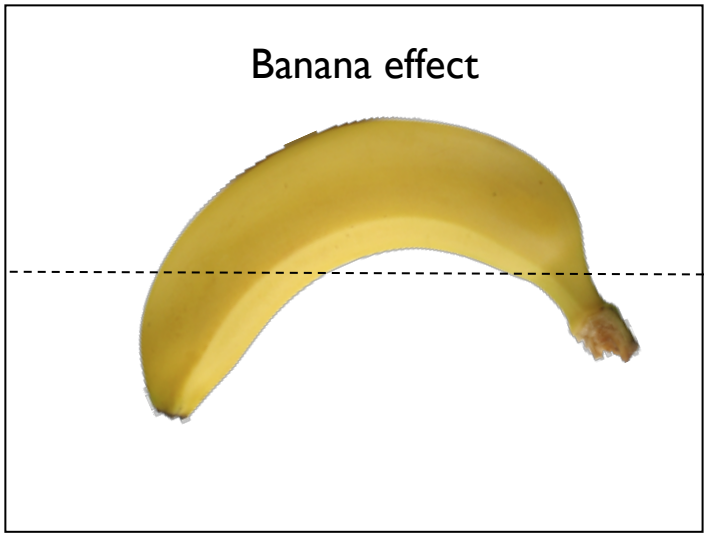


Rotate by
45degree



MA plots





$$A = {}^{1}/_{2} (\log_2 R + \log_2 G)$$

Reasonable Assumption:

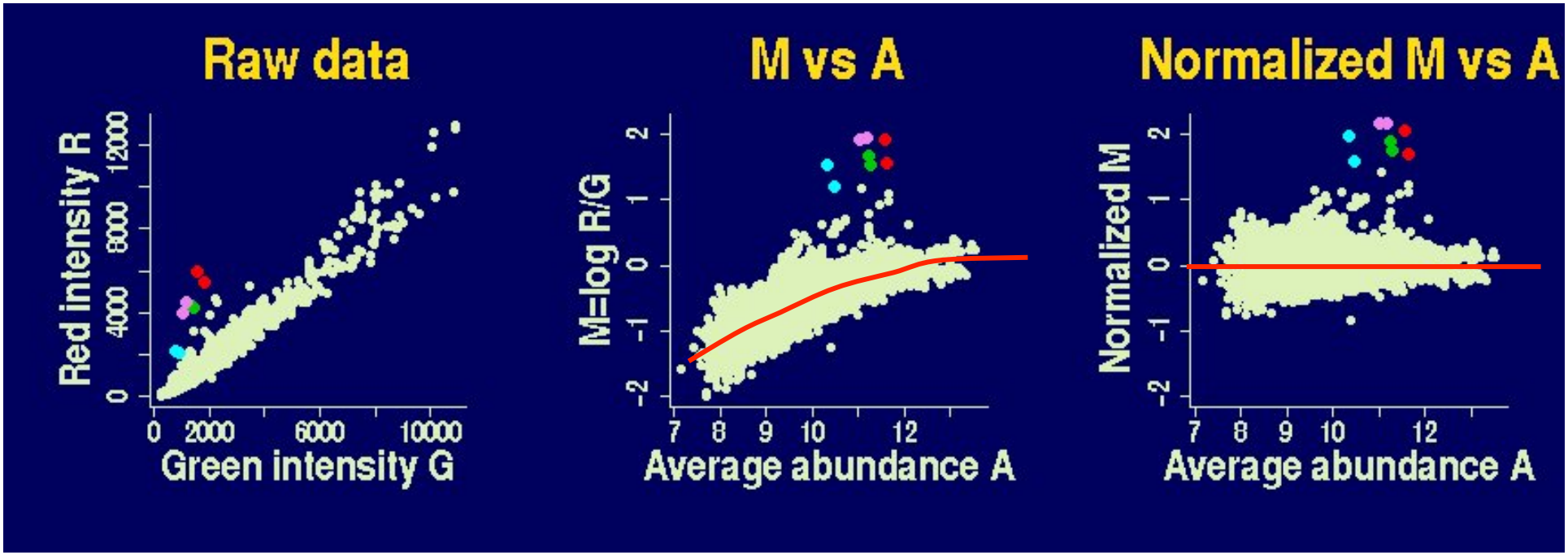
For two colour arrays, in a self-self hybridisation, we expect that for each spot, the intensity in the R channel = that of the G channel

Problem:

This is not necessarily true due to labeling effects, chemistry (dye properties), scanner properties, etc

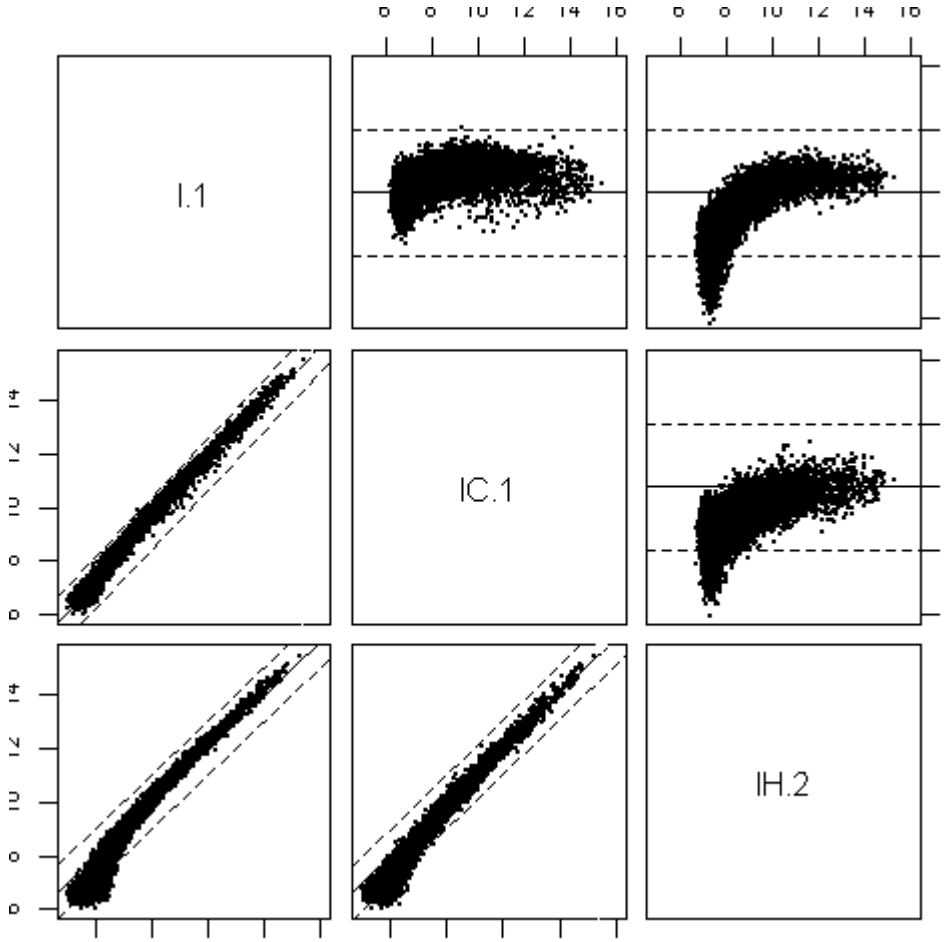
Intensity in one channel (G; Cy3) may be higher than the other (R; Cy5)

Why use an M vs A plot ?



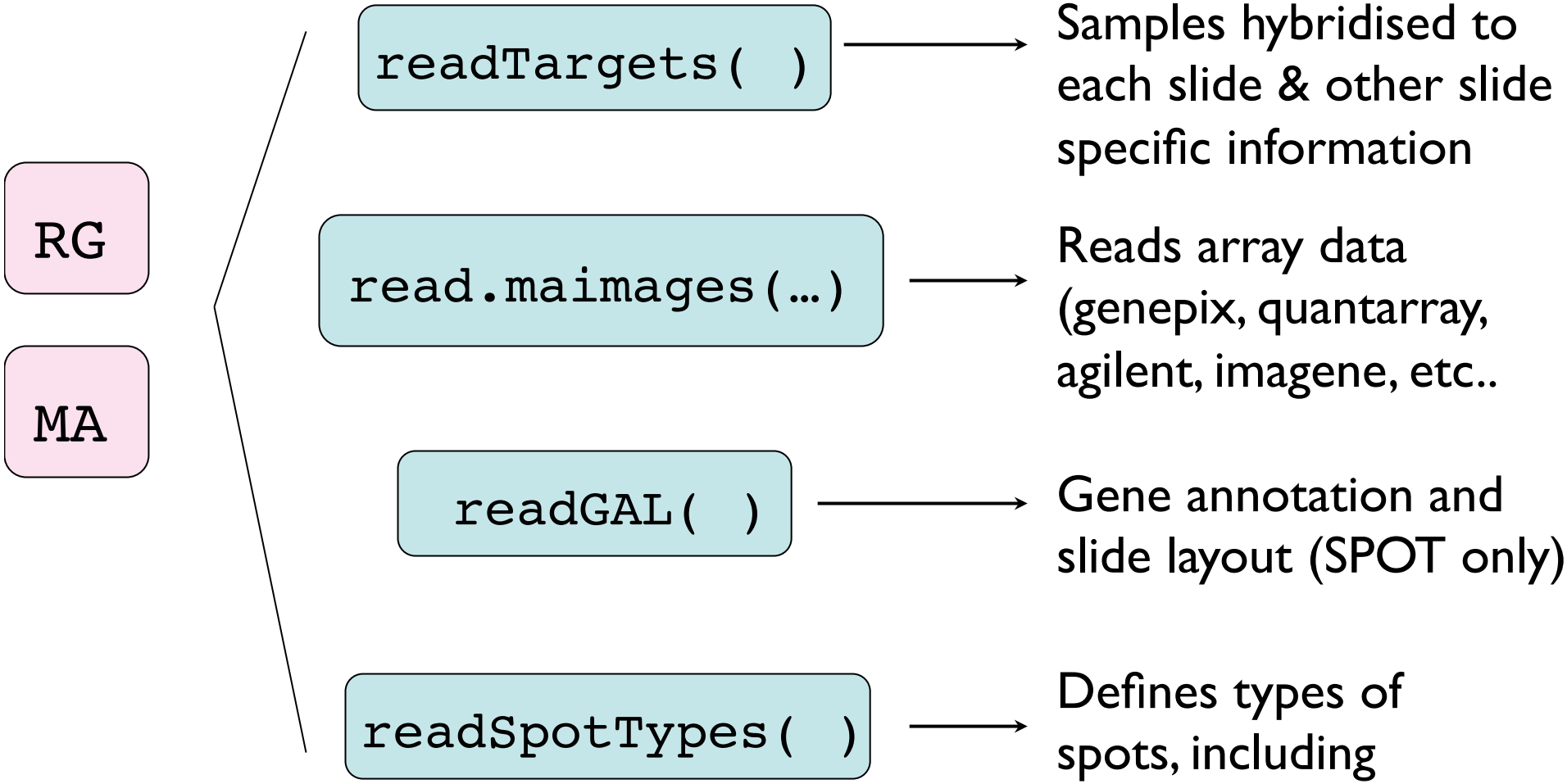
1. Logs stretch out region we are most interested in.
2. Can more clearly see features of the data such as intensity dependent variation, and dye-bias.
3. Differentially expressed genes more easily identified.
4. Intuitive interpretation

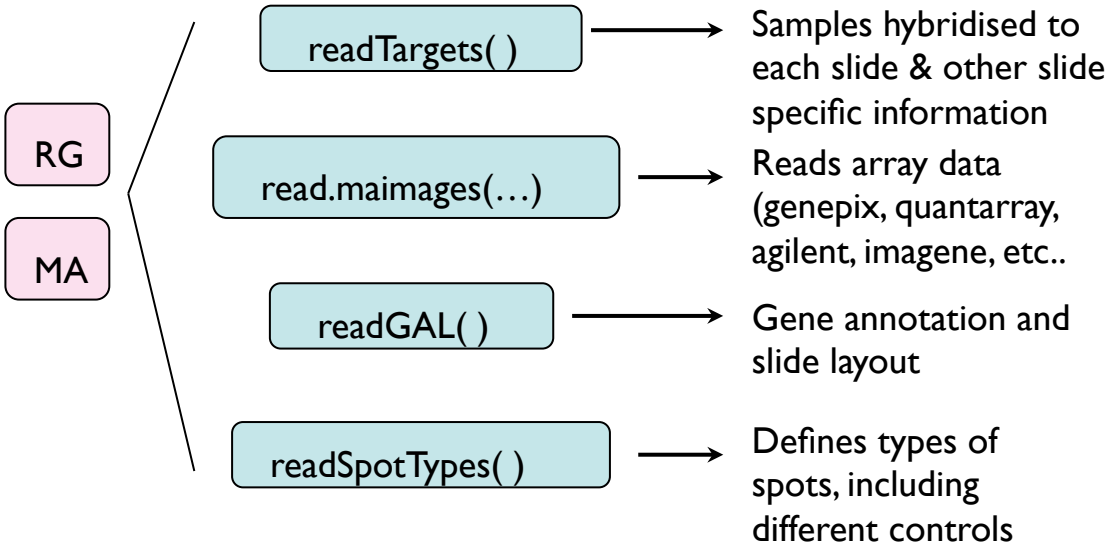
MAXY plots single colour data



Basic exploratory analysis and normalisation in Limma (two-colour data)

LIMMA files / data objects





All this will be covered in the practicals on EDA, normalisation and finding differentially expressed genes using Limma in R.

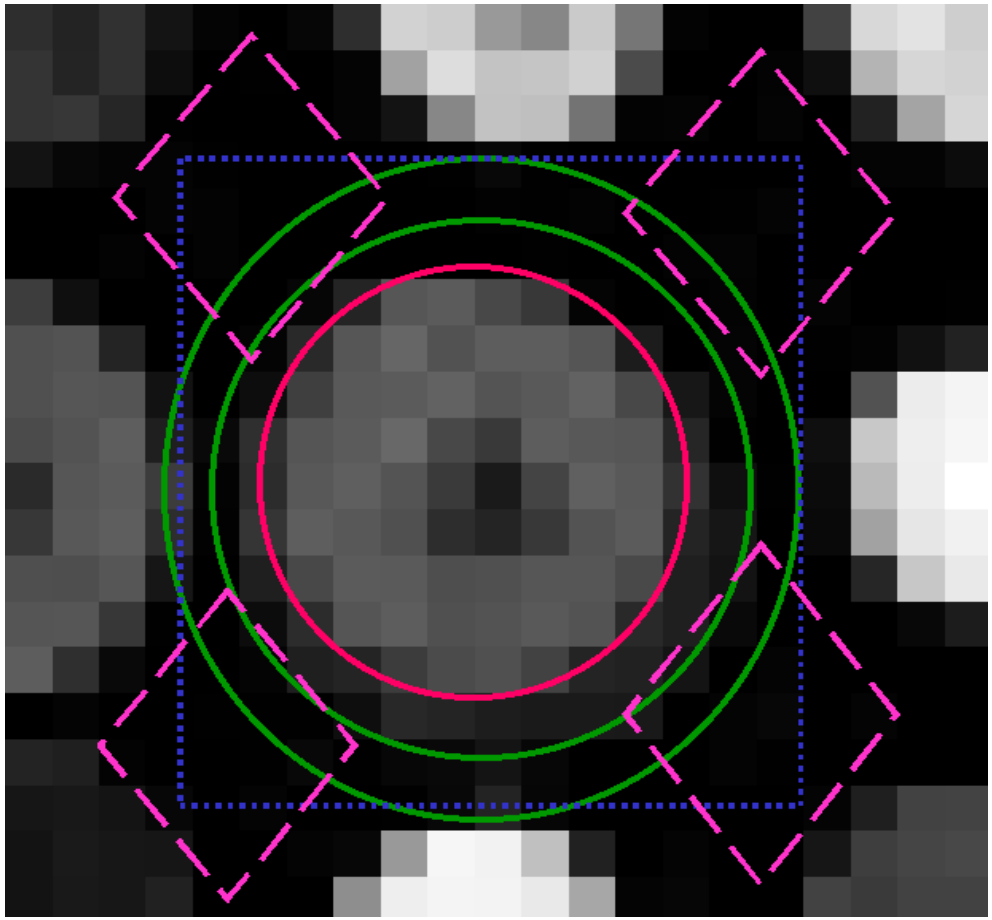
```
backgroundCorrect
```

Local Backgrounds

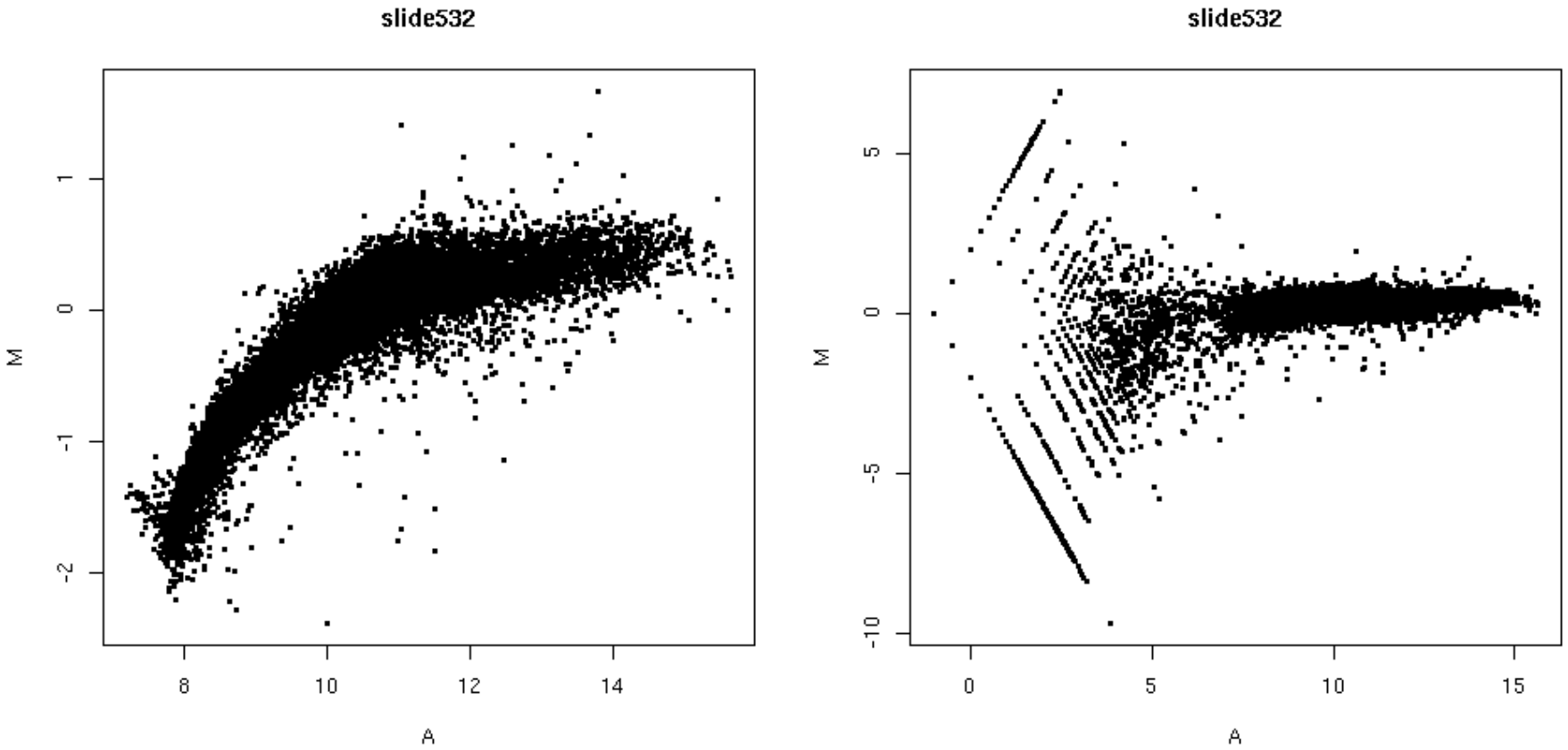
We consider the following as 2 separate problems

- Measure local background signal
- Estimate quantity of background to subtract from spot signal

Our local background estimation method is less likely to be contaminated by neighbouring spots

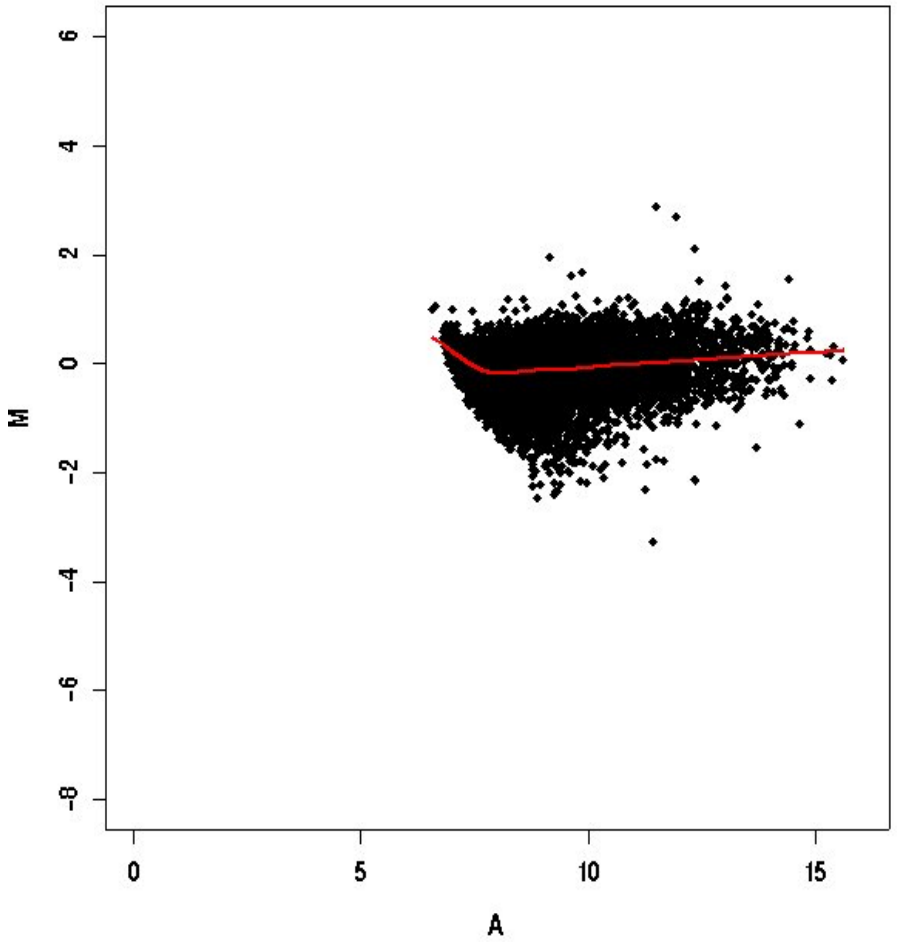


`backgroundCorrect` → EDA check → `plotMA`

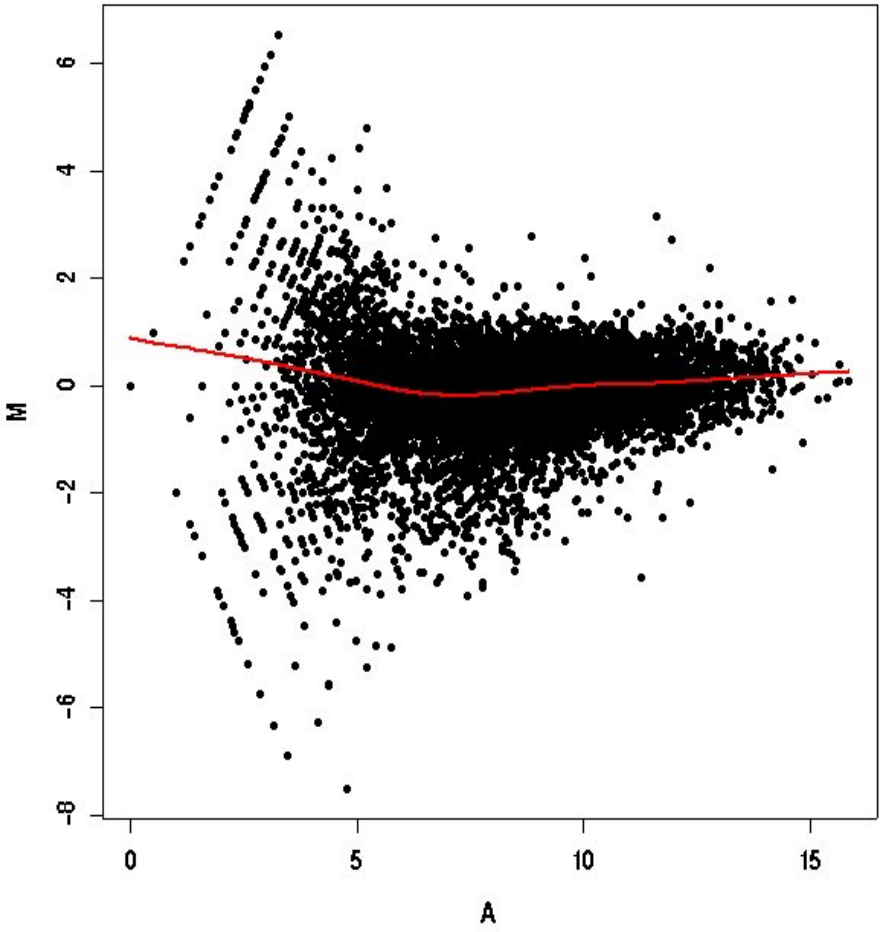


Often large amounts of noise are introduced by background correction. Background levels can be dependent on the image analysis program used.

Background matters

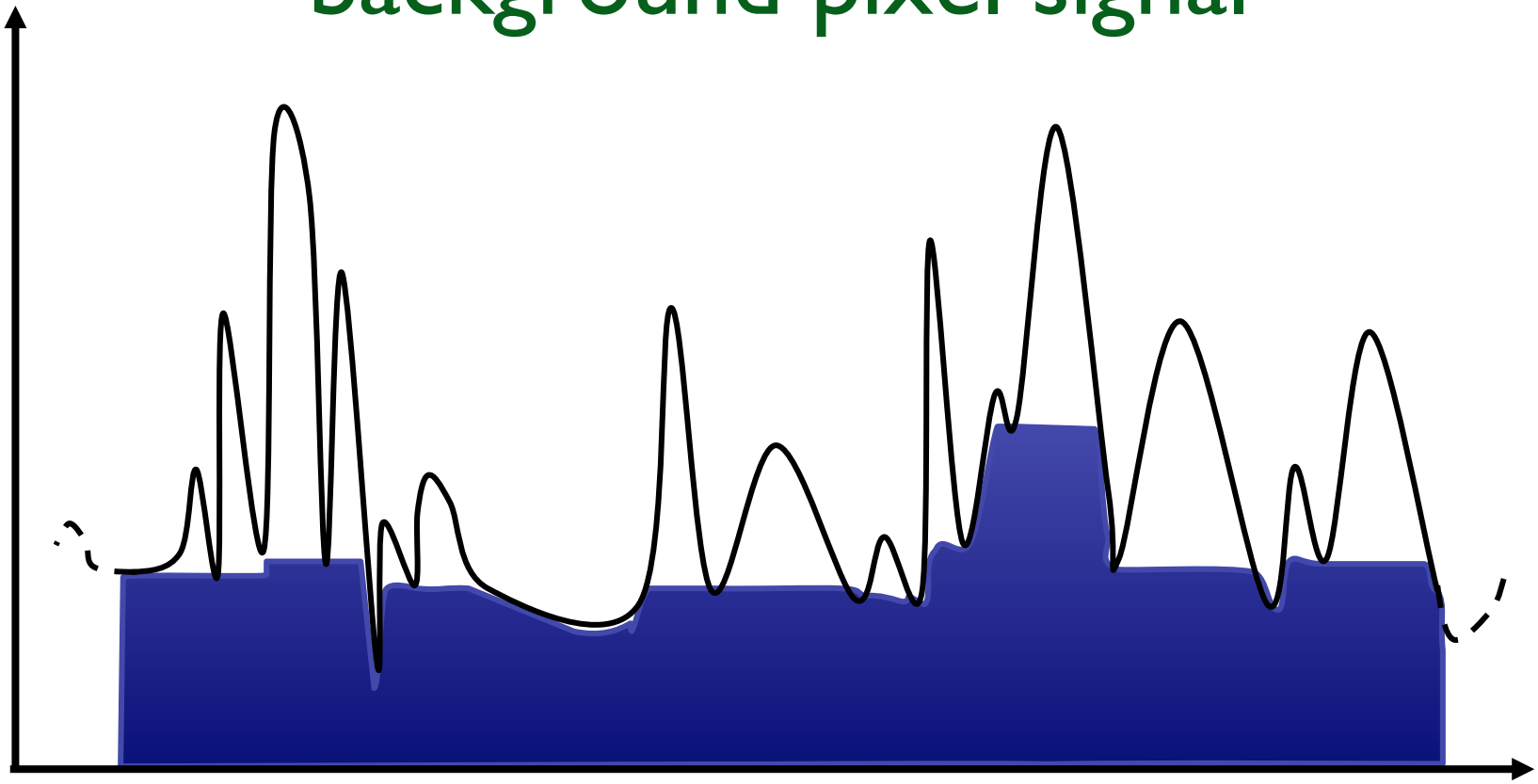


From Spot

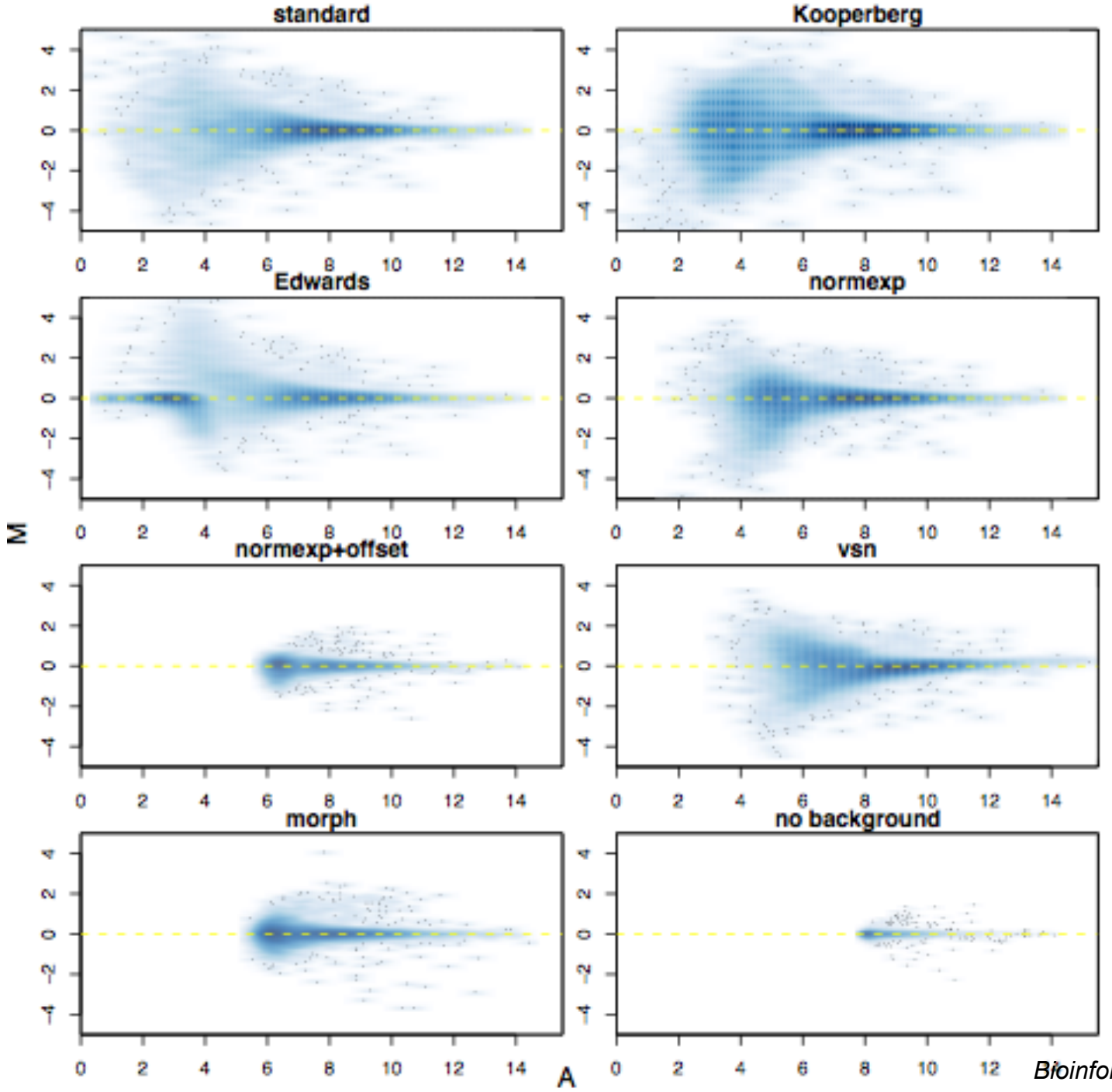


From GenePix

Morphological non-linear filter on background pixel signal

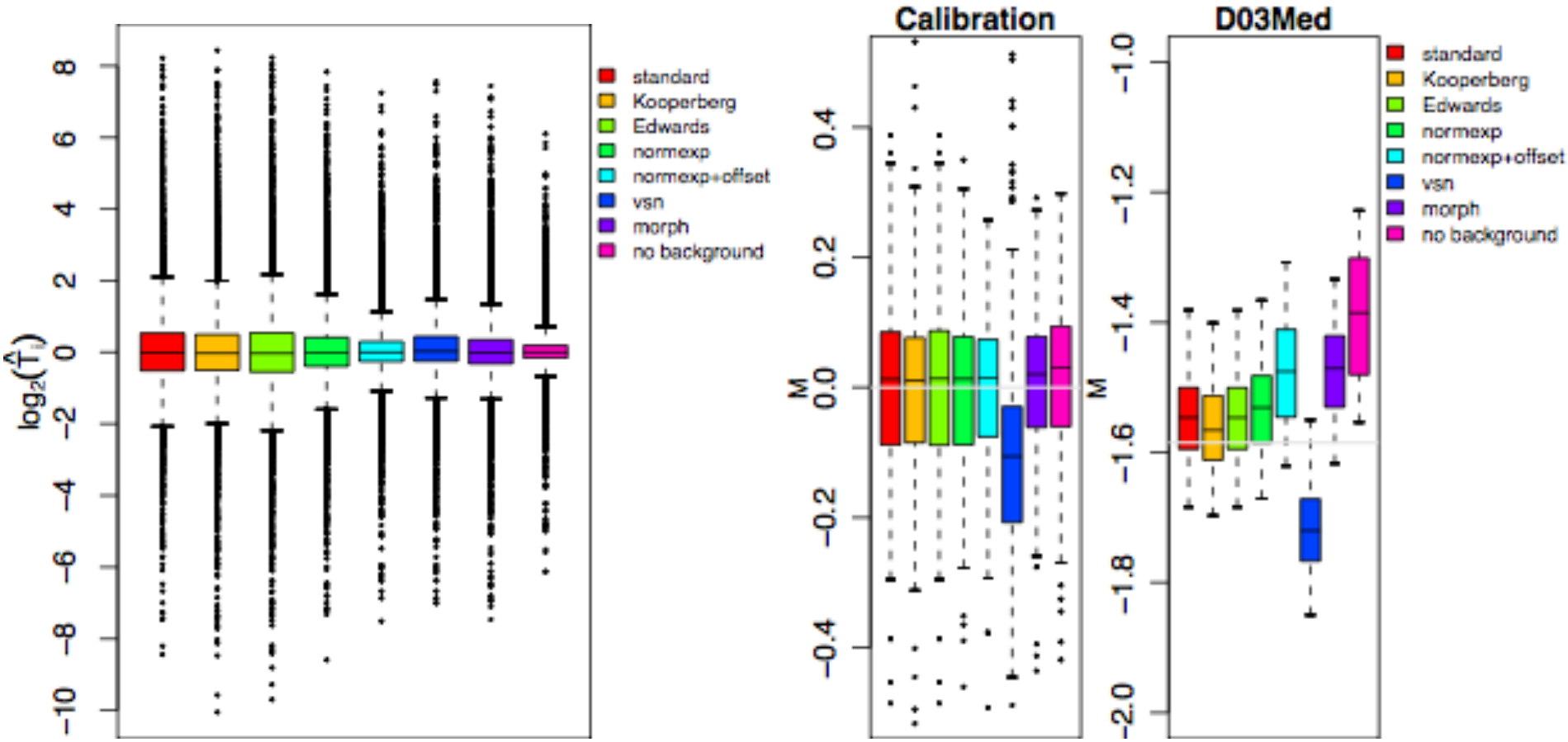


Measures overall baseline background level



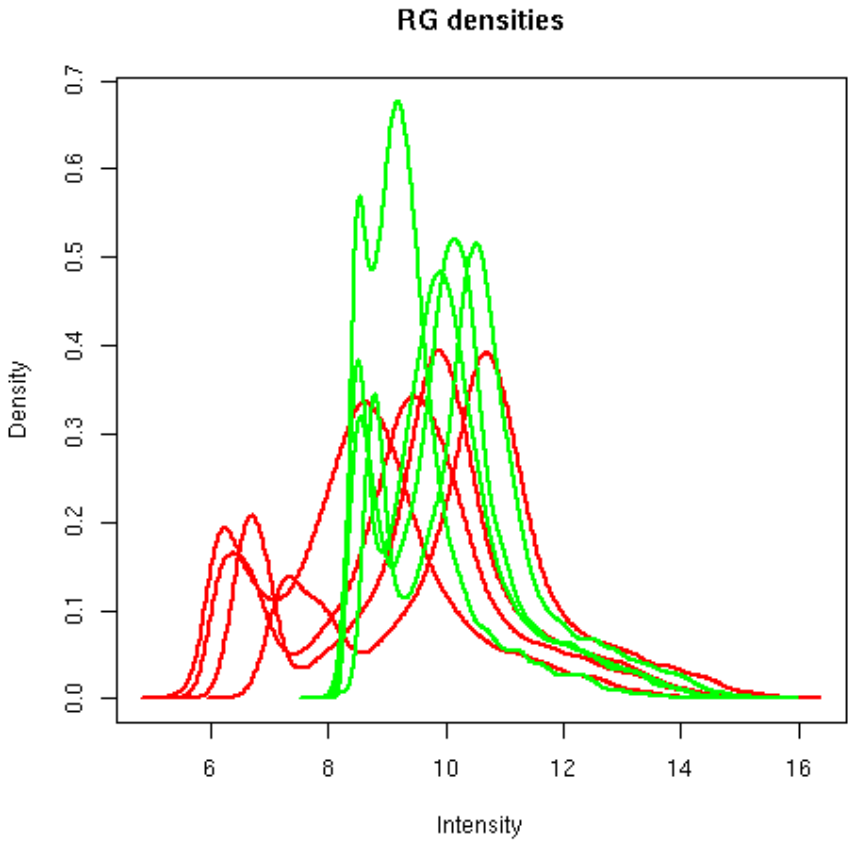
Ritchie M et al
Bioinformatics 2007 Sep 1

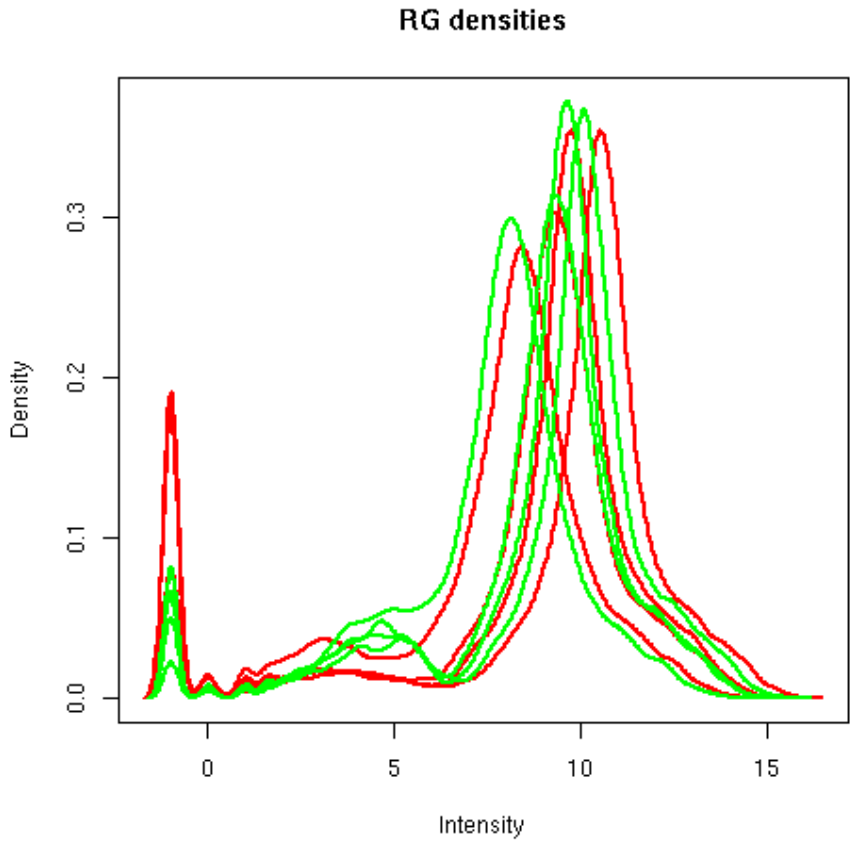
Background matters



Ritchie M et al
Bioinformatics 2007 Sep 1

`backgroundCorrect` → EDA check → `plotDensities`

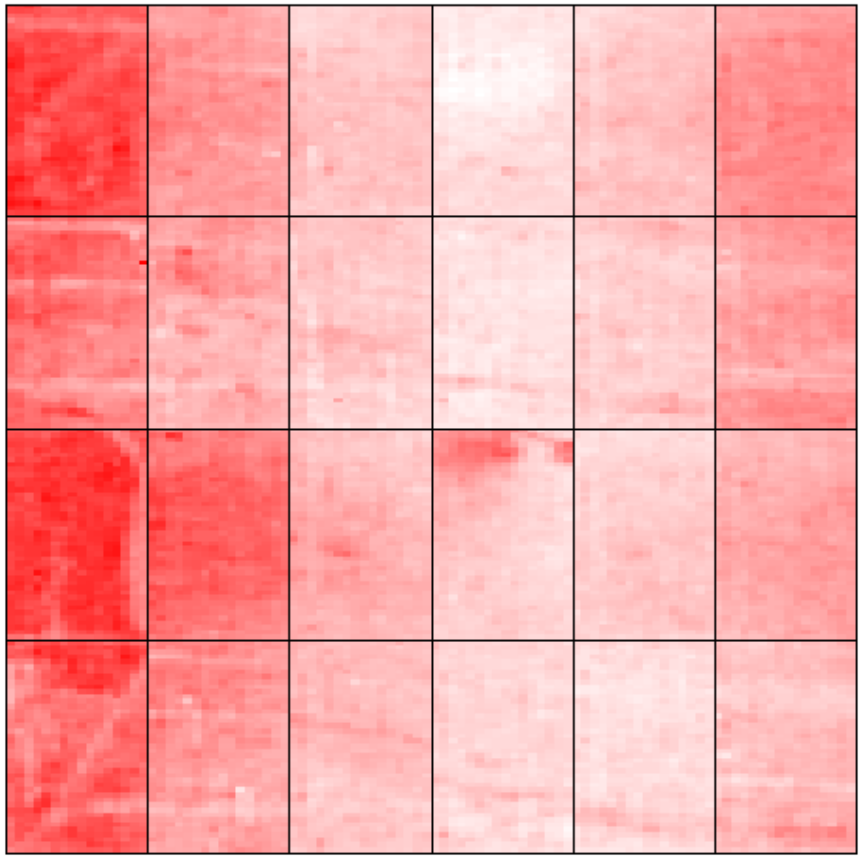


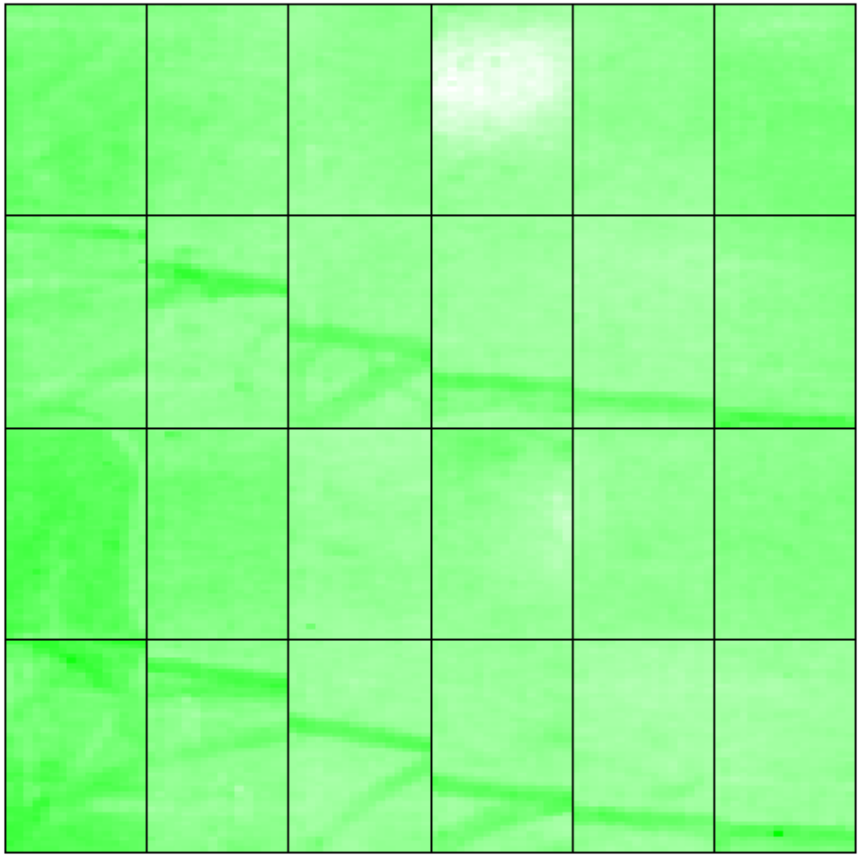


No background correction

Background corrected

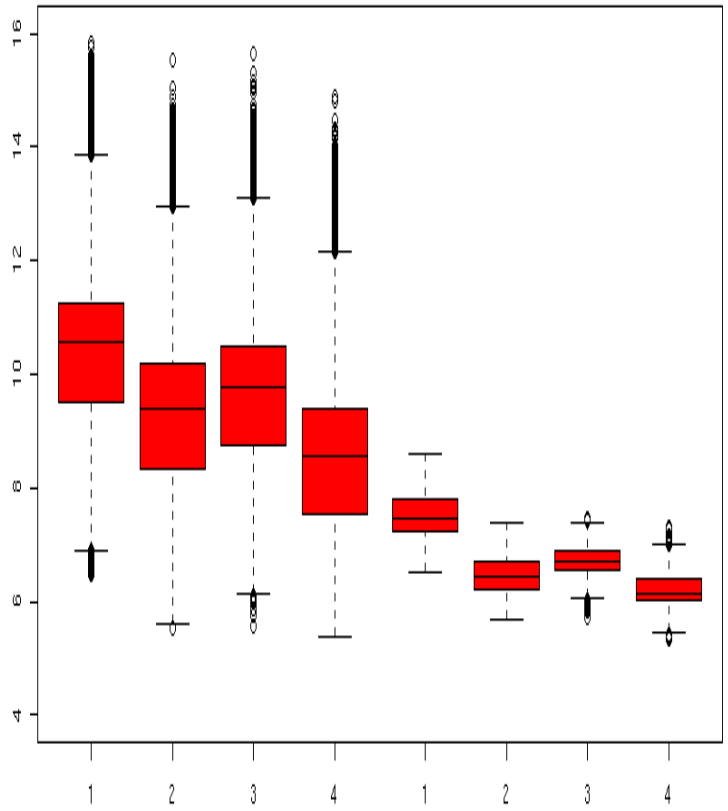
`backgroundCorrect` → EDA check → `imageplot`

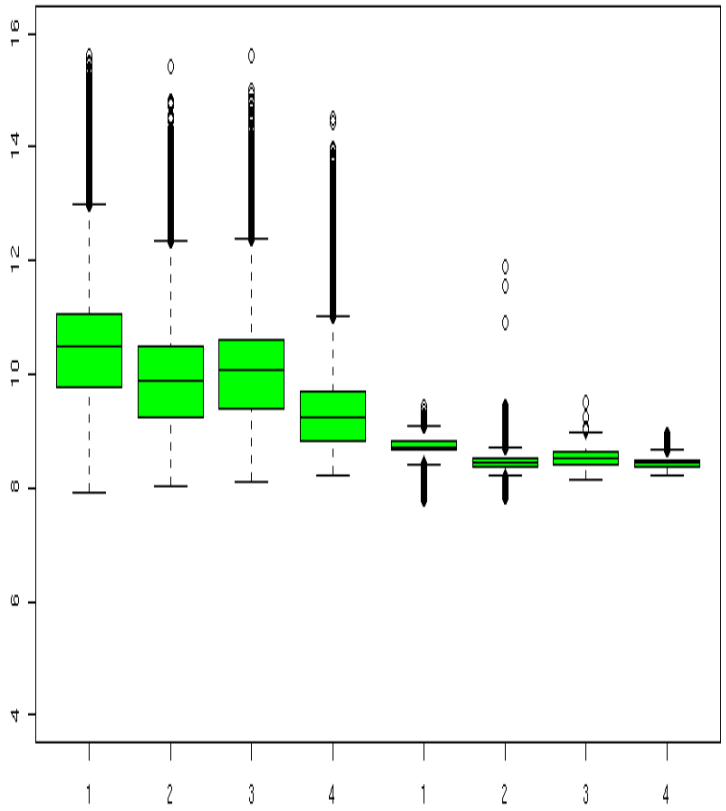




Are there significant spatial effects identifiable in the background?

`backgroundCorrect` → EDA check → `boxplot`





Are the background levels too high?

Is there any difference between Cy5 & Cy3?

backgroundCorrect

`normalizeWithinArrays`

`normalizeWithinArrays`

Why normalise?

To correct for systematic differences between samples on the same slide, or between slides, which do not represent true biological variation between samples.

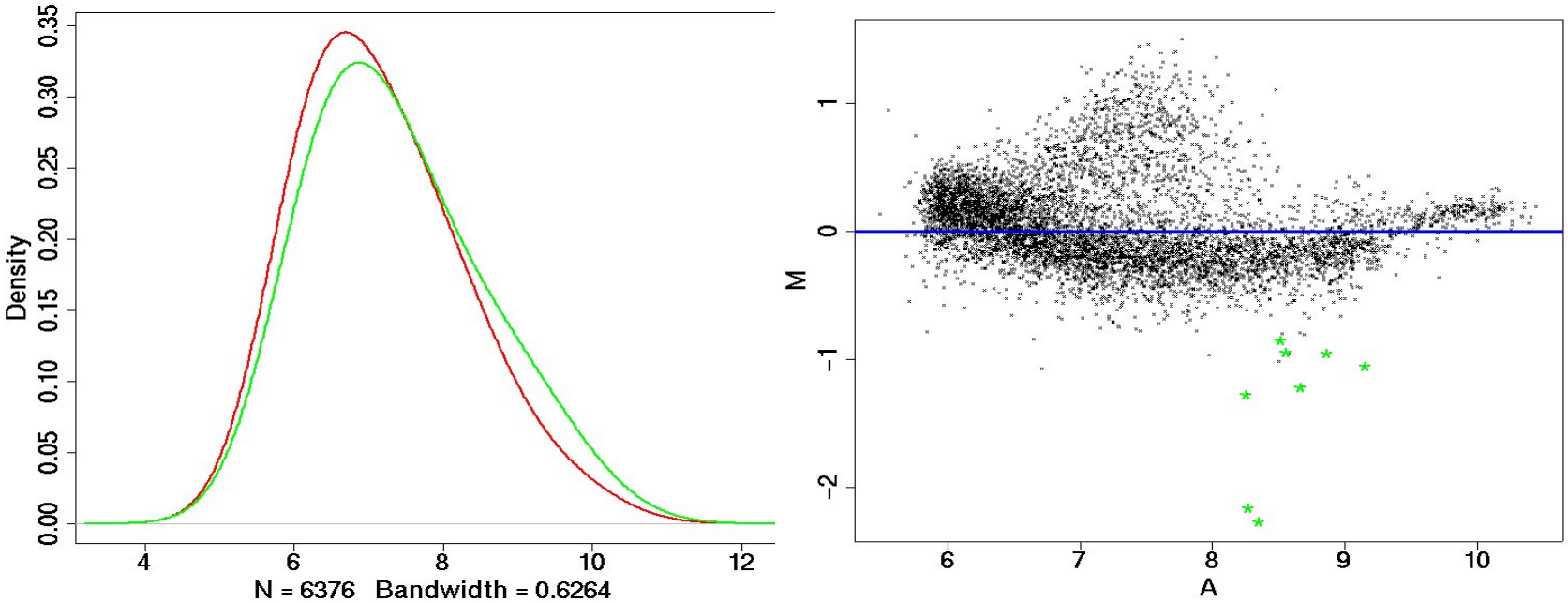
How do we know it is necessary?

By examining replicate and or self-self hybridizations, where no true differential expression is occurring.

We find biases which vary with overall spot intensity, location on the array, dye, plate origin, pins, scanner, scanning parameters,….

`normalizeWithinArrays`

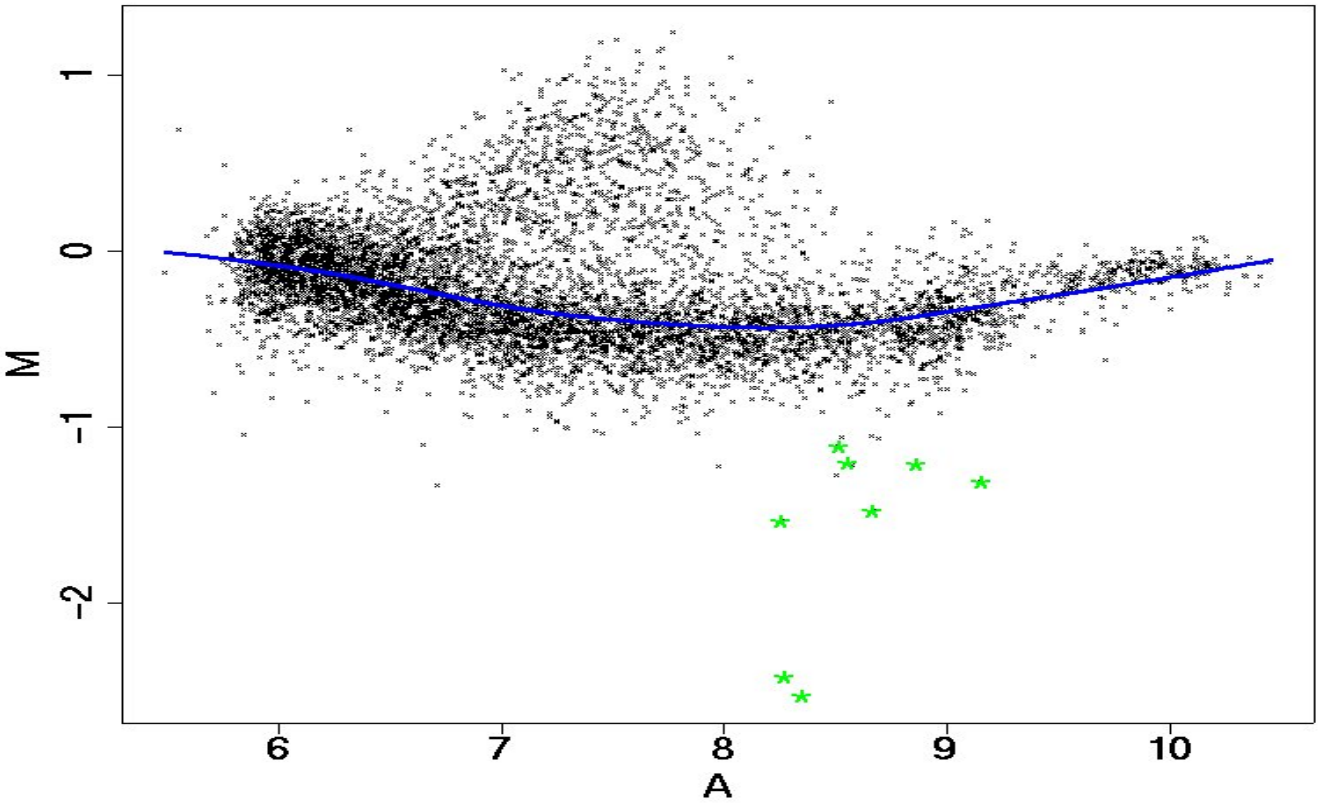
`method="median"`



- Assumption: Changes roughly symmetric
- First panel: smooth density of $\log_2 G$ and $\log_2 R$.
- Second panel: M vs A plot with median put to zero

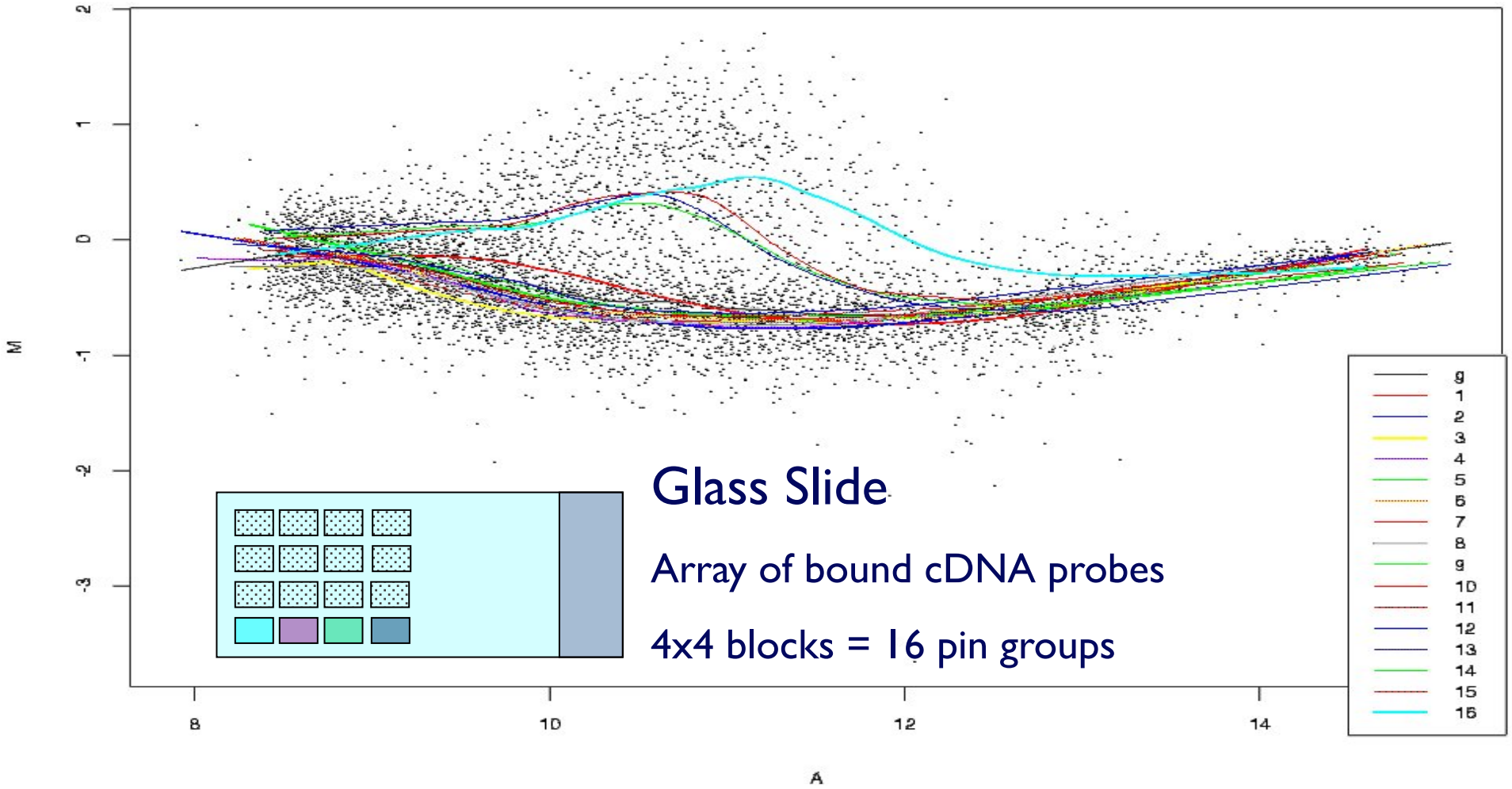
`normalizeWithinArrays`

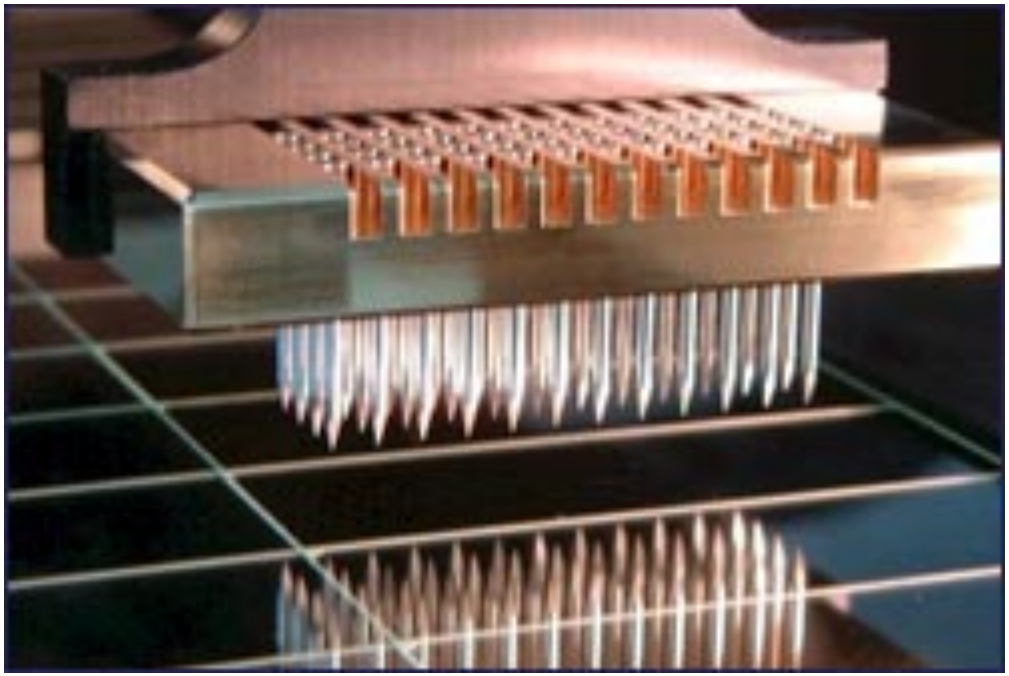
`method="loess"`



Assumption: changes roughly symmetric at all intensities.

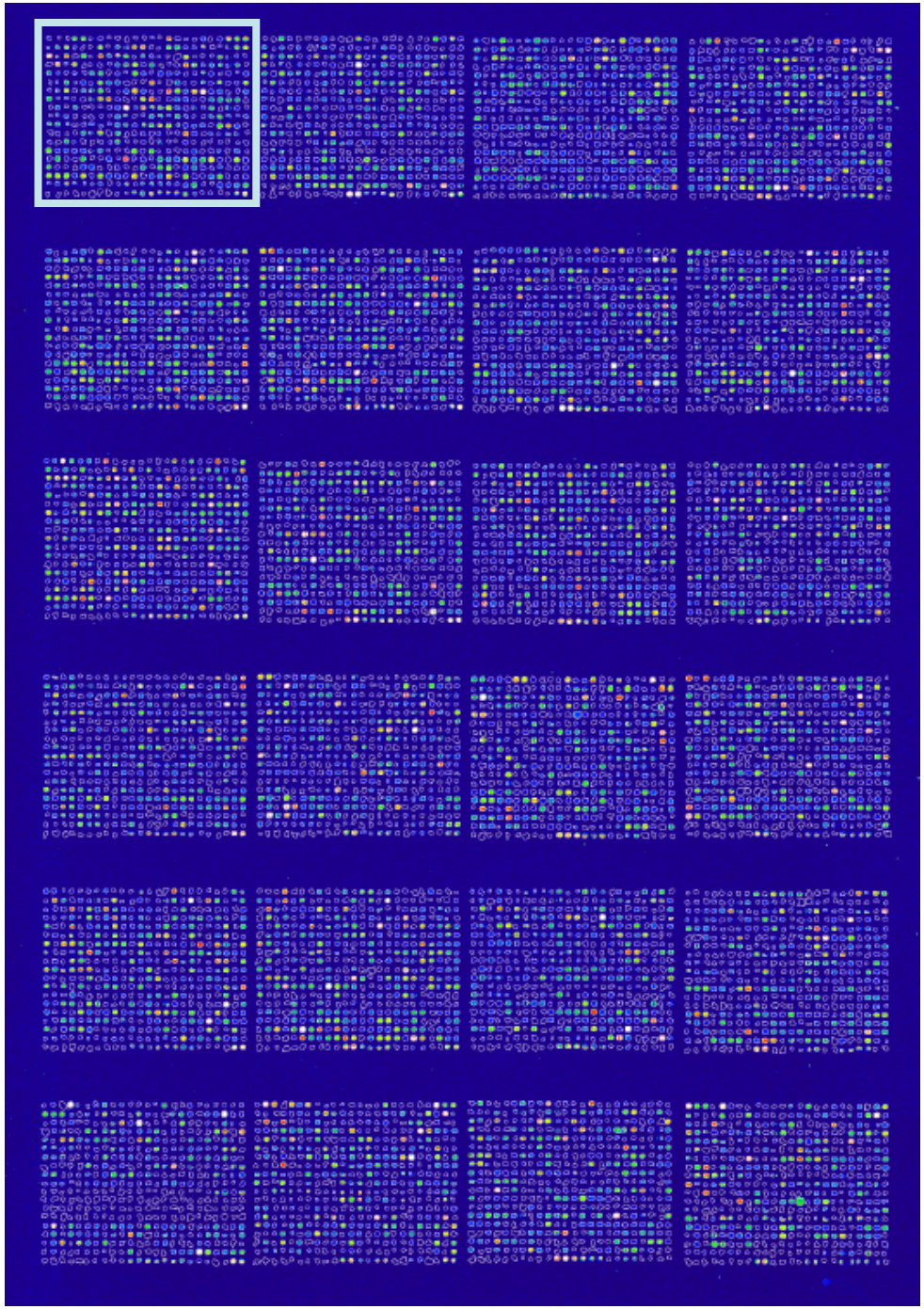
`normalizeWithinArrays` **Default**





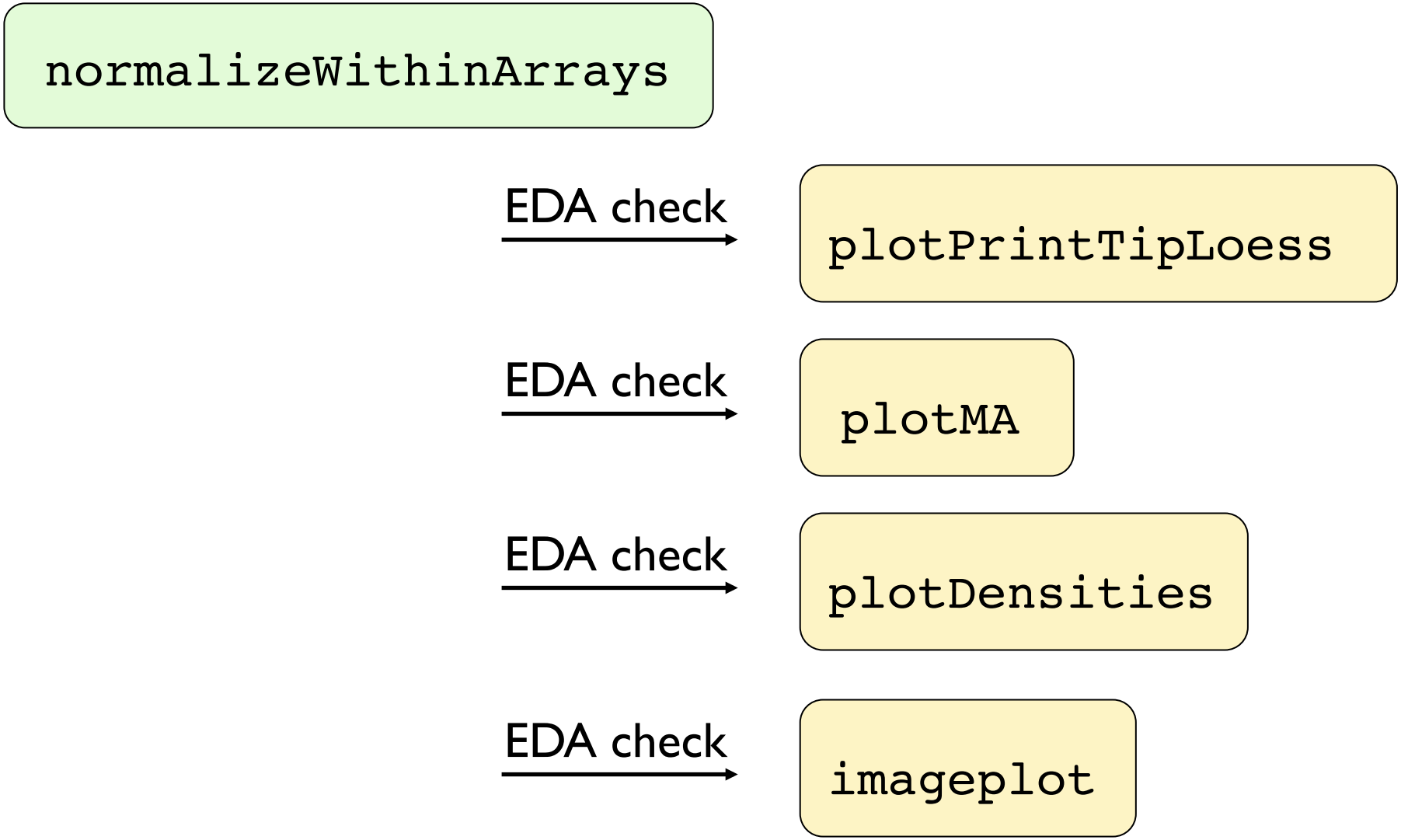
Spots printed in 12 x 4 grid layout corresponding to pins on print-head, traditional way to make microarrays.

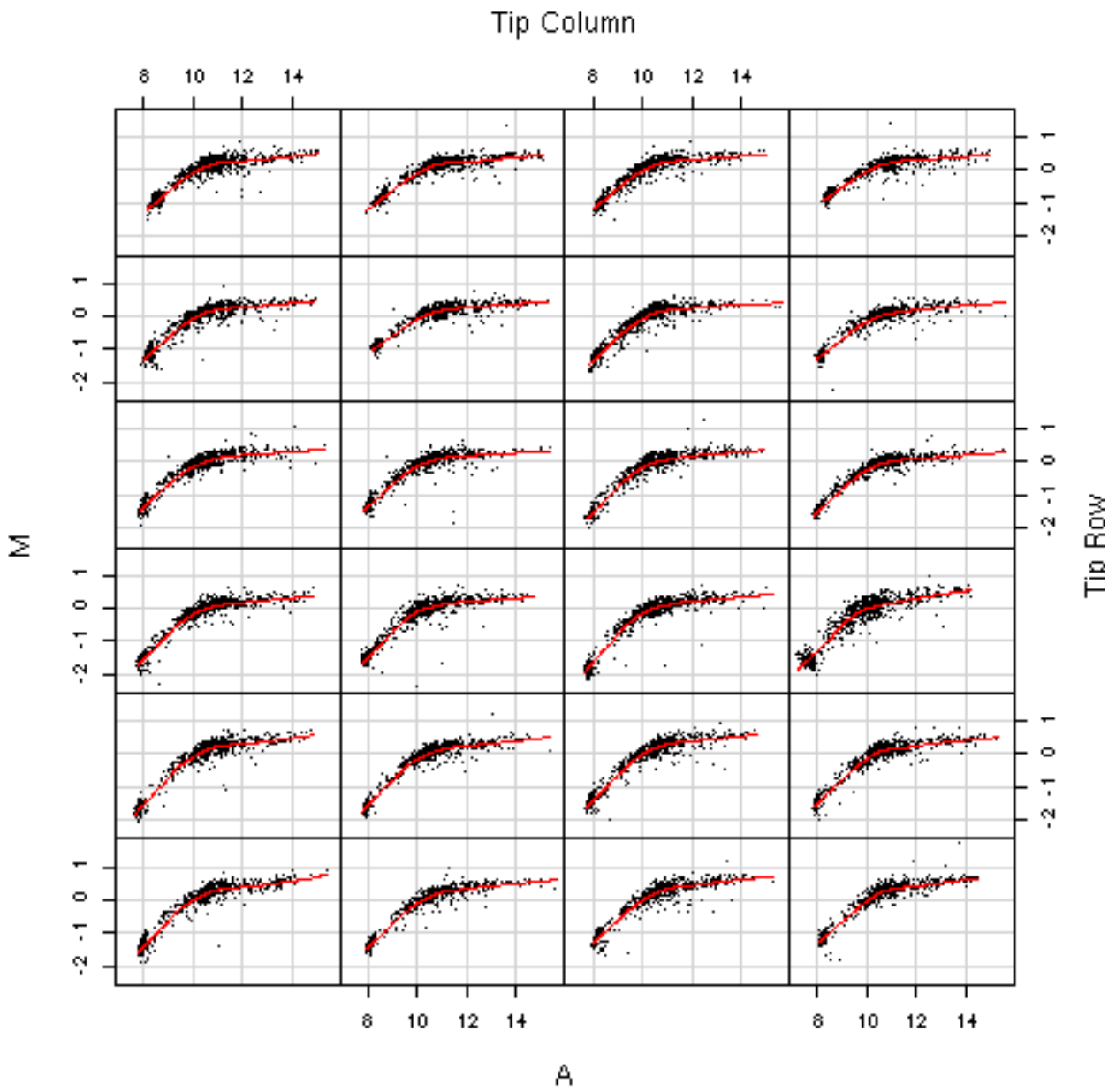
What do pin groups look like?



6 x 4

pin-groups


normalizeWithinArrays
EDA check
plotPrintTipLoess
EDA check
plotMA
EDA check
plotDensities
EDA check
imageplot



Is a print tip group loess normalisation necessary?

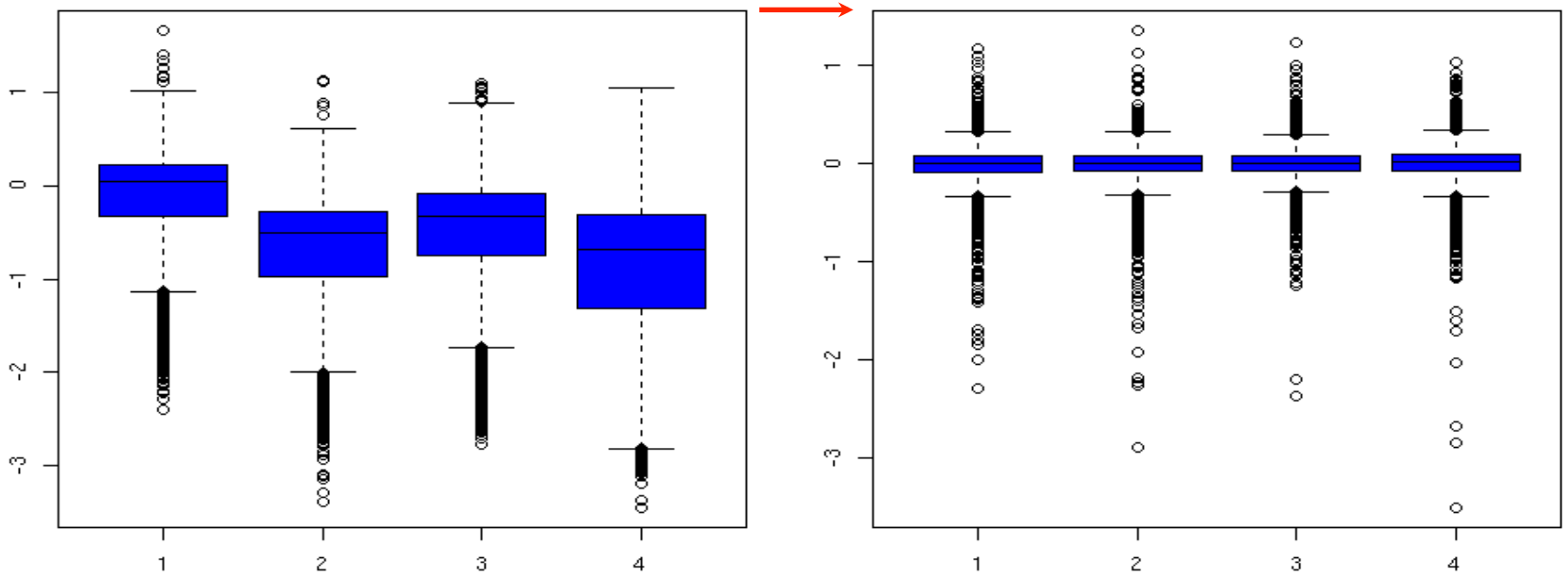
backgroundCorrect

`normalizeWithinArrays`

`normalizeBetweenArrays`

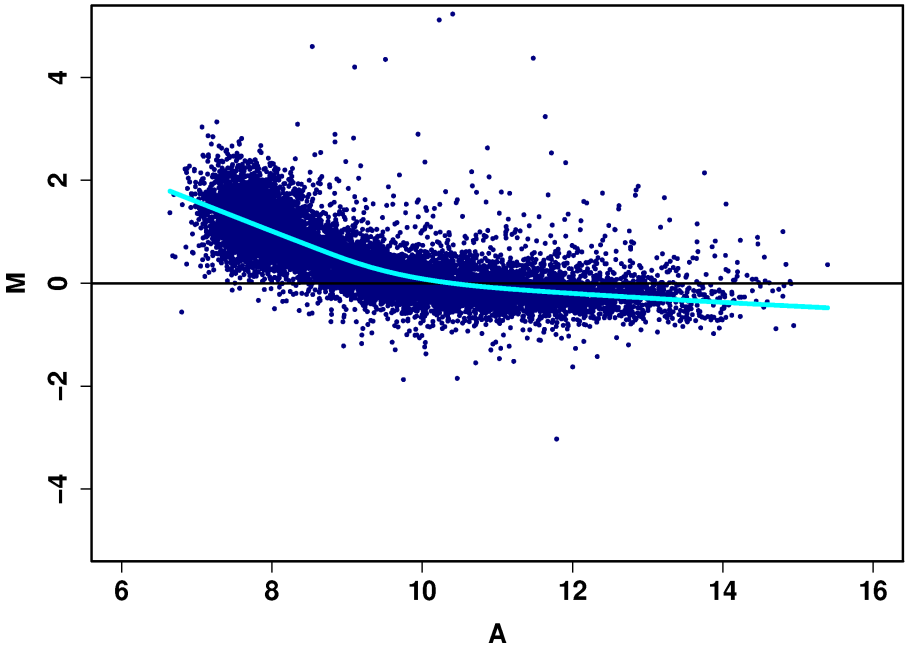
`normalizeBetweenArrays` → EDA check → `boxplot`

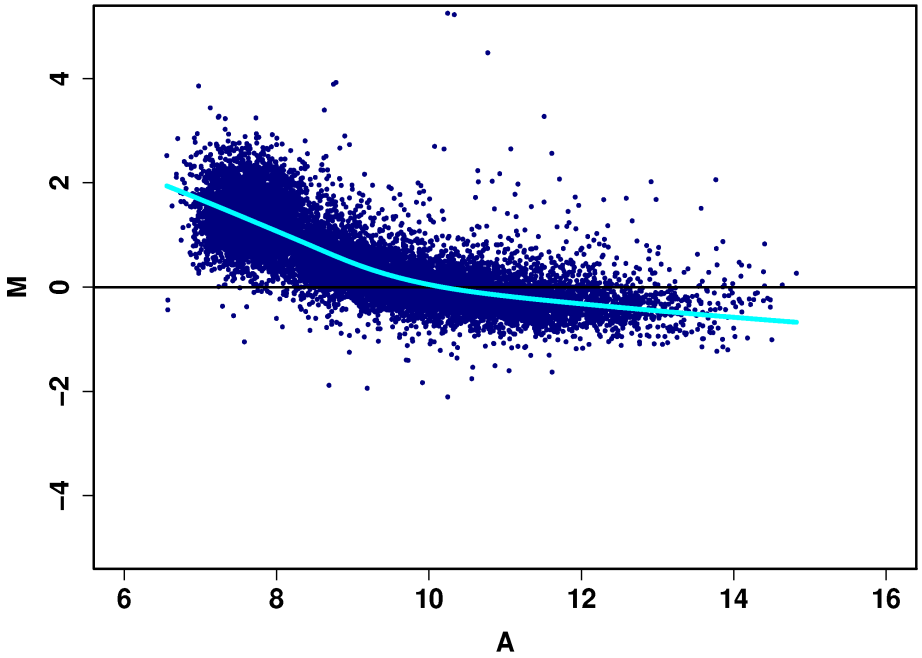
NWA

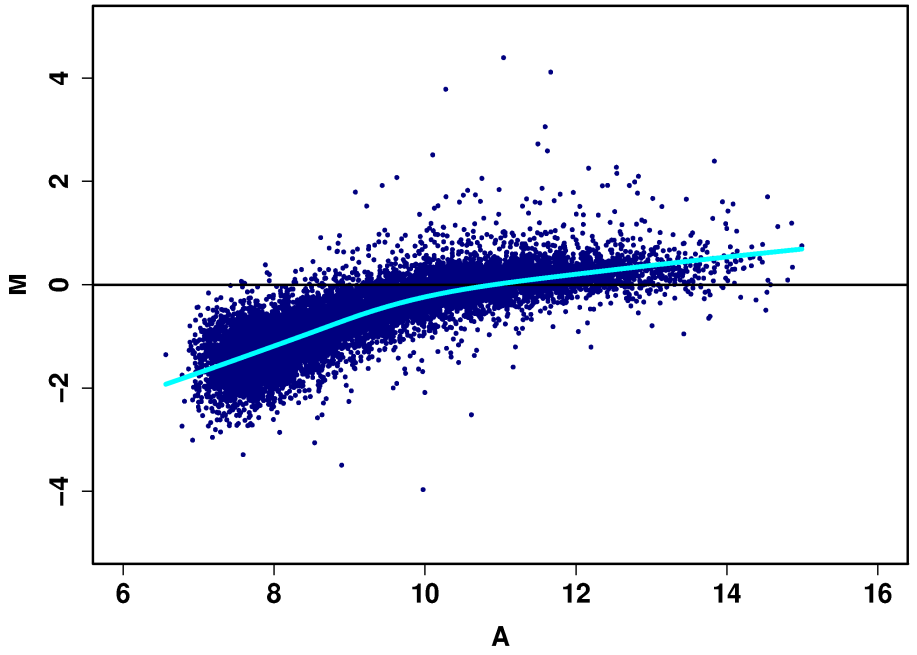


Checking that the scale of M-value measurements are similar between arrays - here we have no need to perform additional normalisation between arrays.

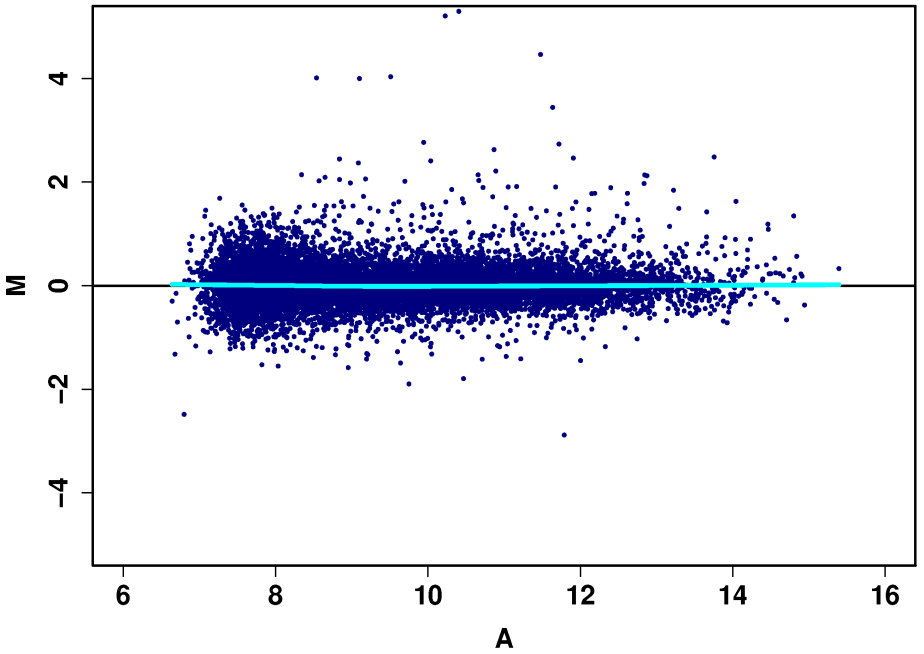
Normalisation enables experiments to be combined/ compared

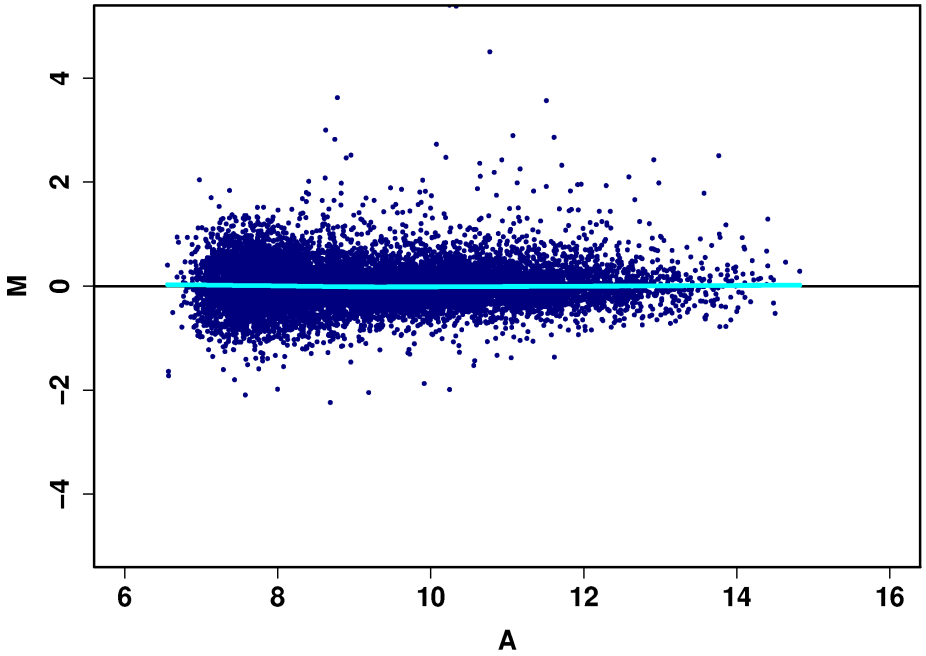


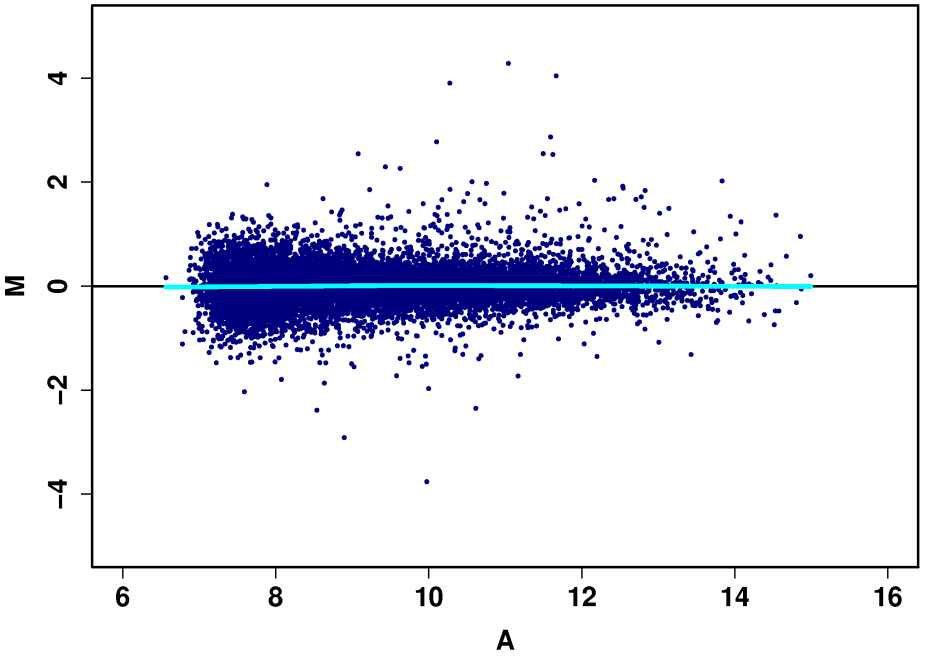


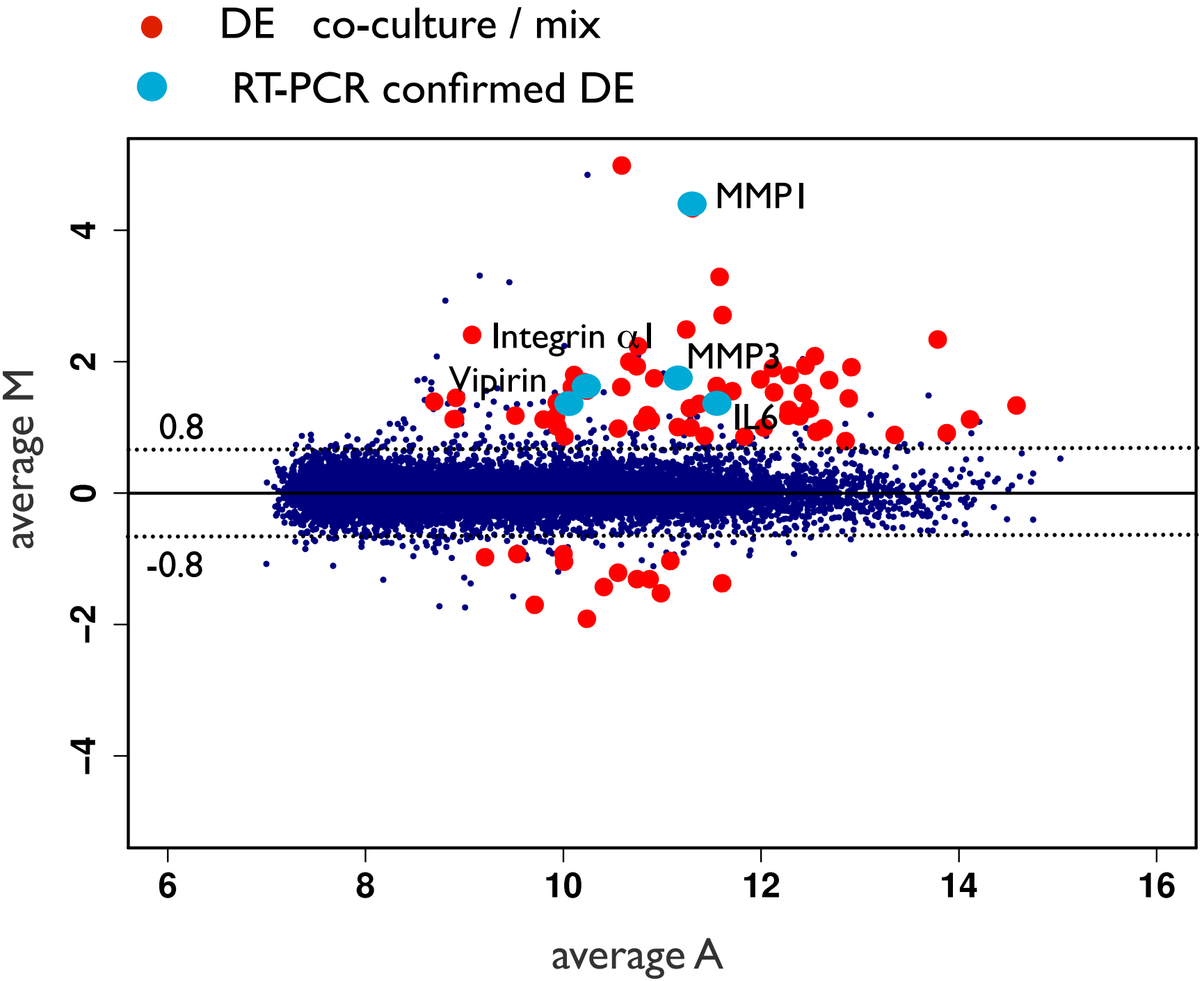


after
normalisation

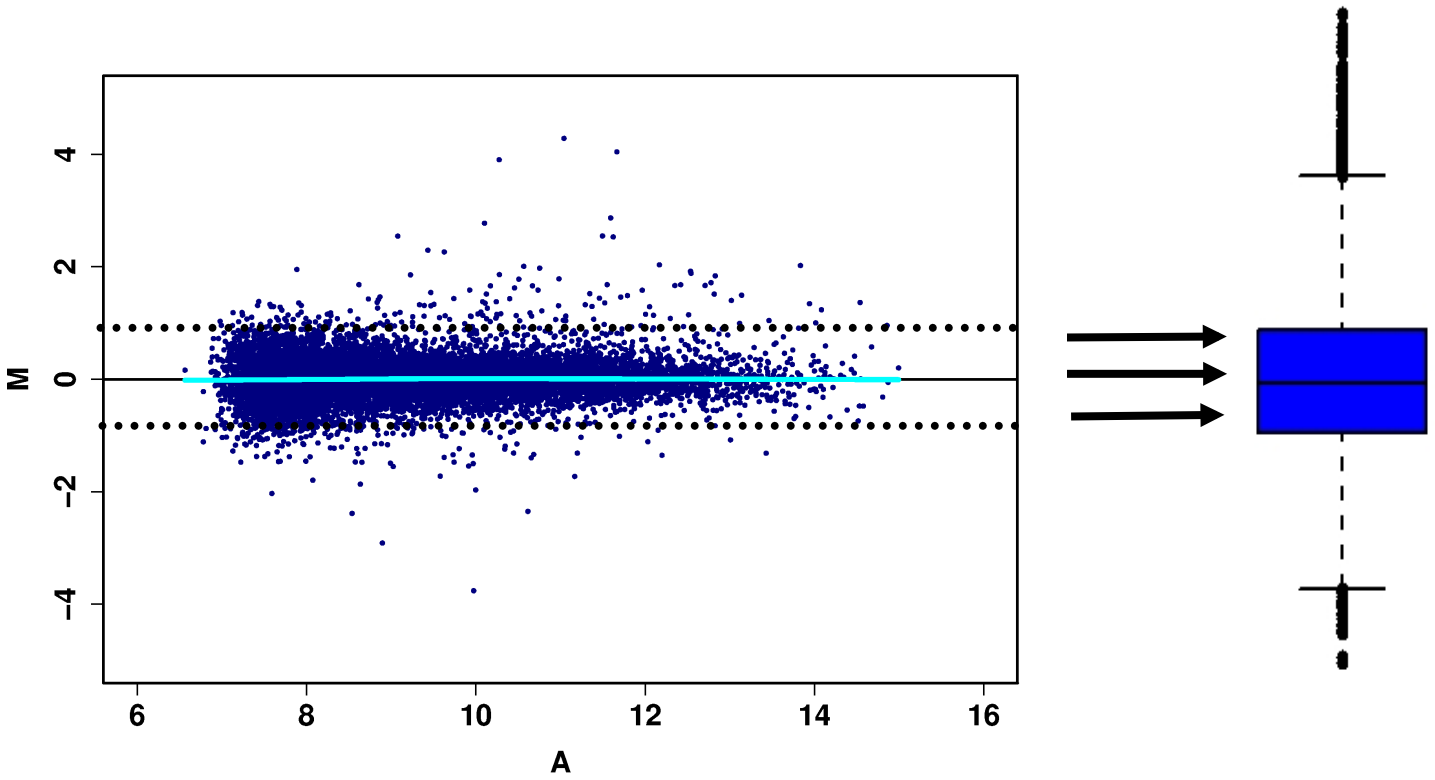






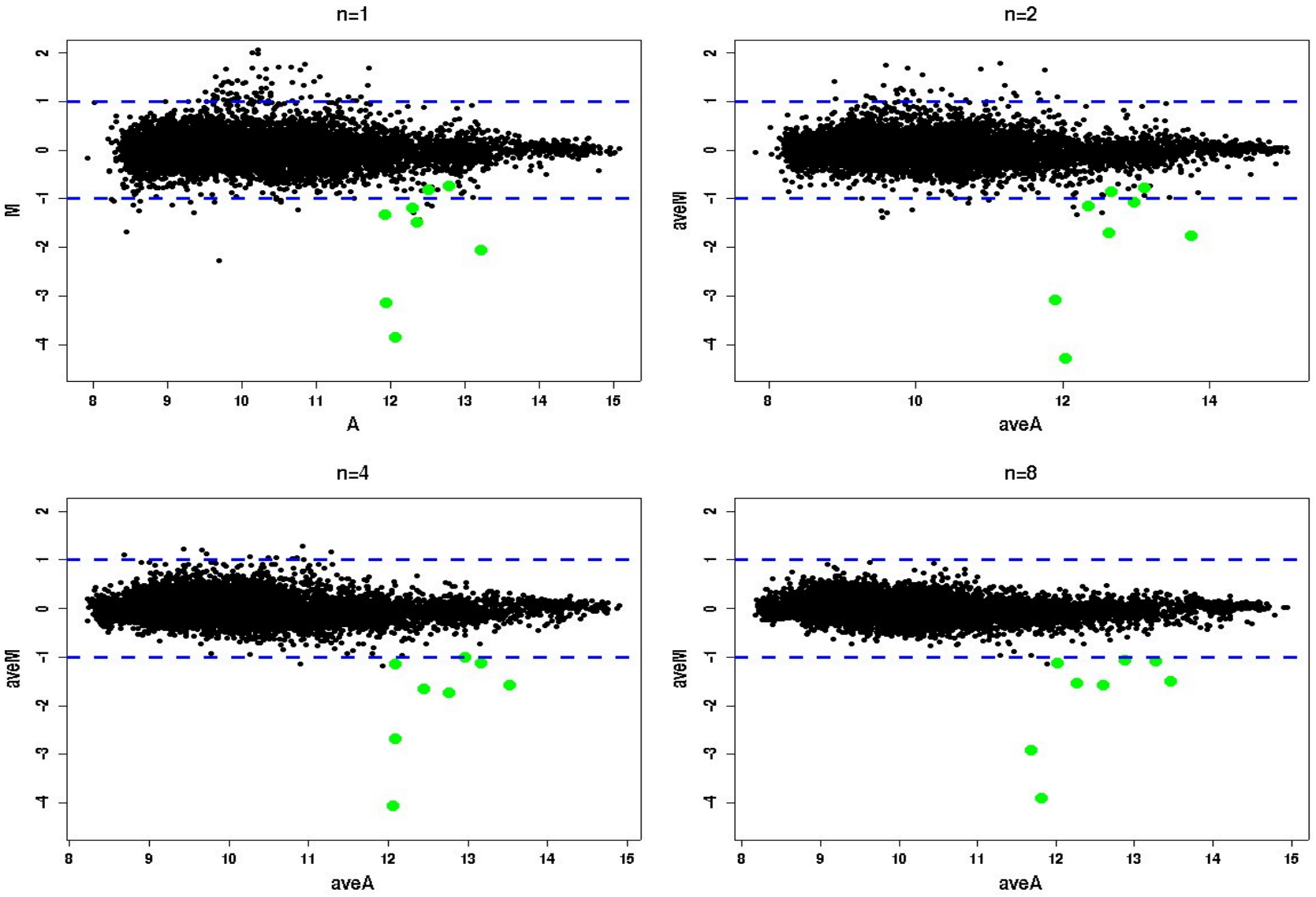

DE co-culture / mix
RT-PCR confirmed DE
MMP1
Integrin αI
MMP3
Vipirin
IL6
0.8
-0.8
average M
average A
4
2
0
−2
−4
6
8
10
12
14
16

Summarising Variability



Variability decreases with more replicates, averaging!

Sample size

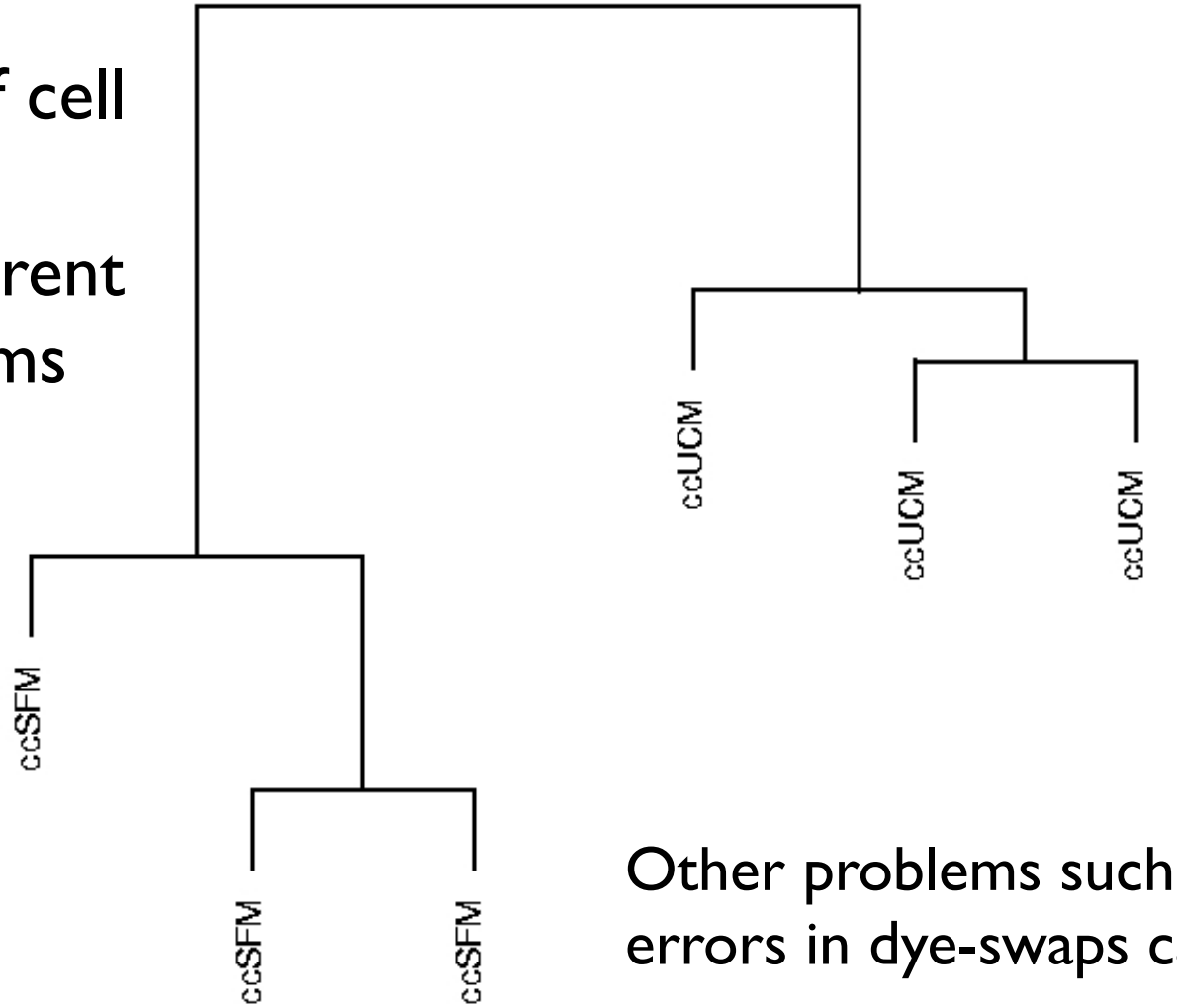


Data provided by Matt Callow

Further exploratory analysis of microarray data

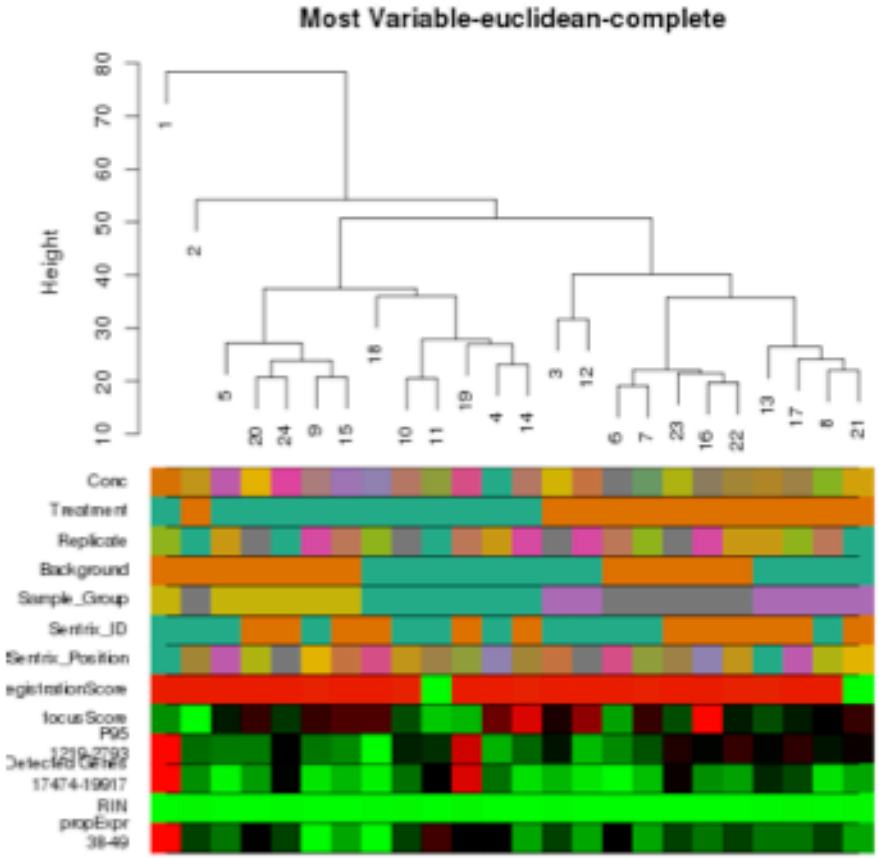
Discovering groups

Replicate arrays of cell lines accidentally grown in two different serum free mediums



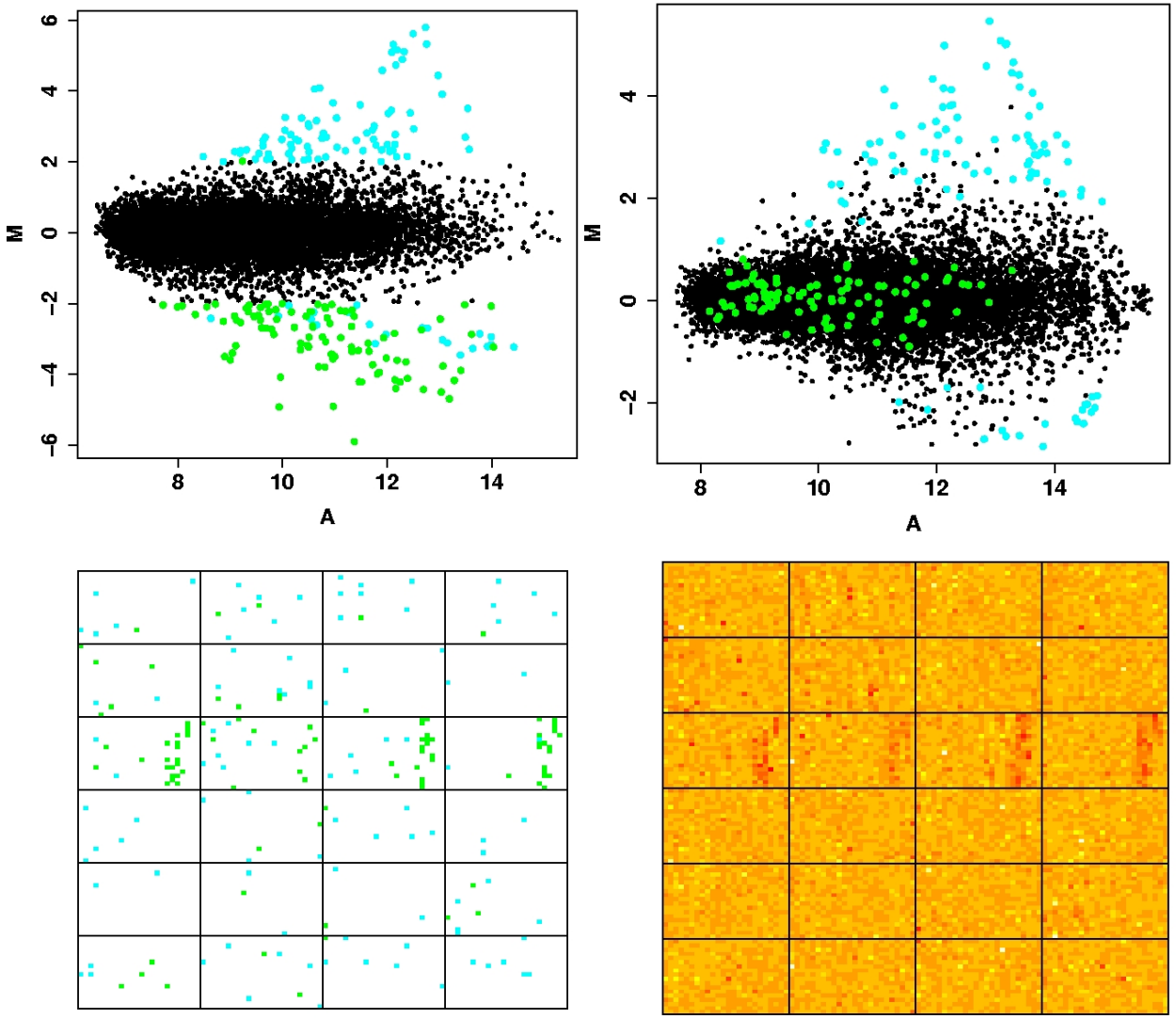
Other problems such as errors in dye-swaps can occur…

Using Clustering for QC



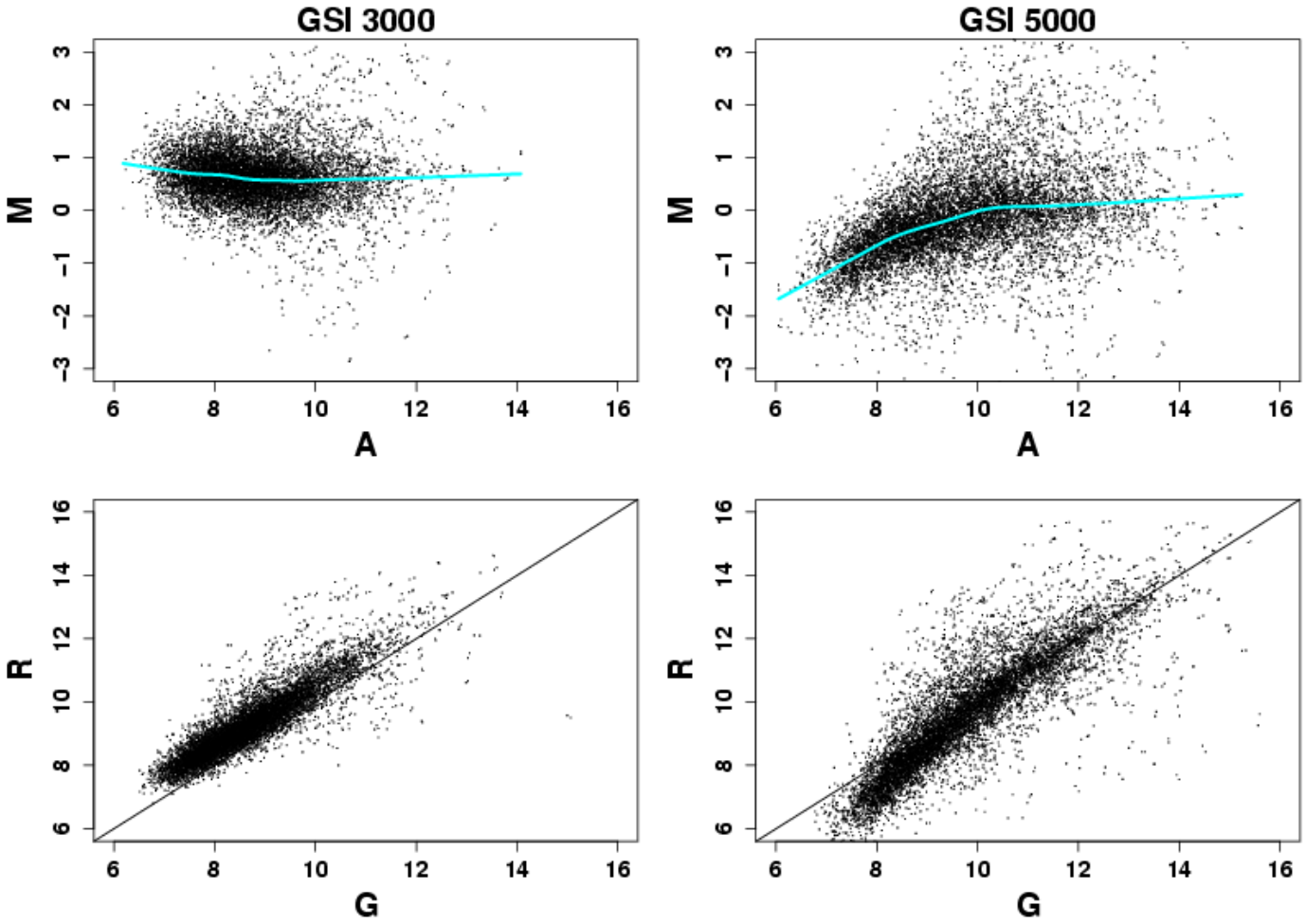

```
library(WGCNA)
d <- dist(t(eset), method = euclidean,])
dend <- hclust(d, method = complete)
plotDendroAndColors(dend, col_matrix,groupLabels=colnames(col_matrix))
```

Finding artifacts…



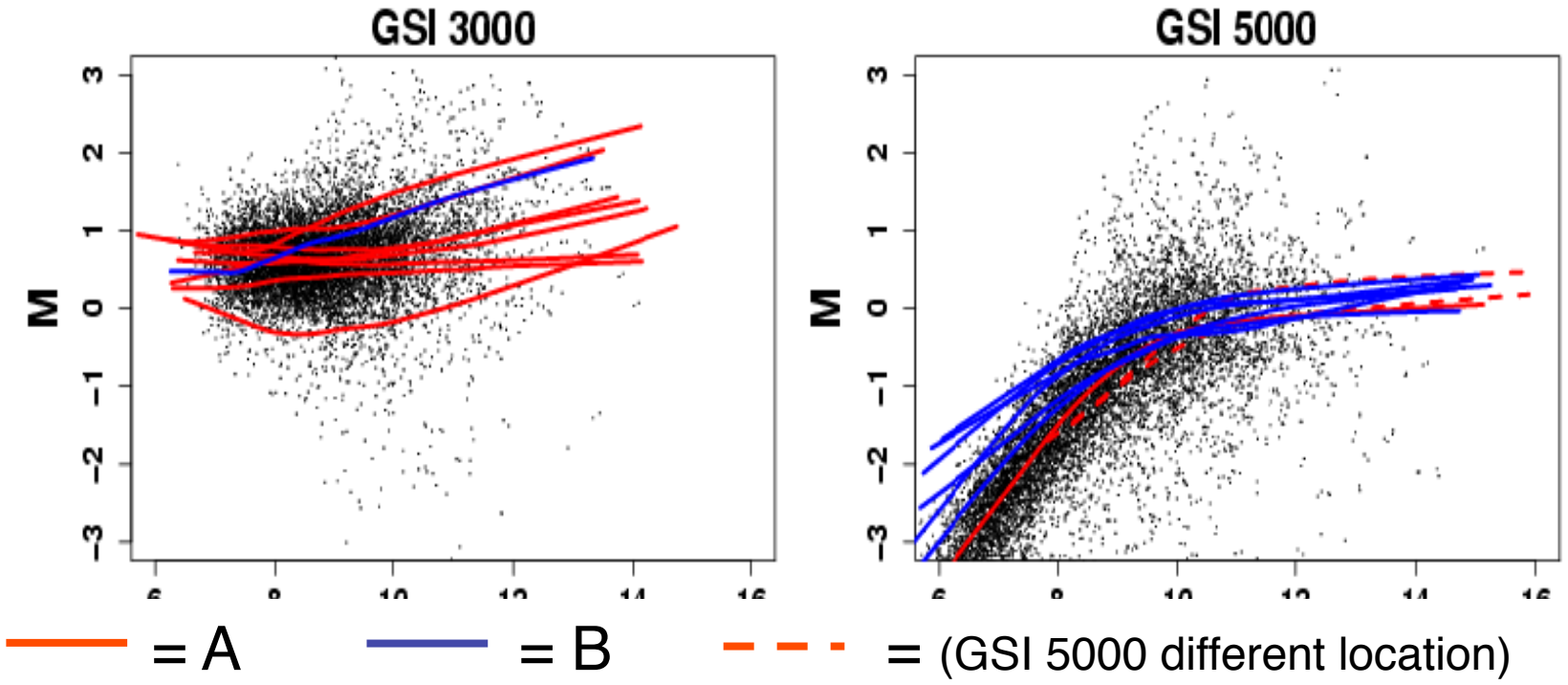
is particularly important when there is little replication!

Unique effects of different scanners -

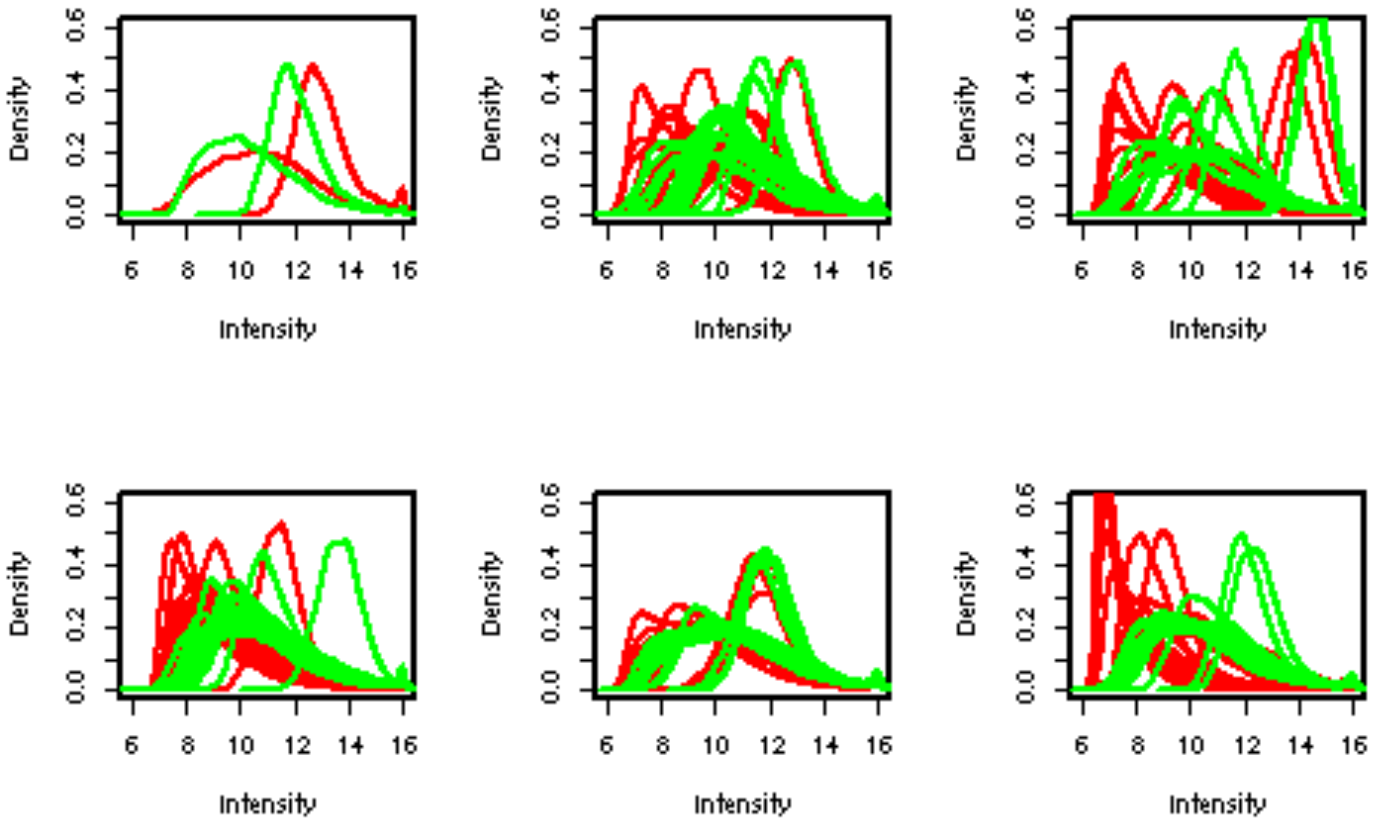


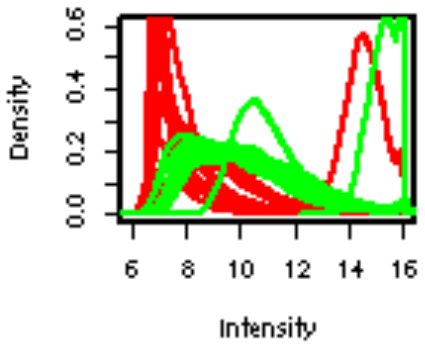
can dramatically affect the quality and features of your data!

And these systematic effects may confound your analysis …



… a simple cluster analysis on this data revealed excellent (but unknowingly biased) separation between groups A and B.





Batches of slides and the actual **day** of hybridisation (washing) can affect your data - **ozone levels** can interfere with the Cy5 dye. Almost every step can potentially introduce significant systematic variation.

Quantitative quality control
Array and spot weights

Material provided by Matthew Ritchie and Gordon Smyth

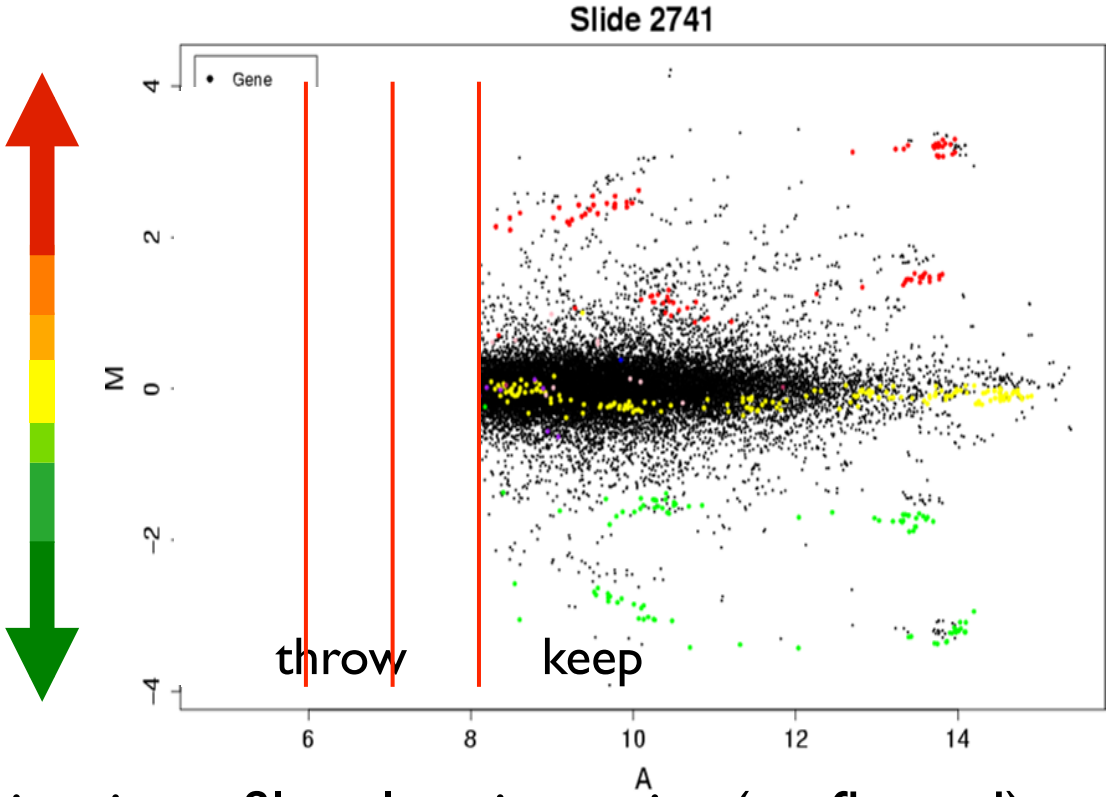
Options for quality control

- Do Nothing (except perhaps removing extreme bad arrays)
- Throw out bad arrays (often good idea as a place to begin)
- Filtering/Flagging bad spots

OR

- Incorporate spot and array quality weights

Filtering spots



Possible option is to filter low intensity (or flagged) spots, usually based on arbitrary cut-off, requires visual inspection and can be time-consuming. You may throw out good data and you can end up with a lot of missing data for some genes across an experiment. This can make downstream analysis (including the interpretation of results) quite difficult.

Array weights

Precision: reproducibility of a measurement.
i.e. Repeatability of log-ratios (M's) over arrays.

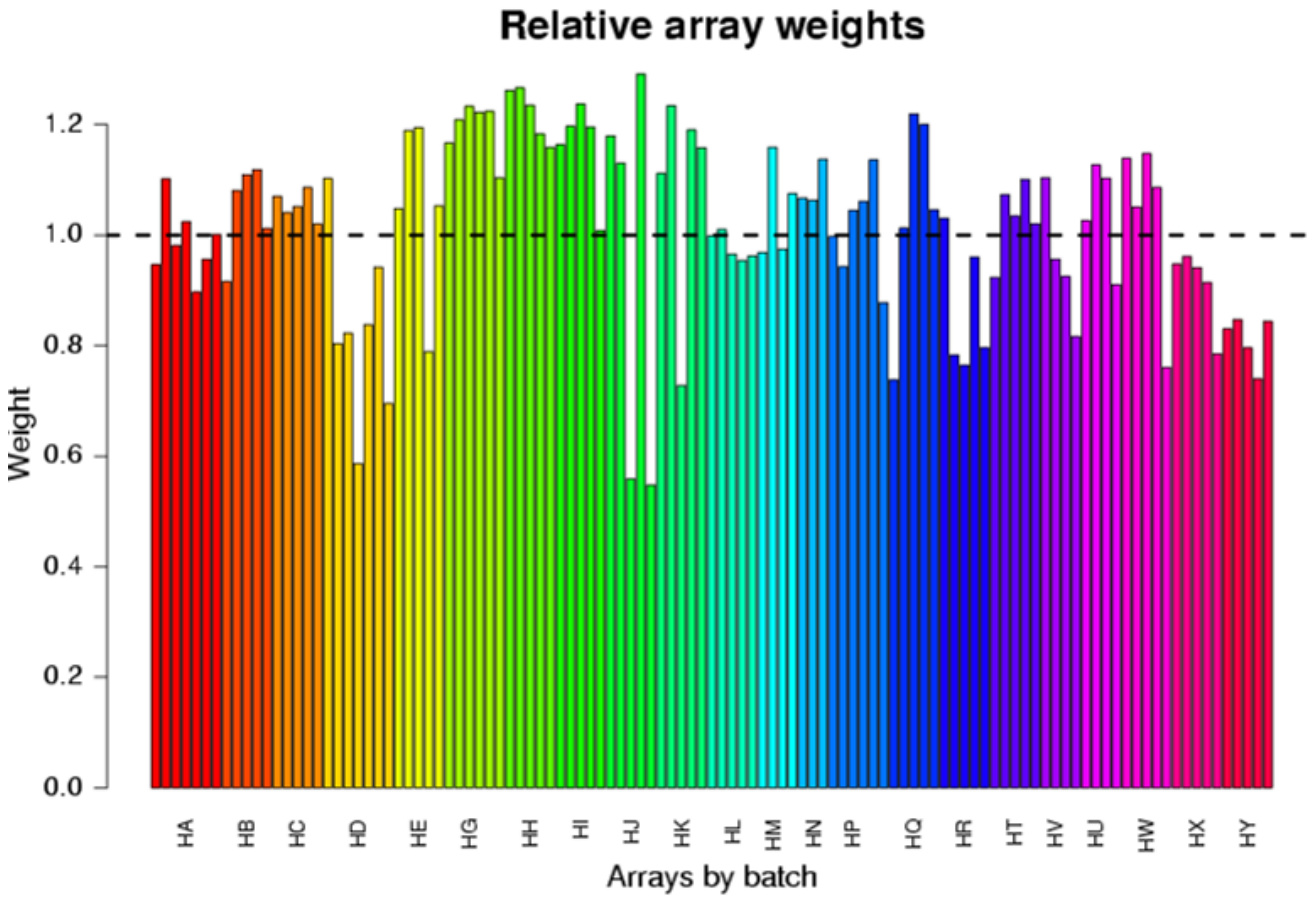
Low Precision — Medium — High Precision

Low Quality — High Quality

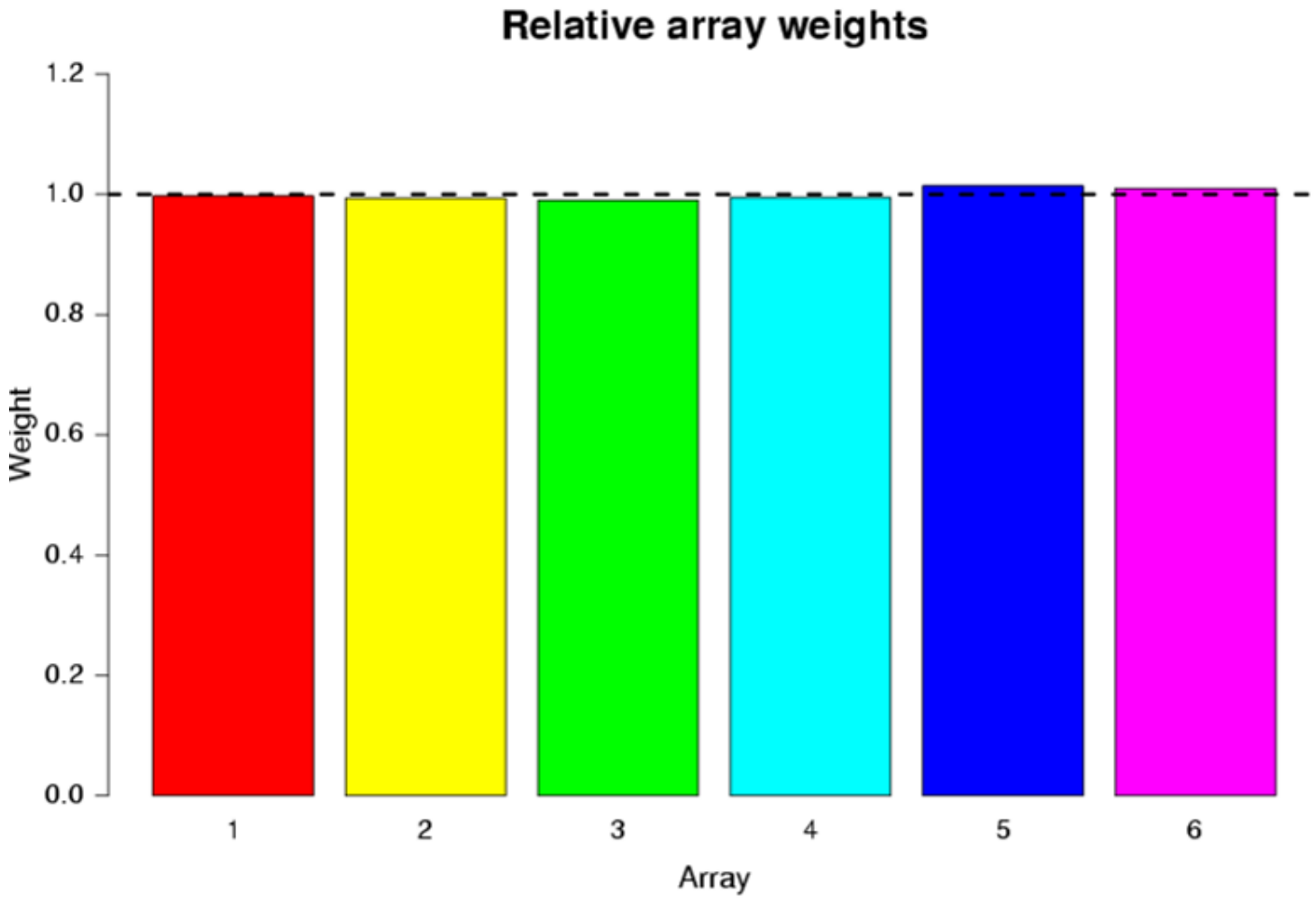
weight (w) = Precision

Uses graduated weighting, requires no cutoff.

Array weights for Quality Control dataset



Array weights for good data



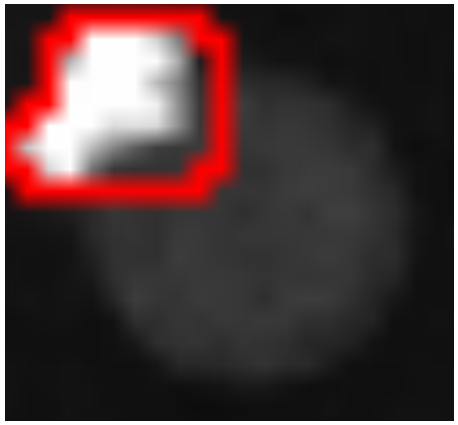
Data from Mireille Lahoud, Shortman Lab

Spot weights

Low Quality Medium High Quality

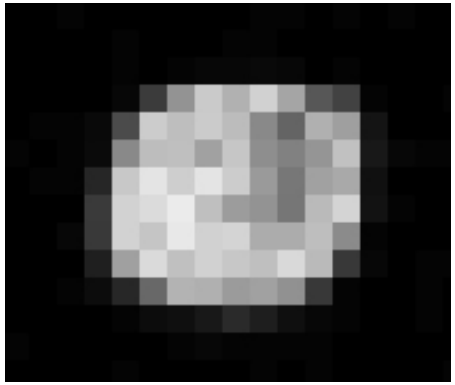
weight (w) reflects spot quality measure

Weighting spots



Area = 50 (165 ideal)

Down weight spot

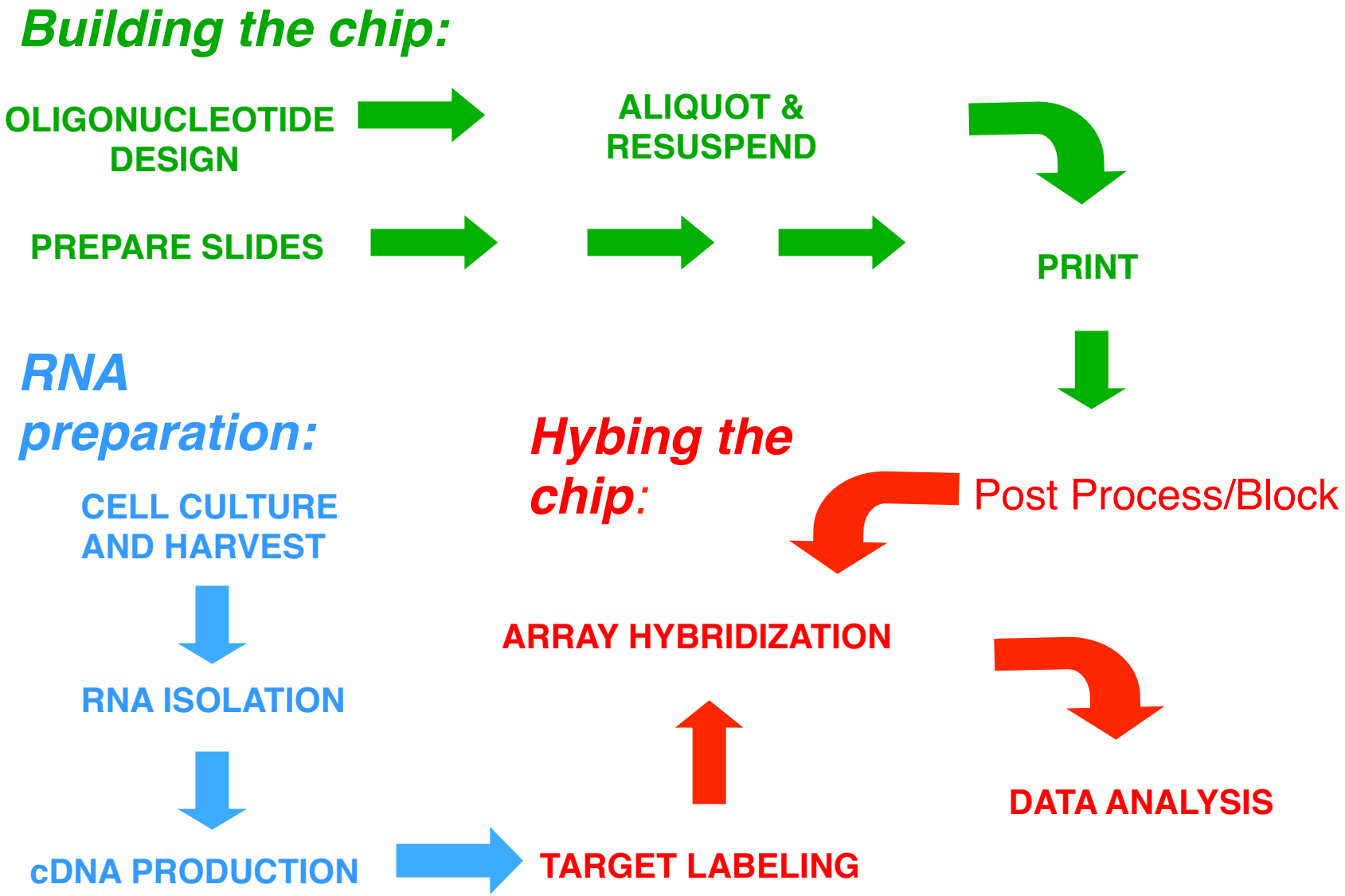


Area = 163 (165 ideal)

Spot given full weight

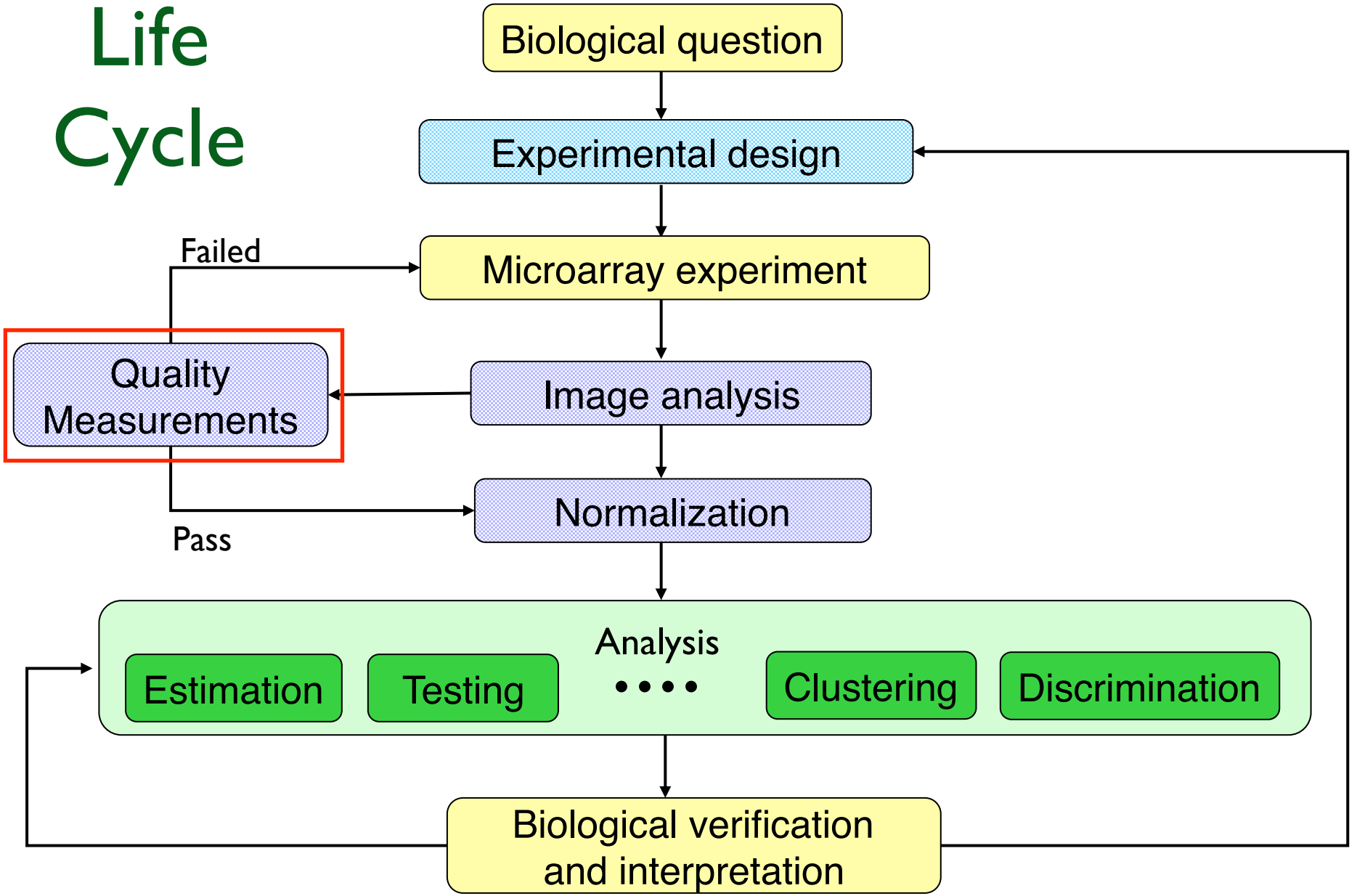
Spot weights are derived from an empirical trend observed in the data. The challenge is to find a measure from which to derive spot weights that is informative for the actual spot "quality".

The process



Taken from Schena & Davis

Life Cycle


Biological question
Experimental design
Failed
Microarray experiment
Quality Measurements
Image analysis
Normalization
Pass
Analysis
Estimation
Testing
• • • •
Clustering
Discrimination
Biological verification and interpretation