Generating quality metrics reports for microarray data sets

Audrey Kauffmann

Introduction

- Microarrays are widely/routinely used
- Technology and protocol improvements → trustworthy
- Variance and noise
  - Technical causes:
    - Platform
    - Lab, experimentalist
    - RNA extraction
    - Amplification, labeling, hybridization, scanning...
  - Biological causes:
    - Tissue itself (cell lines, biopsies, blood...)
    - Tissue contamination
    - Clinical covariates (age, sex, race...)
    - Cell cycle...

Who is concerned?

- Experimentalist investigating a set of samples
  - Choose between different technology platforms, expt. Protocols
  - Decide when to repeat (certain parts of) the experiment
- Statistical collaborator analysing the experiment
  - Decide whether to proceed or to ask the experimentalist to go back to the lab
- Microarray core facility
  - Decide whether to consider their product fit for delivery to customer
  - Customer decides whether to be content (pay the bill)
- Integrative biologist analysing data in a public database
  - Has to choose which experiments
  - Which arrays within an experiment to consider
- Public data(base) provider
  - Put a quality score on each of her offerings

At which step of the analysis?

- Importing the data
- Preprocessing: background correction, normalisation, summarization of probesets
- Differential Expression
- Gene set enrichment analysis

At which step of the analysis?

- Importing the data
- Quality Assessment
- Preprocessing: background correction, normalisation, summarization of probesets
- Quality Assessment
- Differential Expression
- Gene set enrichment analysis

At which step of the analysis?

- Importing the data
- Quality Assessment
- Preprocessing: background correction, normalisation, summarization of probesets
- Quality Assessment → Remove outlier(s)
- Differential Expression
- Gene set enrichment analysis

What aspects to be evaluated? Which quality metrics?

Per Slide

- What are we looking at?
  - Intensity-dependent ratio
  - Detection of spatial effects
- How?
  - MAplots
  - Representation of the chip

Between Slides

- What are we looking at?
  - Homogeneity
  - Outlier samples
  - Biological meaning
- How?
  - Boxplots, density plots
  - Heatmap, PCA

How to easily perform quality assessment?

- **arrayQualityMetrics -** Bioconductor package for Affymetrix, Agilent, Illumina, homemade arrays …
- From an R object ⇒HTML report
- Plots:
  - MA plot and spatial representations
  - Boxplots and density
  - Heatmap and PCA
  - Variance-mean dependency
  - GC content and probe mapping studies
  - Affymetrix only: NUSE, RLE, RNA degradation, QCstats, PM/MM
- Outlier identification

Hands-on

- Run reports

arrayQualityMetrics report – *outlier detection*

QMreport.html - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///home/audrey/tmp/testaqm/EMEXP886raw/QMreport.html

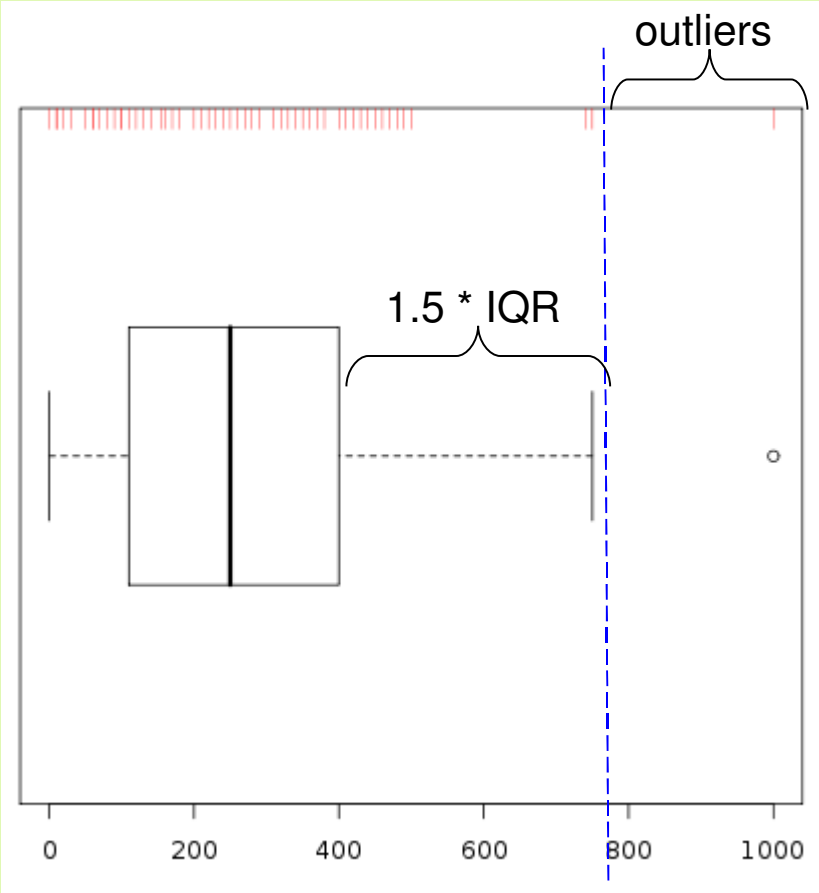
EMEXP886 quality metrics report

Summary

Array #	Array Name	MA plots	Spatial distribution	Boxplots	Heatmap	RLE	NUSE
1	E-MEXP-886-raw-cel-1227302827.cel						*
2	E-MEXP-886-raw-cel-1227302800.cel						
3	E-MEXP-886-raw-cel-1227302845.cel		*				
4	E-MEXP-886-raw-cel-1227302836.cel						
5	E-MEXP-886-raw-cel-1227302791.cel						
6	E-MEXP-886-raw-cel-1227302809.cel	*		*	*		
7	E-MEXP-886-raw-cel-1227302867.cel						
8	E-MEXP-886-raw-cel-1227302889.cel	*		*			
9	E-MEXP-886-raw-cel-1227302818.cel		*				
10	E-MEXP-886-raw-cel-1227302782.cel						

*outlier array

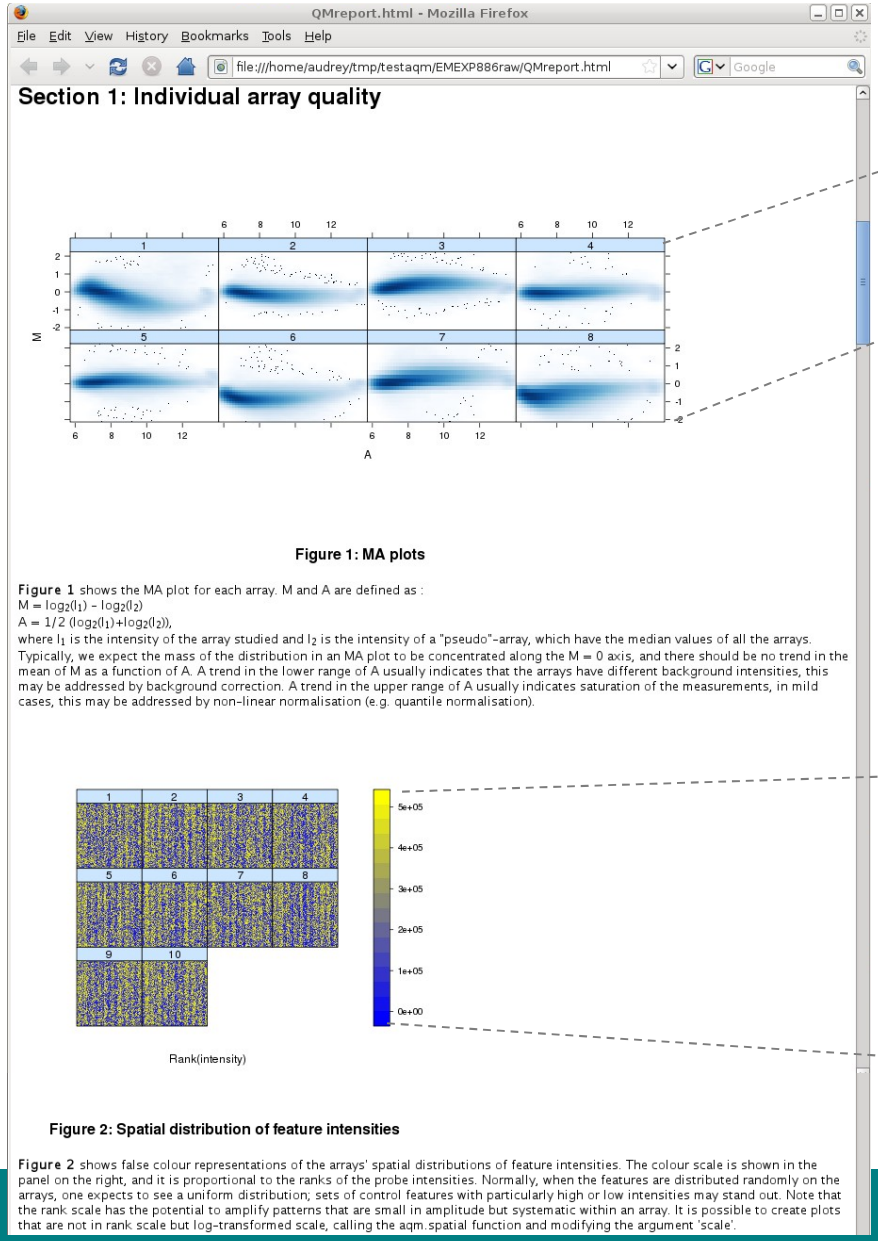
Index



Boxplot of the scores

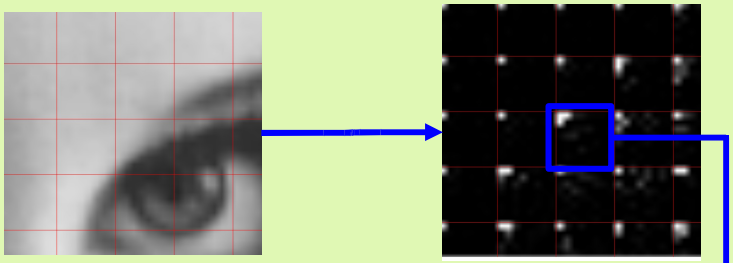
S

arrayQualityMetrics report – *per array*



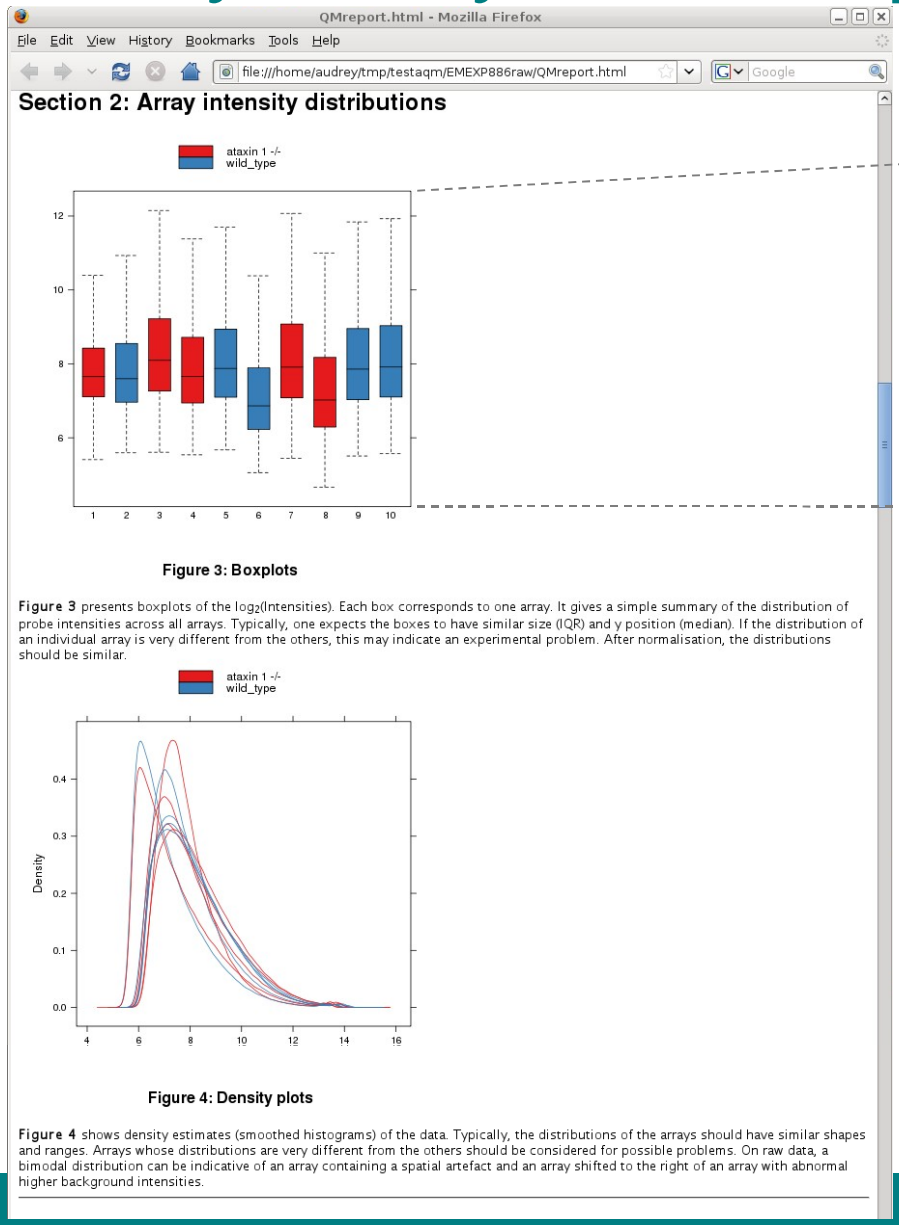
$$S_i = |_{ik}$$

Fourier Transform



$$S_i = \frac{\sum lowFreq_{ik}}{\sum highFreq_{ik}}$$

arrayQualityMetrics report – *intensity distributions*



$$S_i = median\left(I_{ik}\right)$$

$$S_i = IQR\left(I_{ik}\right)$$

arrayQualityMetrics report – *Between arrays*

QMreport.html - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///home/audrey/tmp/testaqm/EMEXP886raw/QMreport.html

Section 3: Between array comparison

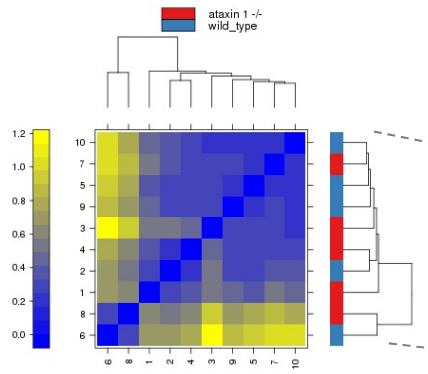


Figure 5: Heatmap representation of the distance between arrays

Figure 5 shows a false colour heatmap of between arrays distances, computed as the mean absolute difference (L_1-distance) of the vector of M-values for each pair of arrays on every probes without any filtering. The colour scale is chosen to cover the range of L_1-distances encountered in the dataset. Arrays for which the sum of the distances to the others is much different from the others, are detected as outlier arrays. The dendrogram on this plot also can serve to check if, the arrays cluster accordingly to a biological meaning.
$d_{xy} = mean|M_{xi}-M_{yi}|$
Here, M_{xi} is the M-value of the i-th probe on the x-th array, without preprocessing. Consider the following decomposition of M_{xi}: $M_{xi} = z_i + \beta_{xi} + \epsilon_{xi}$, where z_i is the probe effect for probe i (the same across all arrays), ϵ_{xi} are i.i.d. random variables with mean zero and β_{xi} is such that for any array x, the majority of values β_{xi} are negligibly small (i. e. close to zero). β_{xi} represents differential expression effects. In this model, all values d_{xy} are (in expectation) the same, namely 2 times the standard deviation of ϵ_{xi}.

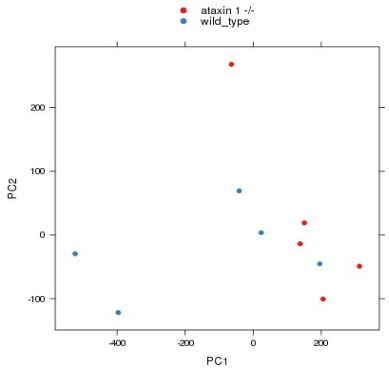


Figure 6: Principal Component Analysis

Figure 6 represents a biplot for the first two principal components from the dataset. The colours correspond to the group of interest given. We expect the arrays to cluster accordingly to a relevant experimental factor. The principal components transformation of a data matrix re-expresses the features using linear combination of the original variables. The first principal component is the linear combination chosen to possess maximal variance, the second is the linear combination orthogonal to the first possessing maximal variance among all orthogonal combination.

For each couple of arrays i and j, k is a probe and the distance between the arrays is:

$$d_{ij} = mean_k\left(\left|I_{ik} - I_{jk}\right|\right)$$

$$S_i = \sum d_{ij}$$

QMreport.html - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///home/audrey/tmp/testaqm/EMEXP886raw/QMreport.html

Section 4: Variance mean dependence

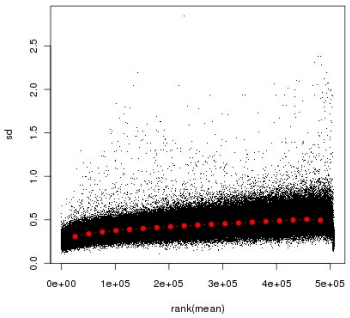


Figure 7: Standard deviation versus rank of the mean

For each feature, **Figure 7** shows the standard deviation of the intensities across arrays on the y-axis versus the rank of their mean on the x-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

arrayQualityMetrics report – *Affymetrix plots*

QMreport.html - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///home/audrey/tmp/testaqm/EMEXP886raw/QMreport.html

Section 5: Affymetrix specific plots

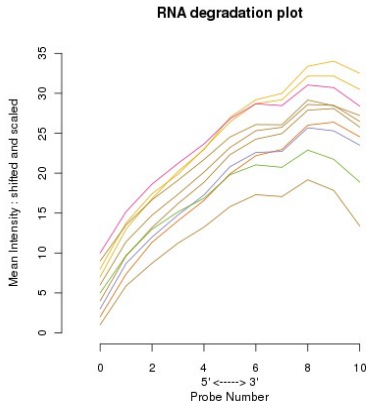


Figure 8: RNA degradation plot

In **Figure 8** , a RNA digestion plot is computed on normalised data (so that standard deviation is equal to 1). In this plot each array is represented by a single line. It is important to identify any array(s) that has a slope which is very different from the others. The indication is that the RNA used for that array has potentially been handled quite differently from the other arrays. This diagnostic plot is based on tools provided in the **affy** package.

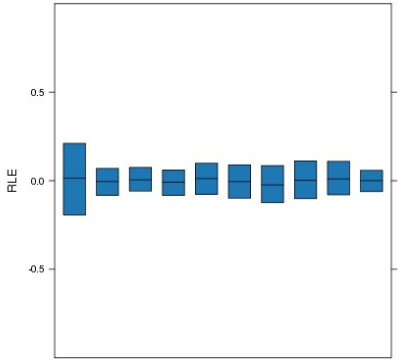


Figure 9: Relative Log Expression plot

Figure 9 is a Relative Log Expression (RLE) plot. RLE are performed on preprocessed data (background correction and quantile normalisation). An array that has problems will either have larger spread, or will not be centred at M = 0, or both. This diagnostic plot is based on tools provided in the **affyPLM** package.

QMreport.html - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///home/audrey/tmp/testaqm/EMEXP886raw/QMreport.html

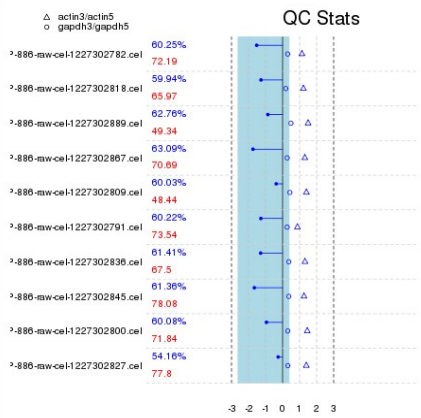


Figure 11: Diagnostic plot recommended by Affymetrix

Figure 11 represents the diagnostic plot recommended by Affymetrix. It is fully described in the package **simpleaffy**'s vignette. Any metrics that is shown in red is out of the manufacturer's specific boundaries and suggests a potential problem, any metrics shown in blue is fine.

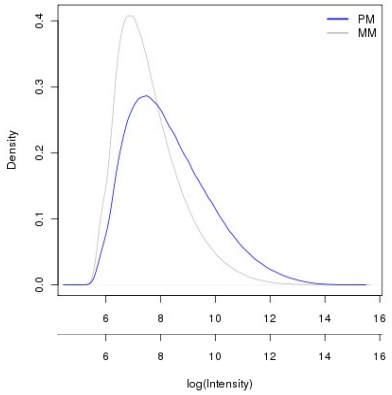
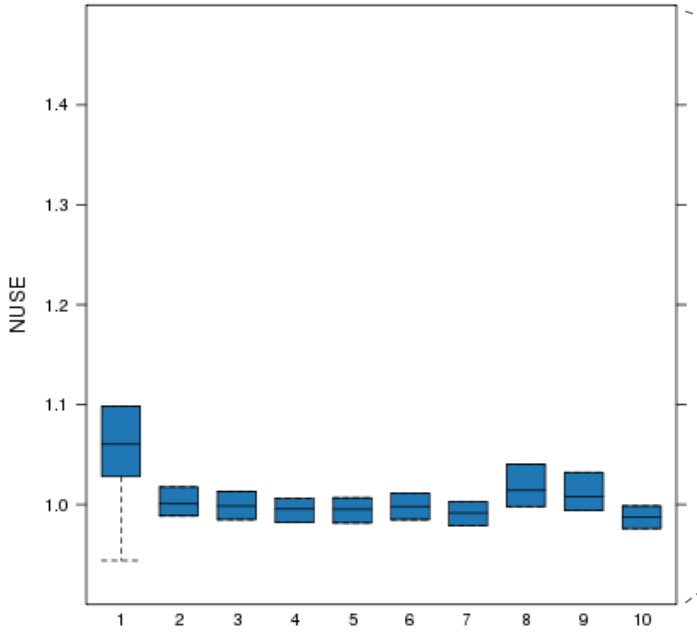


Figure 12: Perfect matches and mismatches

Figure 12 shows the density distributions of the $\log_2$ intensities grouped by the matching of the probes. Blue, density estimate of intensities of perfect match probes (PM) and grey the mismatch probes (MM). We expect that, MM probes having poorer hybridization than PM probes, the PM curve should be shifted to the right of the MM curve.

This report has been created with arrayQualityMetrics 2.0.23 under R version 2.10.0 Under development (unstable) (2009-03-31 r48256)

arrayQualityMetrics report – *Affymetrix NUSE: Normalised Unscaled Standard Error*



$$S_i = median\left(I_{ik}\right)$$
$$S_i = IQR\left(I_{ik}\right)$$

$$NUSE(\beta_{kj}) = \frac{SE(\beta_{kj})}{med_j(SE(\beta_{kj}))}$$

- Fitting a probe level model (gene *k*, array *j*)
- Differences in variability between genes. An array with elevated SE (standard error) relative to the other arrays is of lower quality

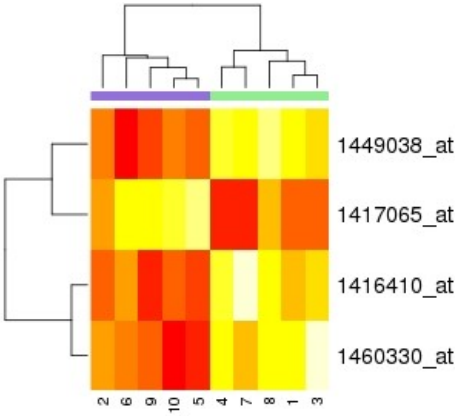
Why is outlier detection important? Example

- ArrayExpress experiment E-MEXP-886, cerebellar gene expression:
  - 5 WT mice (15 weeks of age)
  - 5 Atnx1 KO mice (15 weeks of age)
- Affymetrix MOE-430A (mouse) Genechip
- Ataxin 1 (Atxn1): protein of unknown function associated with cerebellar neurodegeneration in SpinoCerebellar Ataxia type 1 (SCA1), which impairs the of eye movement

Results from E-MEXP-886 analysis

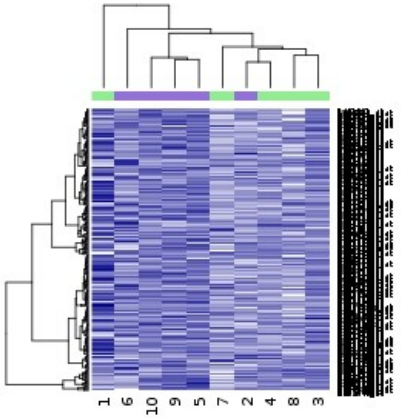
- Moderated t-test

	# Genes	
	P < 0.01	P < 0.001
10 samples	34	4



- Most enriched KEGG Pathway

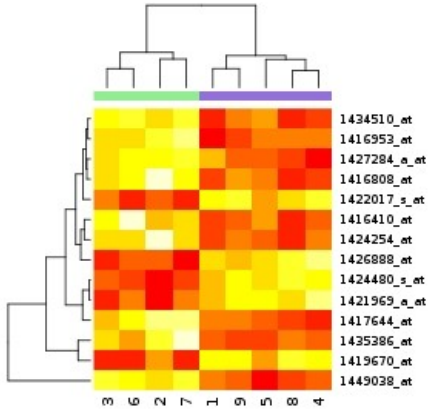
	Neuroactive ligand-receptor interaction	
	# Significant genes	Corrected t-value
10 samples	4	-5.65



Outlier array's impact on results

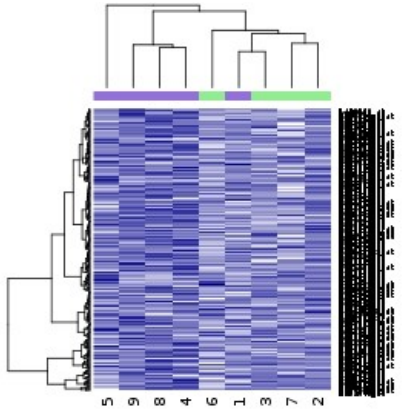
- Moderated t-test (limma)

	# Genes	
	P < 0.01	P < 0.001
10 samples	34	4
Without array 1	190	14



- Most enriched KEGG Pathway

	Neuroactive ligand-receptor interaction	
	# Significant genes	Corrected t-value
10 samples	4	-5.65
Without array 1	23	-11.53



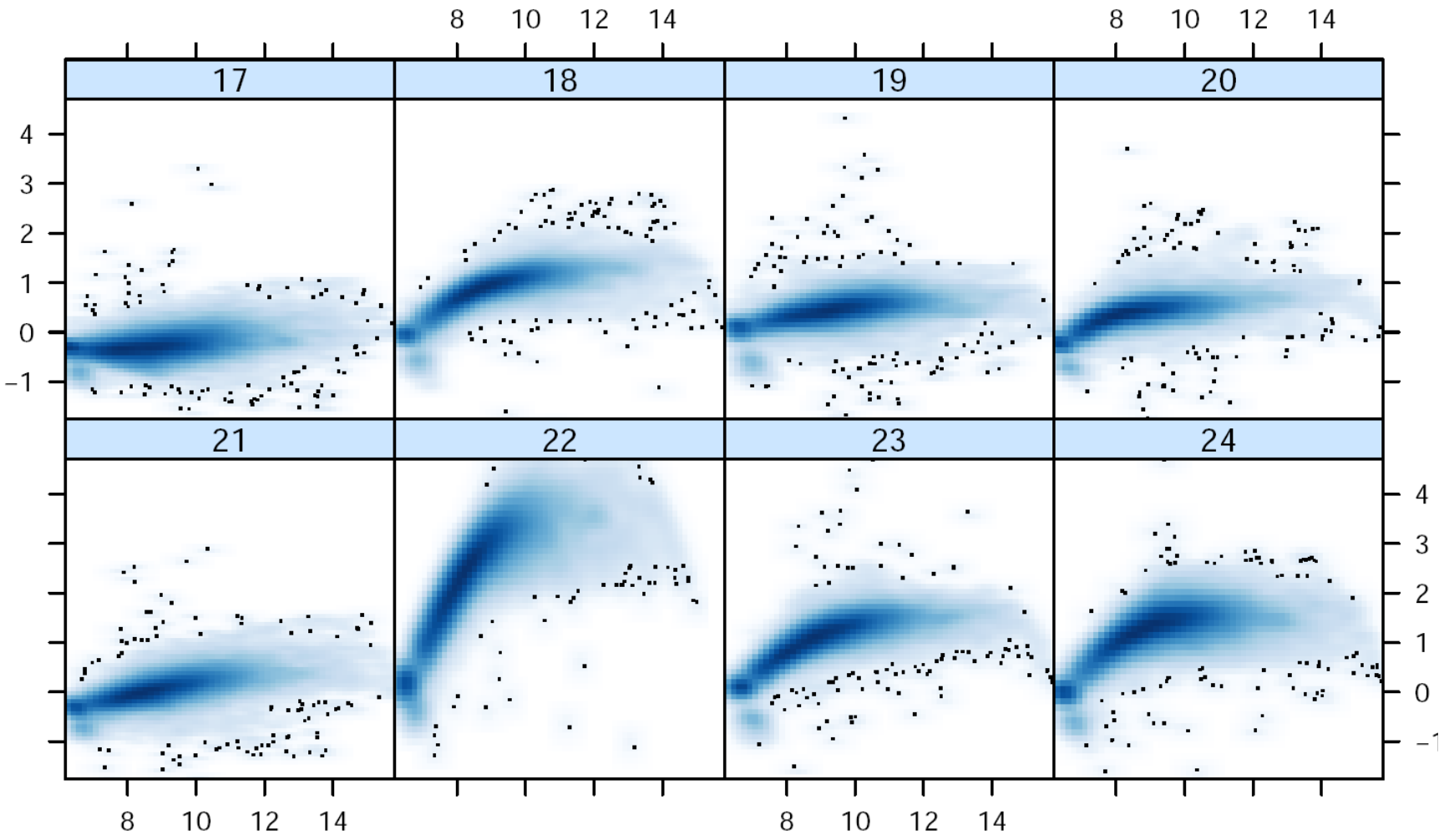
Validation of outlier's array specificity

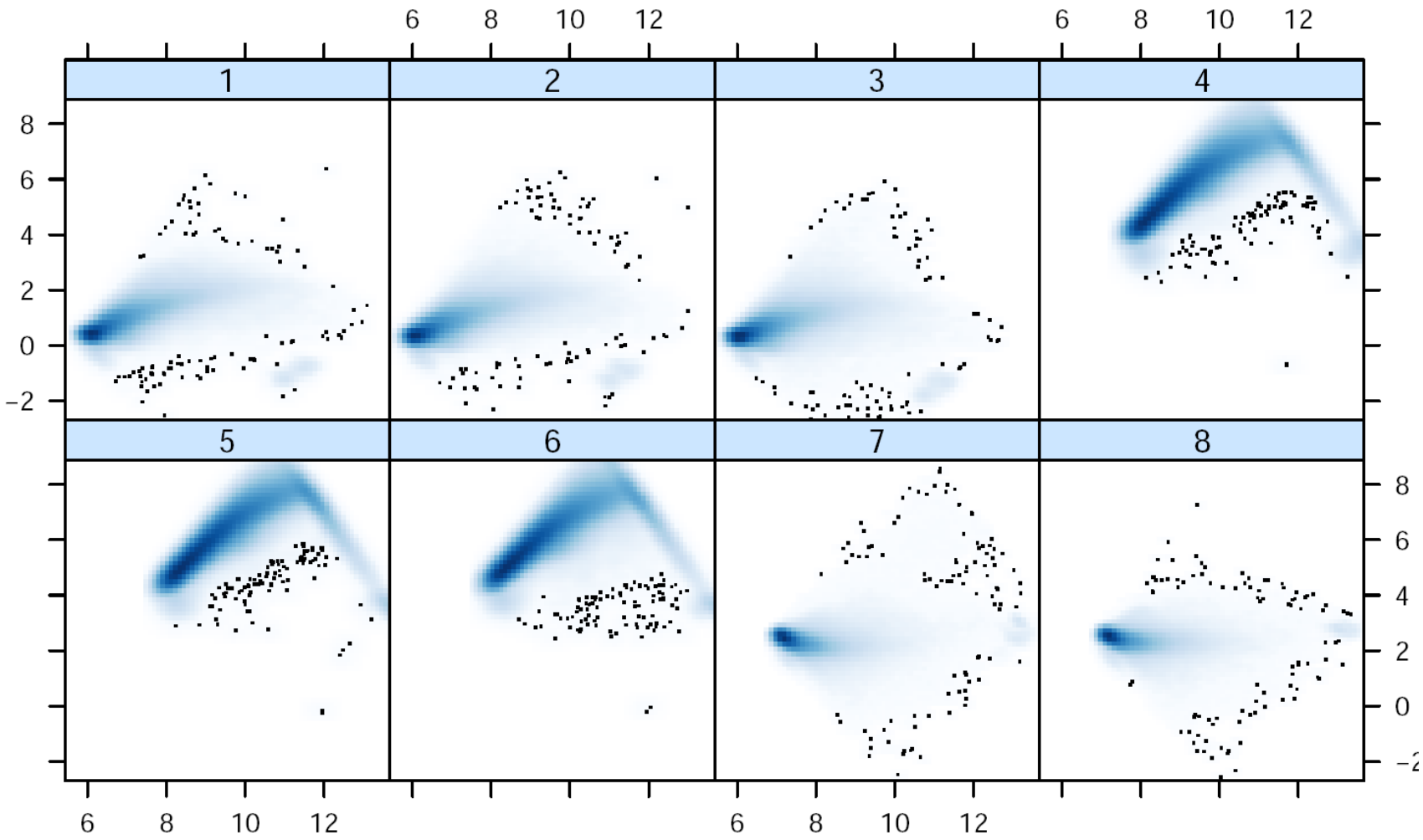
- Array #1 specific effect?

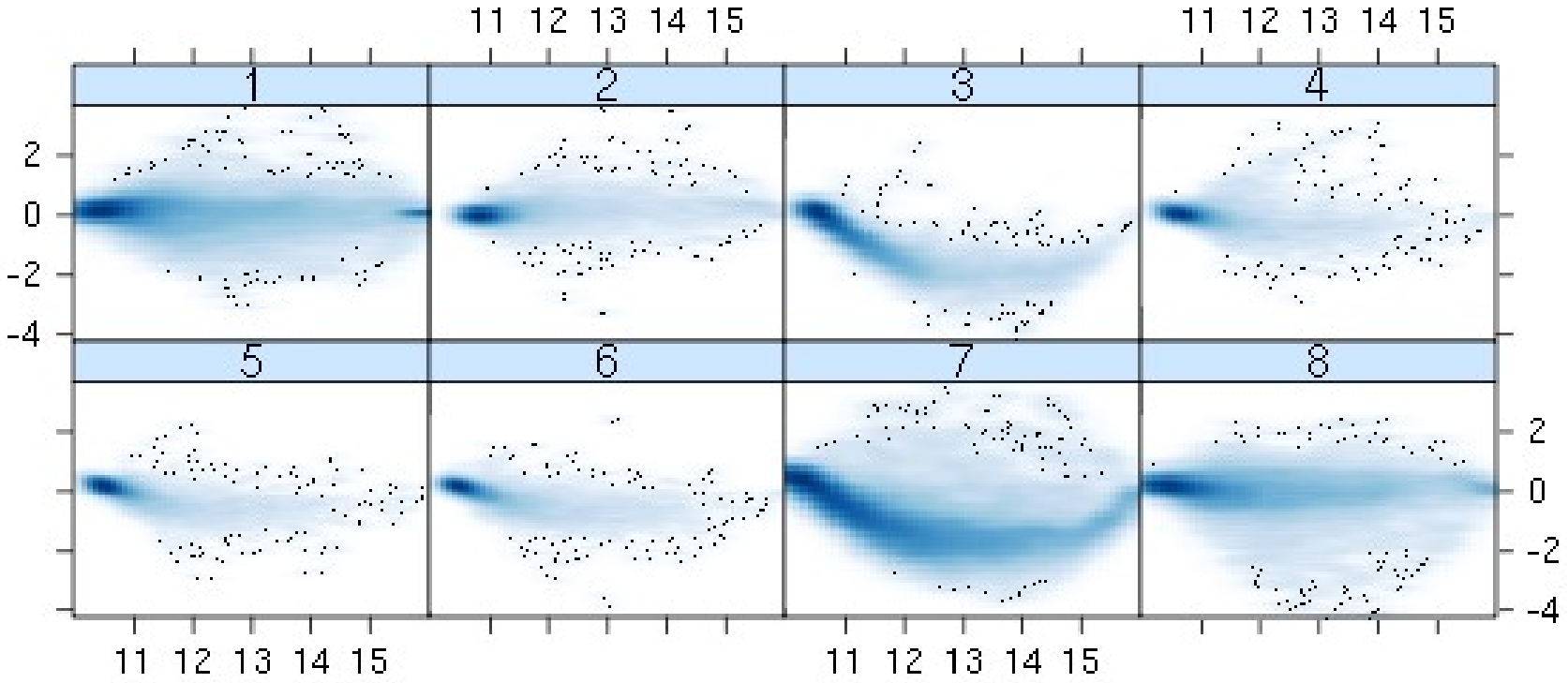
⇒ Remove one by one each array and perform the moderated t-test

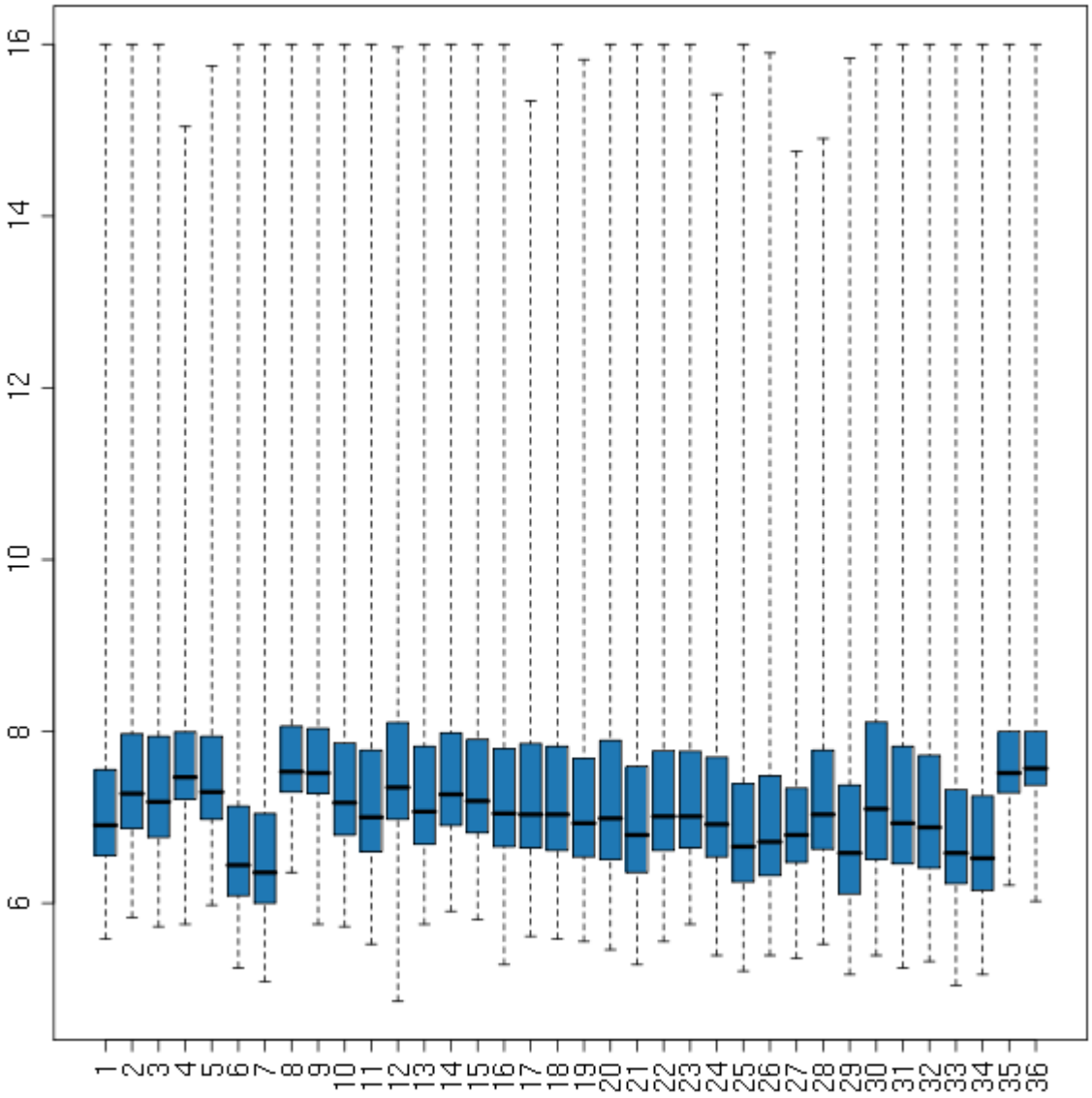
	# Genes	
	P < 0.01	P < 0.001
10 samples	34	4
Without sample 1	**190**	**14**
Without sample 2	39	3
Without sample 3	29	2
Without sample 4	21	1
Without sample 5	12	1
Without sample 6	87	5
Without sample 7	23	4
Without sample 8	34	4
Without sample 9	17	2
Without sample 10	23	2

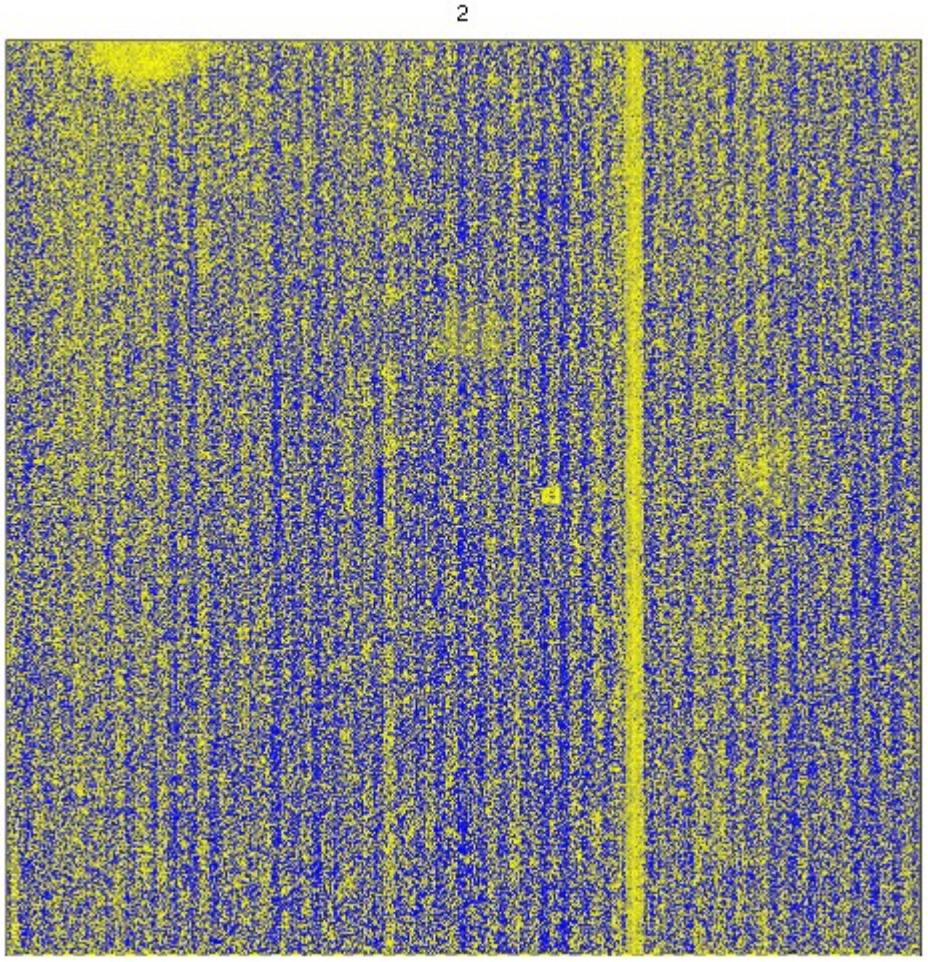
Horror Picture Show

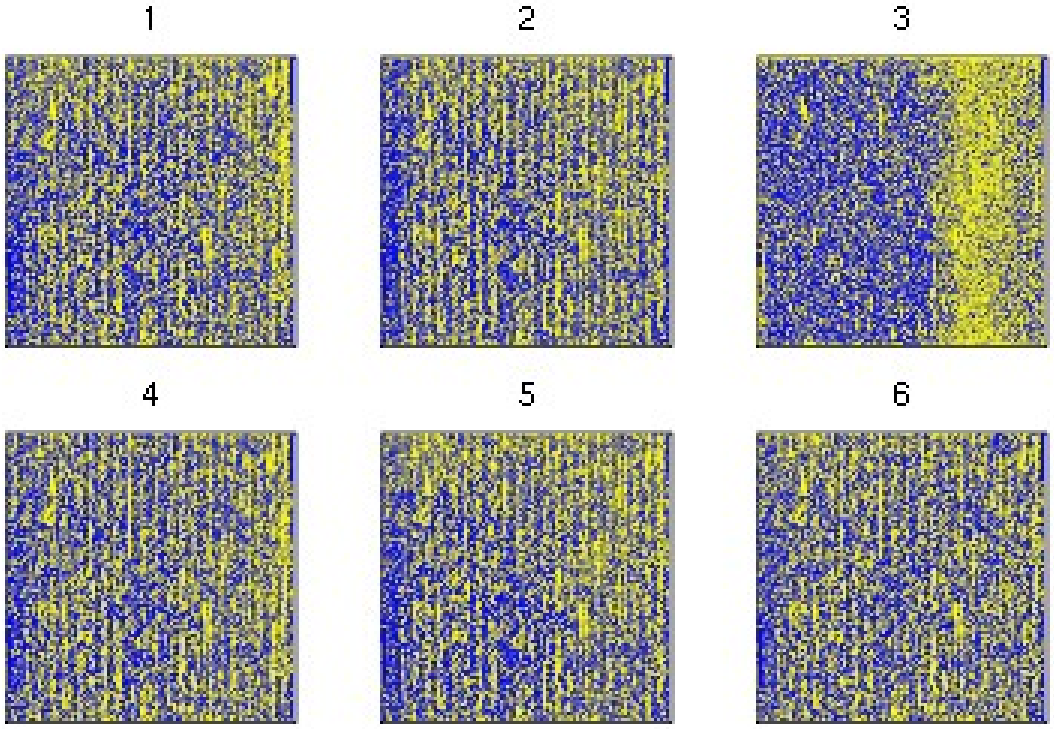


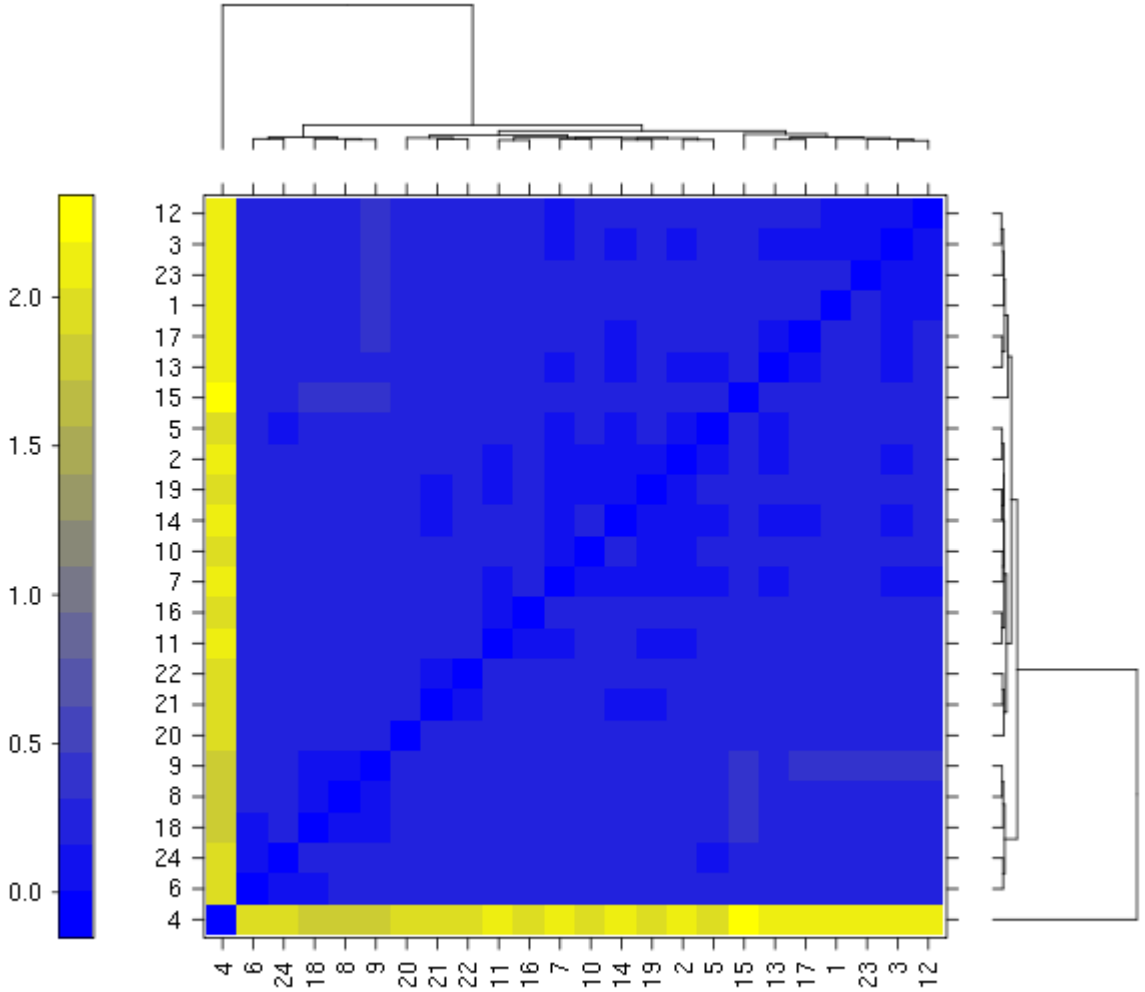

1
2
3
4
5
6
7
8
6
8
10
12
8
6
4
2
0
−2

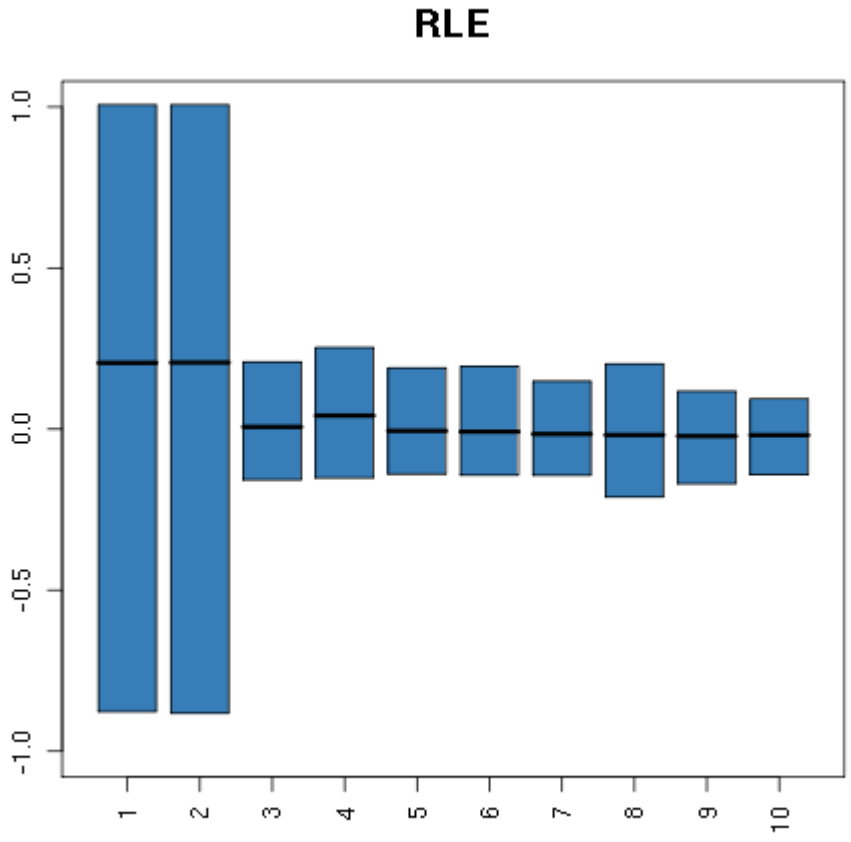

11 12 13 14 15
11 12 13 14 15
1
2
3
4
2
0
-2
-4
5
6
7
8
2
0
-2
-4
11 12 13 14 15
11 12 13 14 15

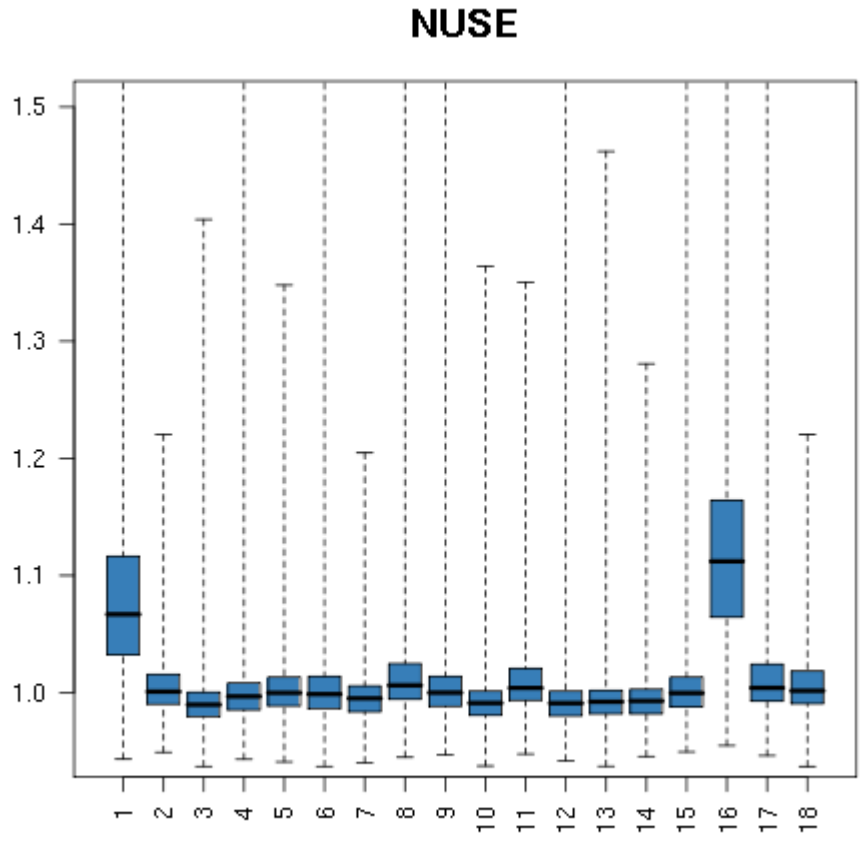

16
14
12
10
8
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36


2


1
2
3
4
5
6


2.0
1.5
1.0
0.5
0.0
12 3 23 1 17 13 15 5 2 19 14 10 7 16 11 22 21 20 9 8 18 24 6 4
4 6 24 18 8 9 20 21 22 11 16 7 10 14 19 2 5 15 13 17 1 23 3 12


RLE
1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
6
7
8
9
10


NUSE
1.5
1.4
1.3
1.2
1.1
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Conclusions

- Quality assessment is important
  - Still needed
  - Give a first “taste” of the data
  - By removing outlier, the statistical power is increased

- arrayQualityMetrics package
  - One command line
  - Before preprocessing: to decide which normalisation
  - After normalisation: to check normalisation efficiency
  - Comprehensive report
  - Outlier detection