

Array Quality Metrics

Audrey Kauffmann

Microarray data quality into question

- 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...

At which step of the analysis?

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

At which step of the analysis?

- Importing the data
- Quality Assessment
- Preprocessing: background correction, normalisation, summarisation 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, summarisation 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?

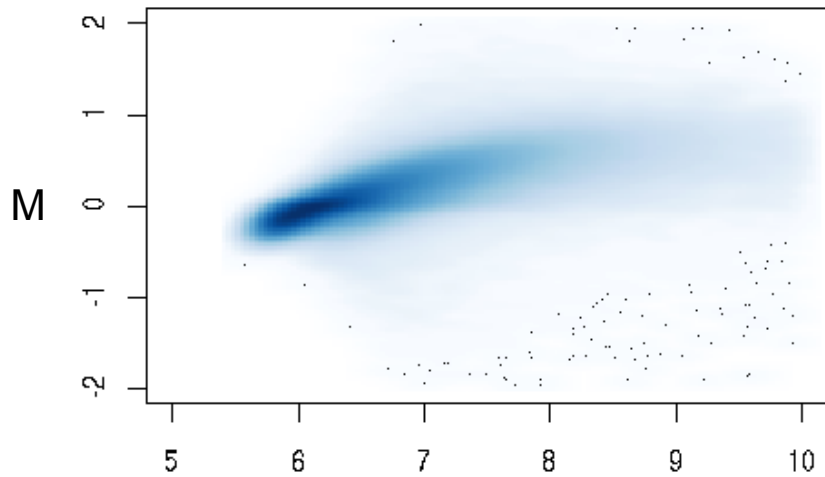
- **arrayQualityMetrics**, Bioconductor package for:
 - Affymetrix, Agilent, Illumina, homemade arrays etc...
- From an R object \Rightarrow HTML report
- **Plots:**
 - MA plot and spatial representations
 - Boxplots and density
 - Heatmap and PCA
 - Variance-mean dependency
 - GC content and probe mapping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

Functions

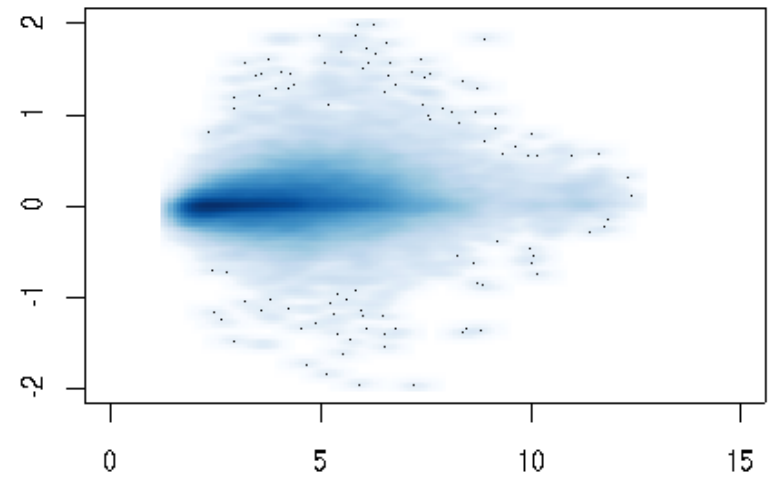
- `arrayQualityMetrics(expressionset, outdir, force, do.logtransform, intgroup, groupprep, spatial, sN)`
- aQM « modules »:
 - `aqm.prepdata()`
 - `aqm.maplot()`
 - `aqm.density()`
 - ...
 - `aqm.plot()`
 - `aqm.writereport()`

MA plot

Before normalisation



After normalisation

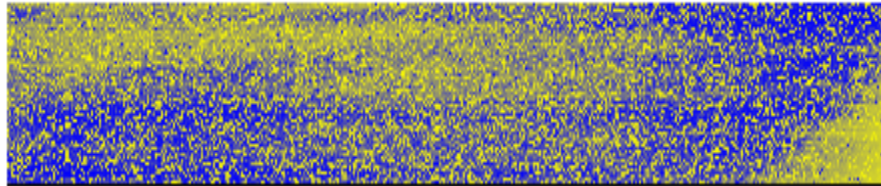
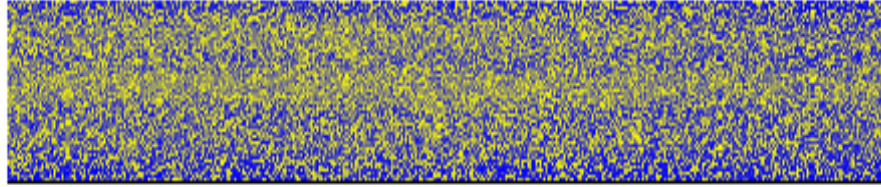


A

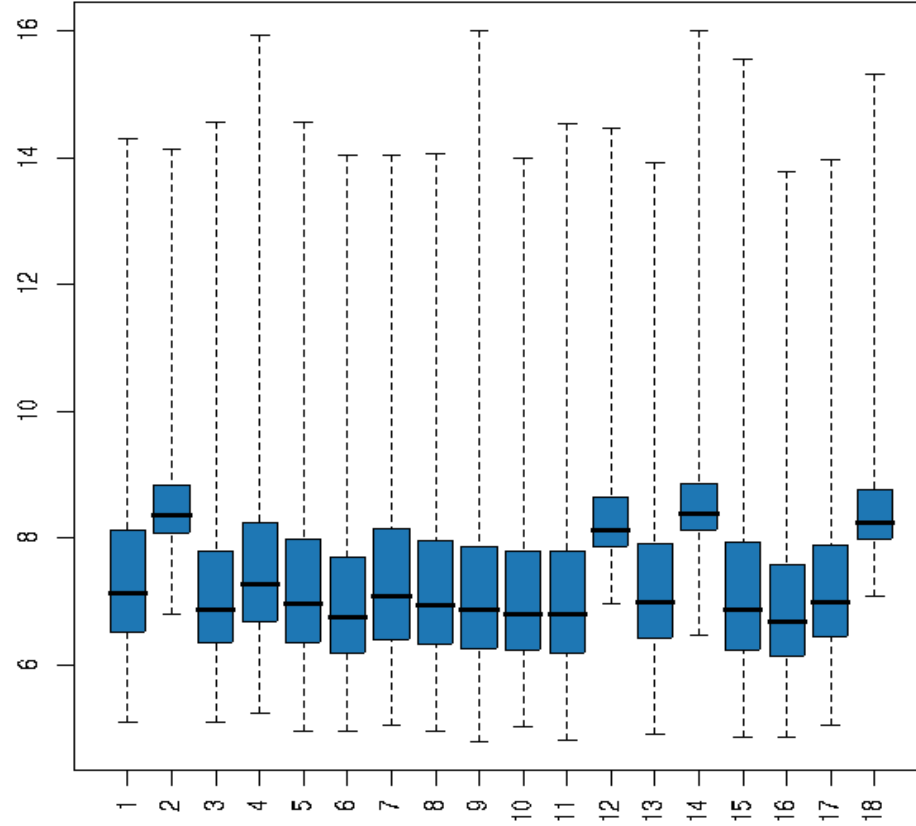
$$M = \log_2(I_1) - \log_2(I_2)$$

$$A = 1/2 (\log_2(I_1) + \log_2(I_2))$$

Spatial representations

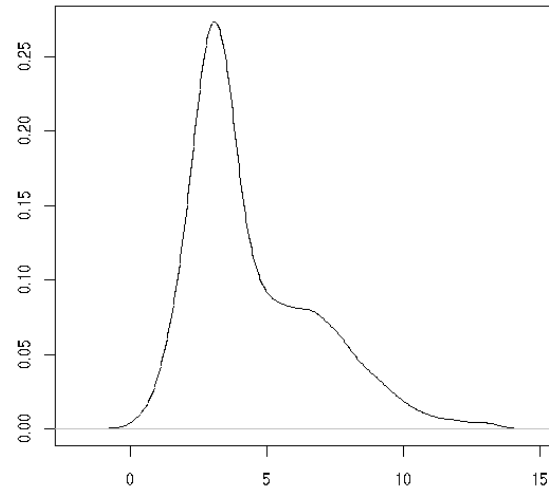
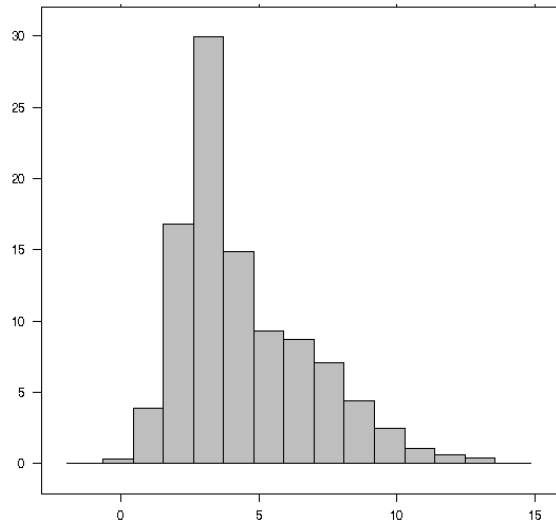


Boxplot



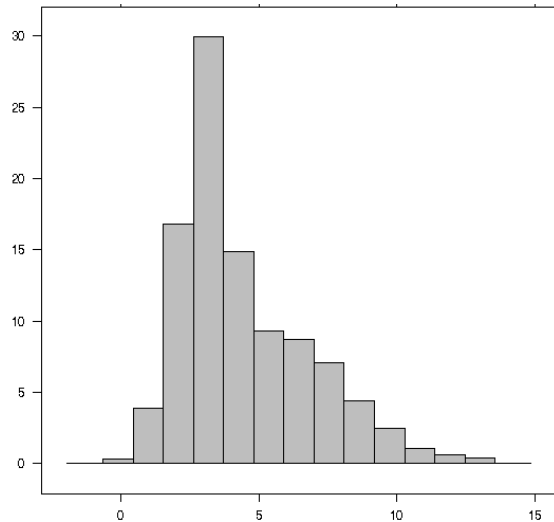
Density plot

- Histograms: graphical representation of frequencies, discrete values
- Density: estimate of the histogram, continuous values

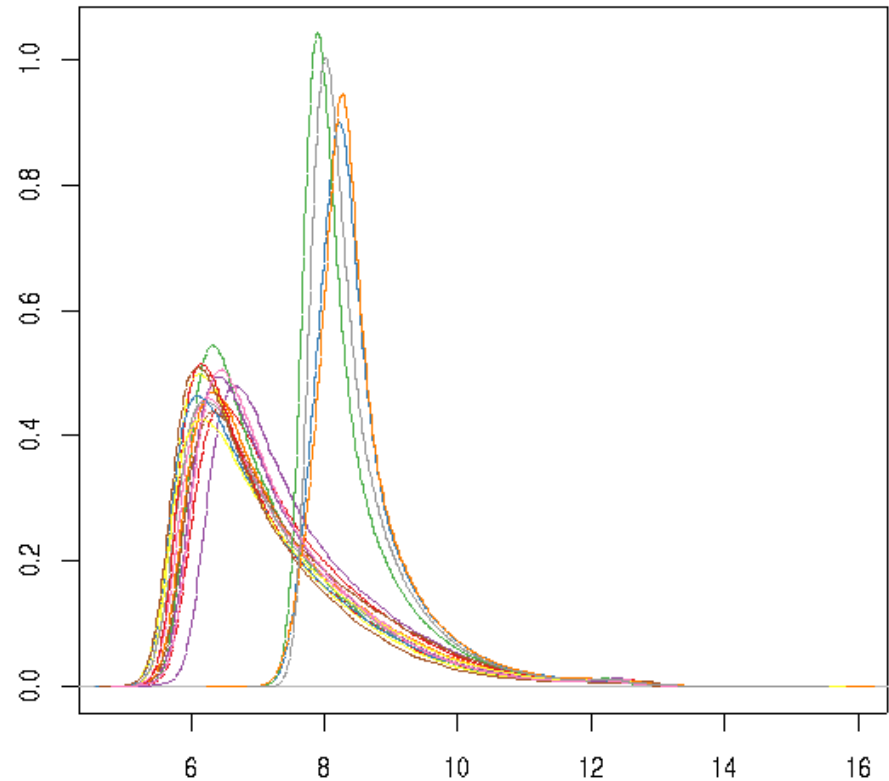


Density plot

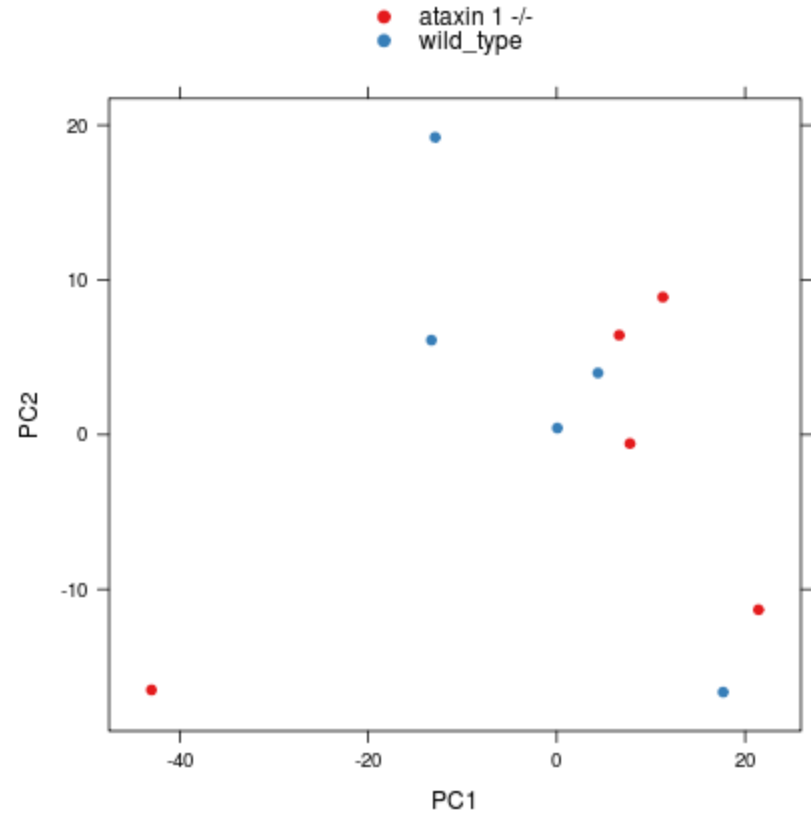
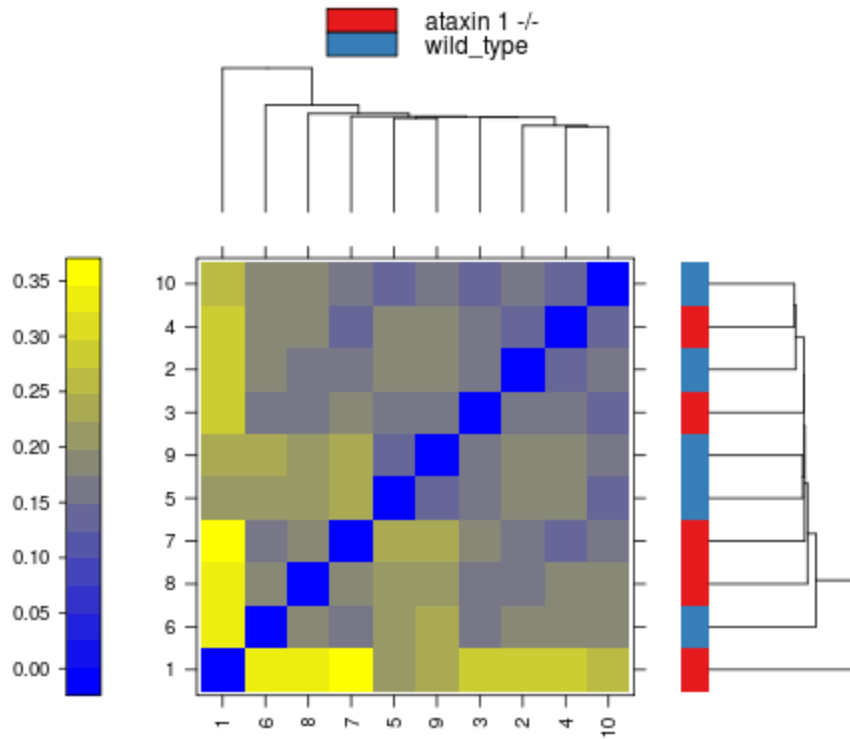
- Histograms: graphical representation of frequencies, discrete values



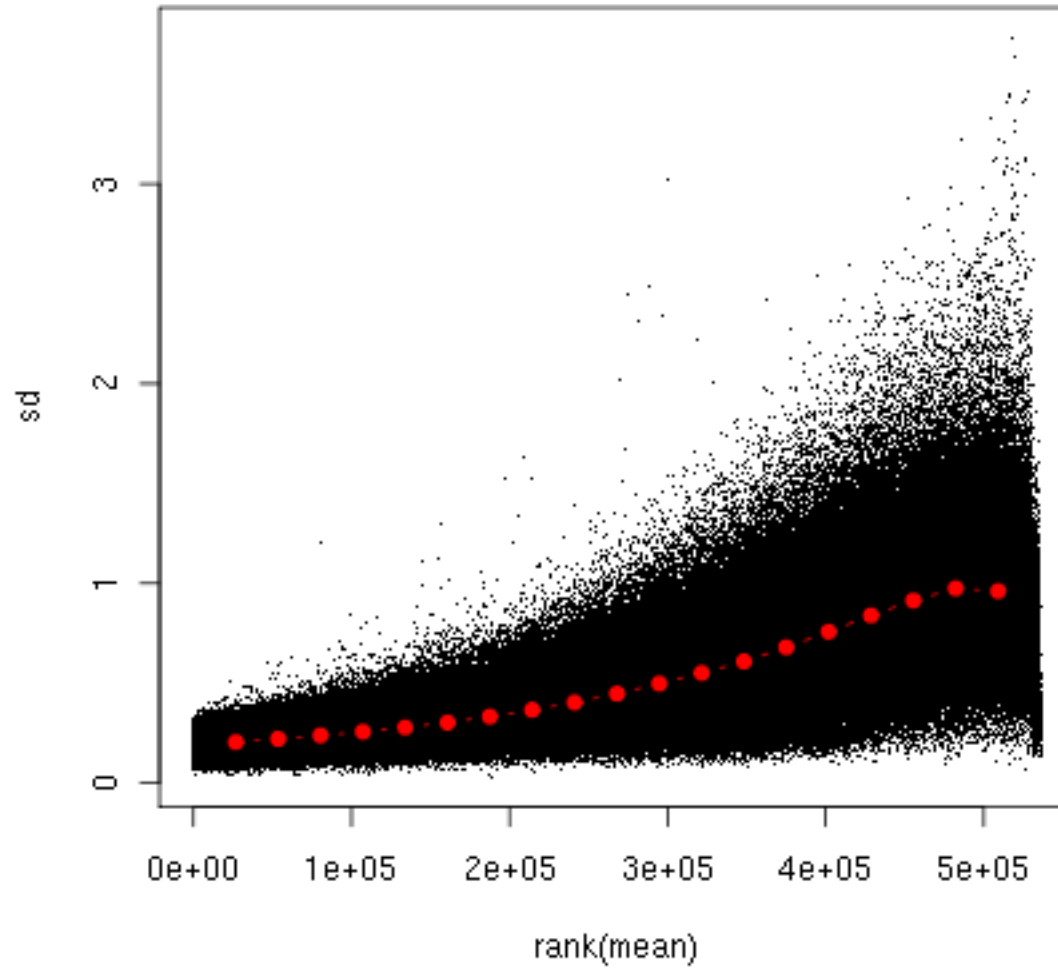
- Density: estimate of the histogram, continuous values



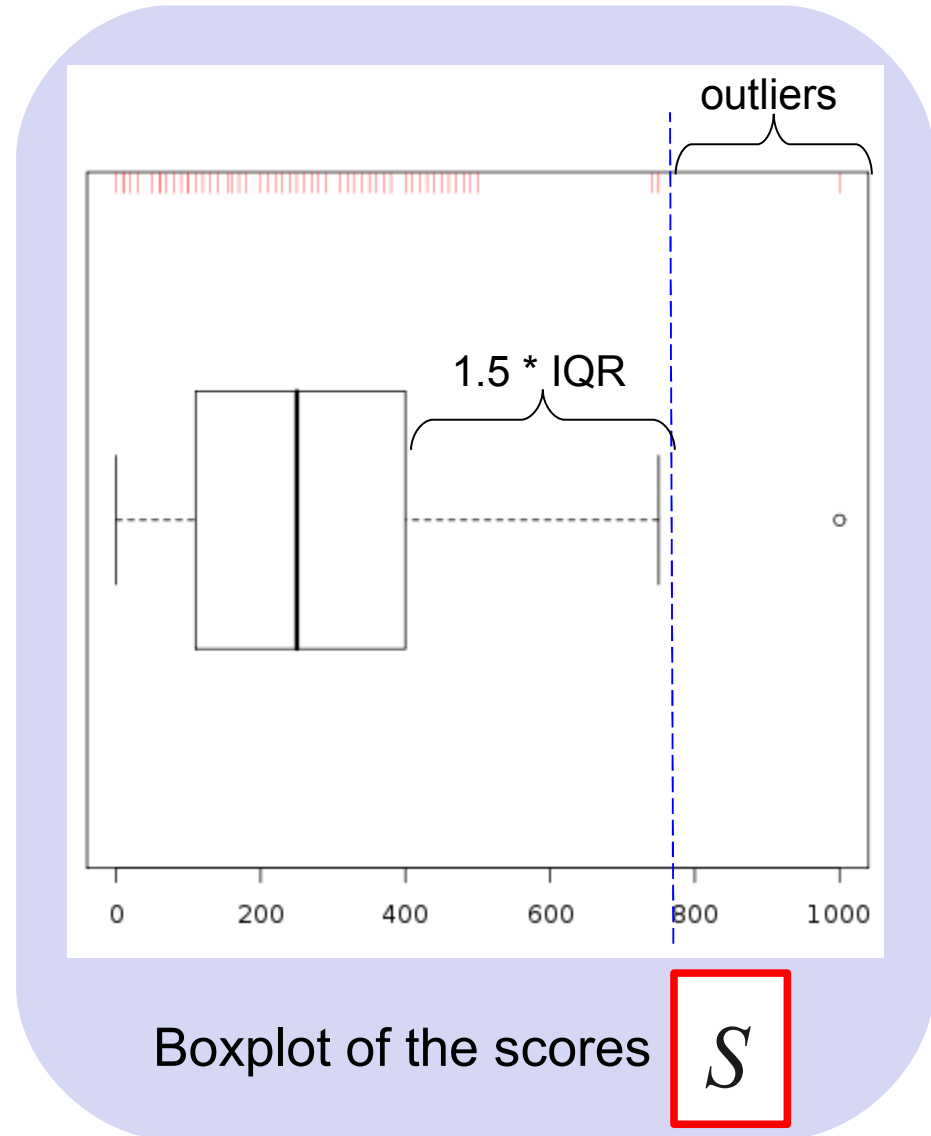
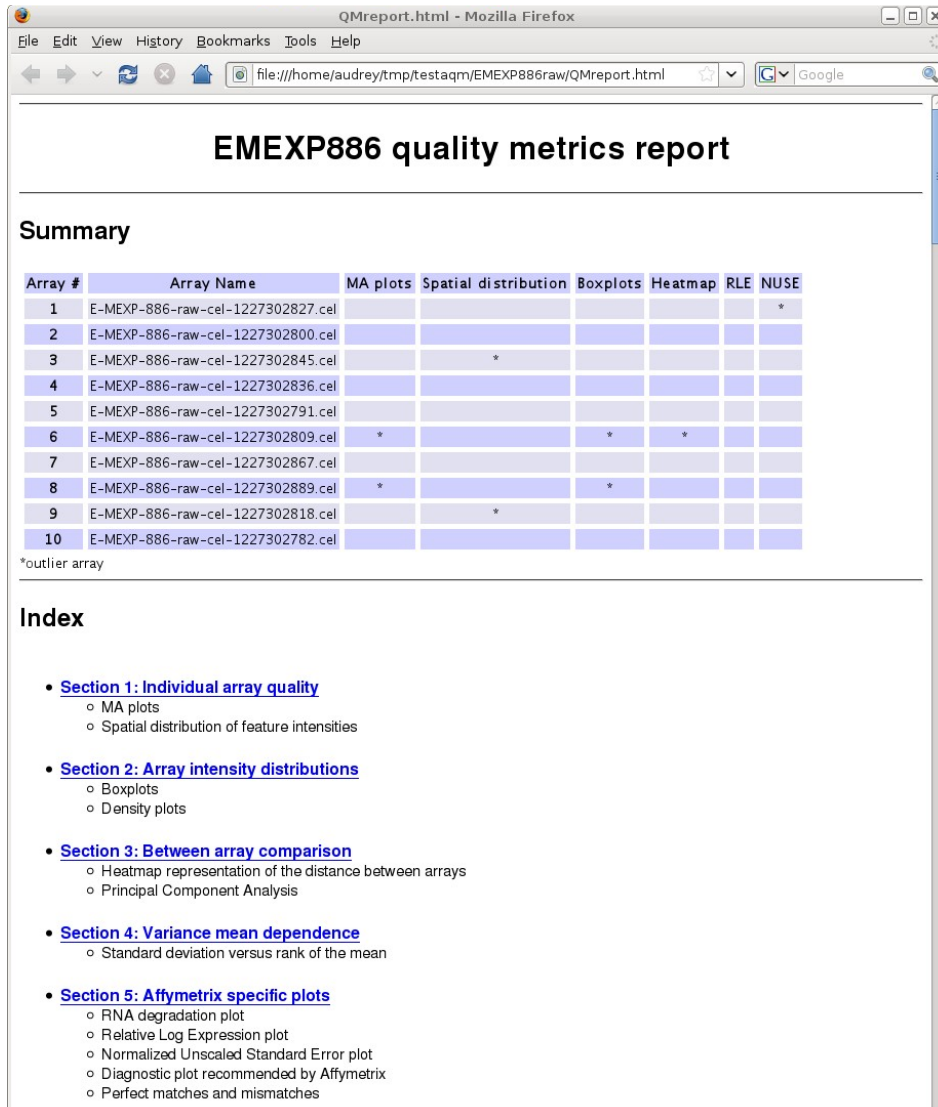
Heatmap and PCA



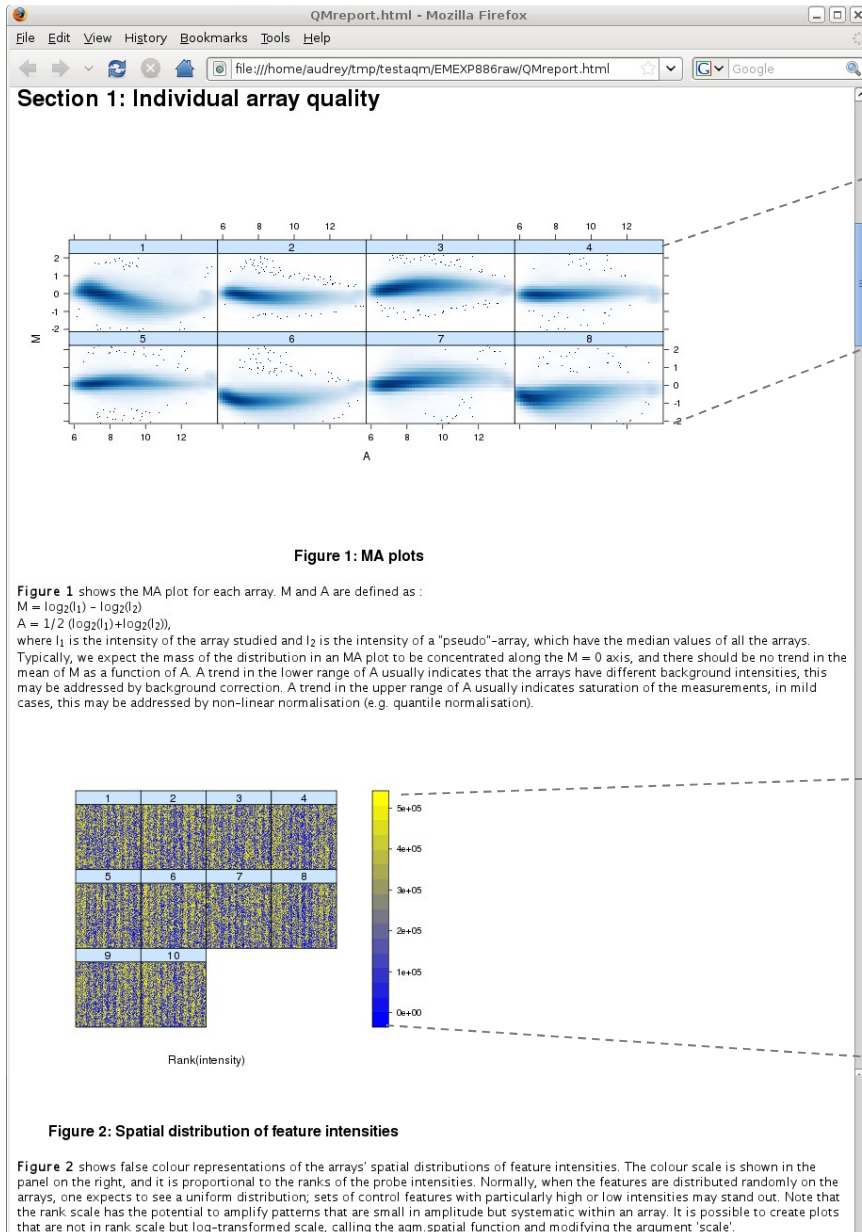
Variance Mean Dependency



arrayQualityMetrics report - *Outlier detection*

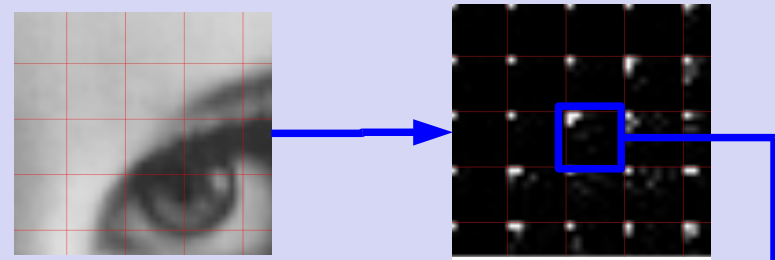


arrayQualityMetrics report - Per array



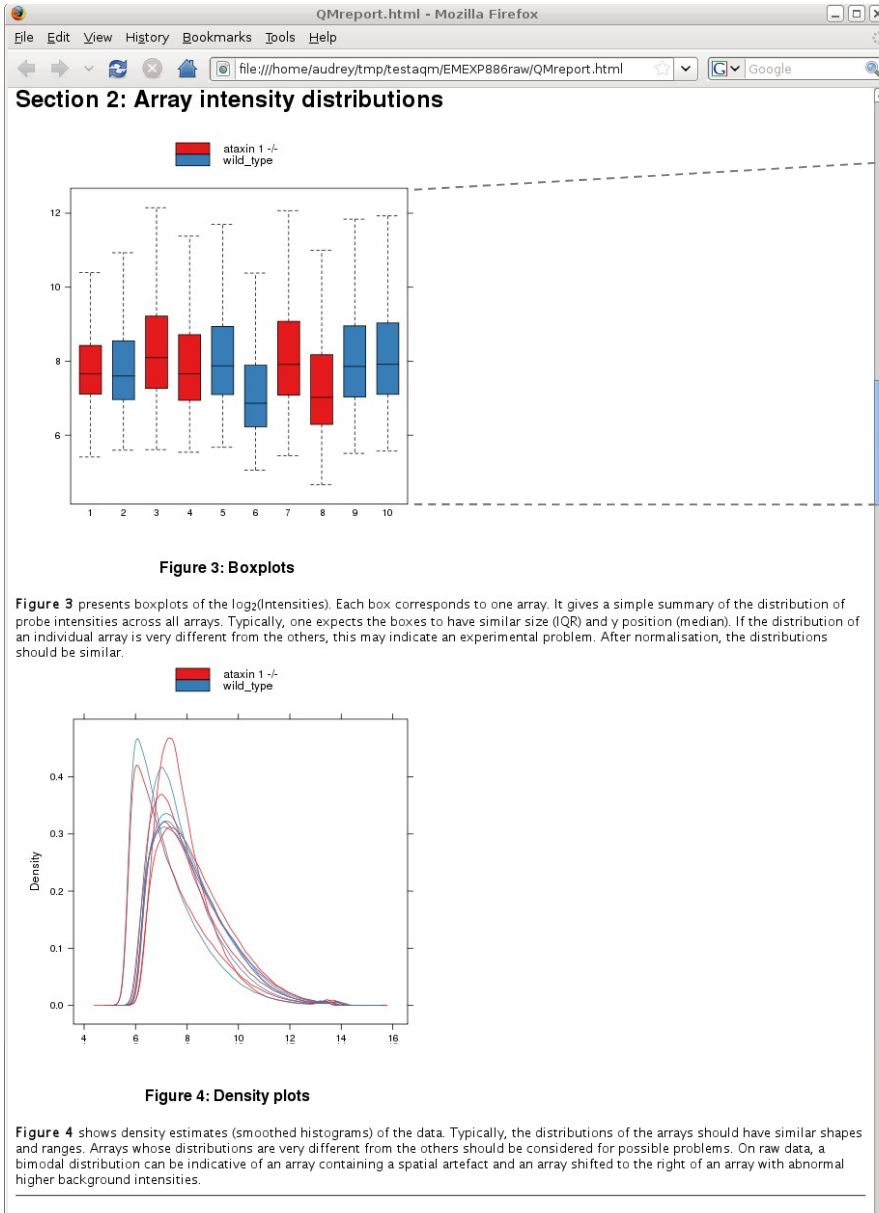
$$S_i = \sum |M_{ik}|$$

Fourier Transform



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

arrayQualityMetrics report - Intensity distributions



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

arrayQualityMetrics report - *Between arrays*

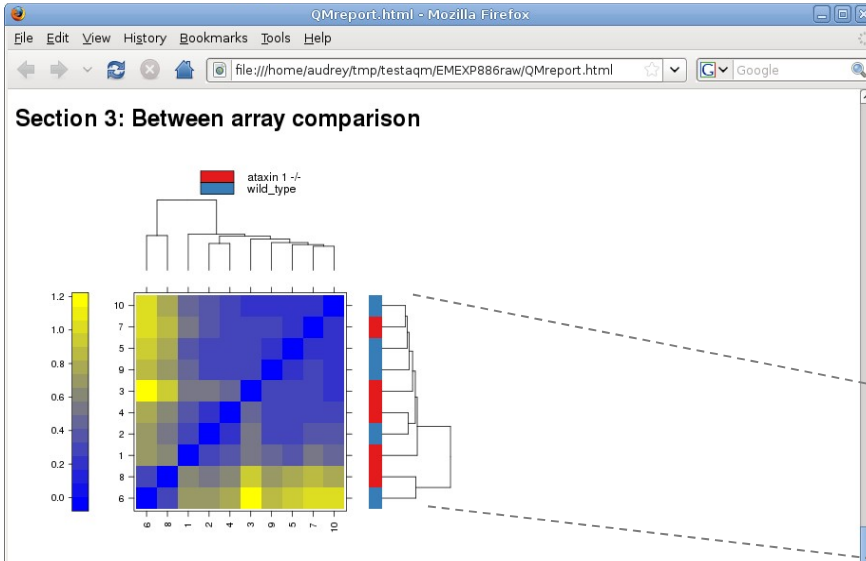


Figure 5: Heatmap representation of the distance between arrays

Figure 5 shows a false colour heatmap of between arrays distances, computed as the mean absolute difference (L₁-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 L₁-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.

$d_{xy} = \text{mean}|M_{xi} - M_{yj}|$
 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} .

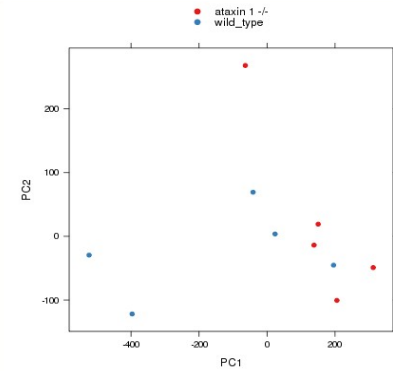


Figure 6: Principal Component Analysis

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:

$$d_{ij} = \text{mean}_k (|I_{ik} - I_{jk}|)$$

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

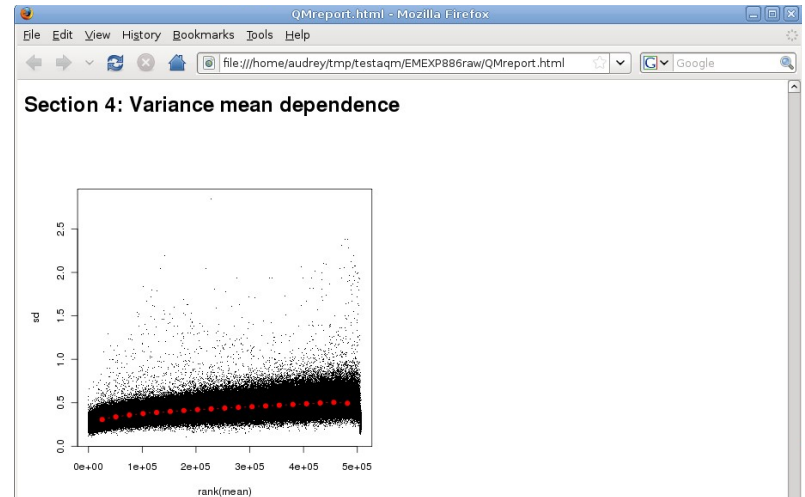


Figure 7: Standard deviation versus rank of the mean

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

Section 5: Affymetrix specific plots

RNA degradation plot

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.

Relative Log Expression plot

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

QC Stats

Probe	QC Stat
act1n3/act1n5	60.25%
gapdh3/gapdh5	72.19
^886-mw-cel-h1227302782.cel	59.94%
^886-mw-cel-h1227302818.cel	65.97
^886-mw-cel-h1227302889.cel	62.76%
^886-mw-cel-h1227302867.cel	49.34
^886-mw-cel-h1227302809.cel	63.09%
^886-mw-cel-h1227302791.cel	70.69
^886-mw-cel-h1227302836.cel	60.03%
^886-mw-cel-h1227302845.cel	48.44
^886-mw-cel-h1227302800.cel	60.22%
^886-mw-cel-h1227302827.cel	73.54
^886-mw-cel-h1227302827.cel	61.41%
^886-mw-cel-h1227302845.cel	67.5
^886-mw-cel-h1227302800.cel	61.36%
^886-mw-cel-h1227302800.cel	78.08
^886-mw-cel-h1227302800.cel	60.08%
^886-mw-cel-h1227302827.cel	71.84
^886-mw-cel-h1227302827.cel	54.16%
^886-mw-cel-h1227302827.cel	77.8

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.

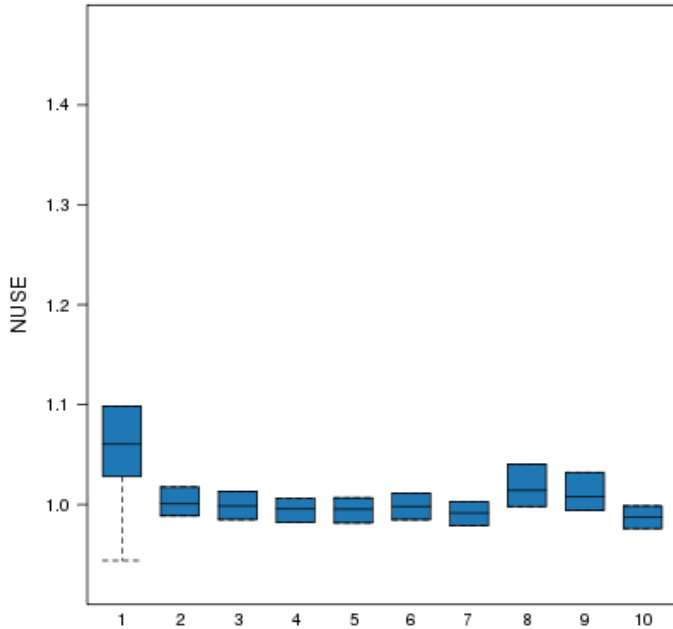
Perfect matches and mismatches

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



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

$$NUSE(\beta_{ik}) = \frac{SE(\beta_{ik})}{\text{med}_i(SE(\beta_{ik}))}$$

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

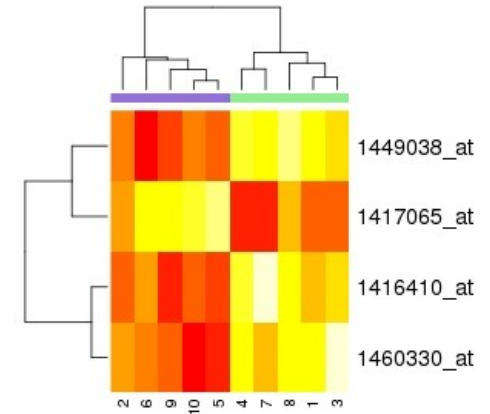
Why is outlier detection important? Example

- ArrayExpress experiment E-MEXP-886, cerebellar gene expression:
 - 5 WT mice (15 weeks of age)
 - 5 Atxn1 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

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

Quality report after normalisation - outlier detection

QMreport.html - Mozilla Firefox

file:///home/audrey/Desktop/EMEXP886norm/QMreport.html

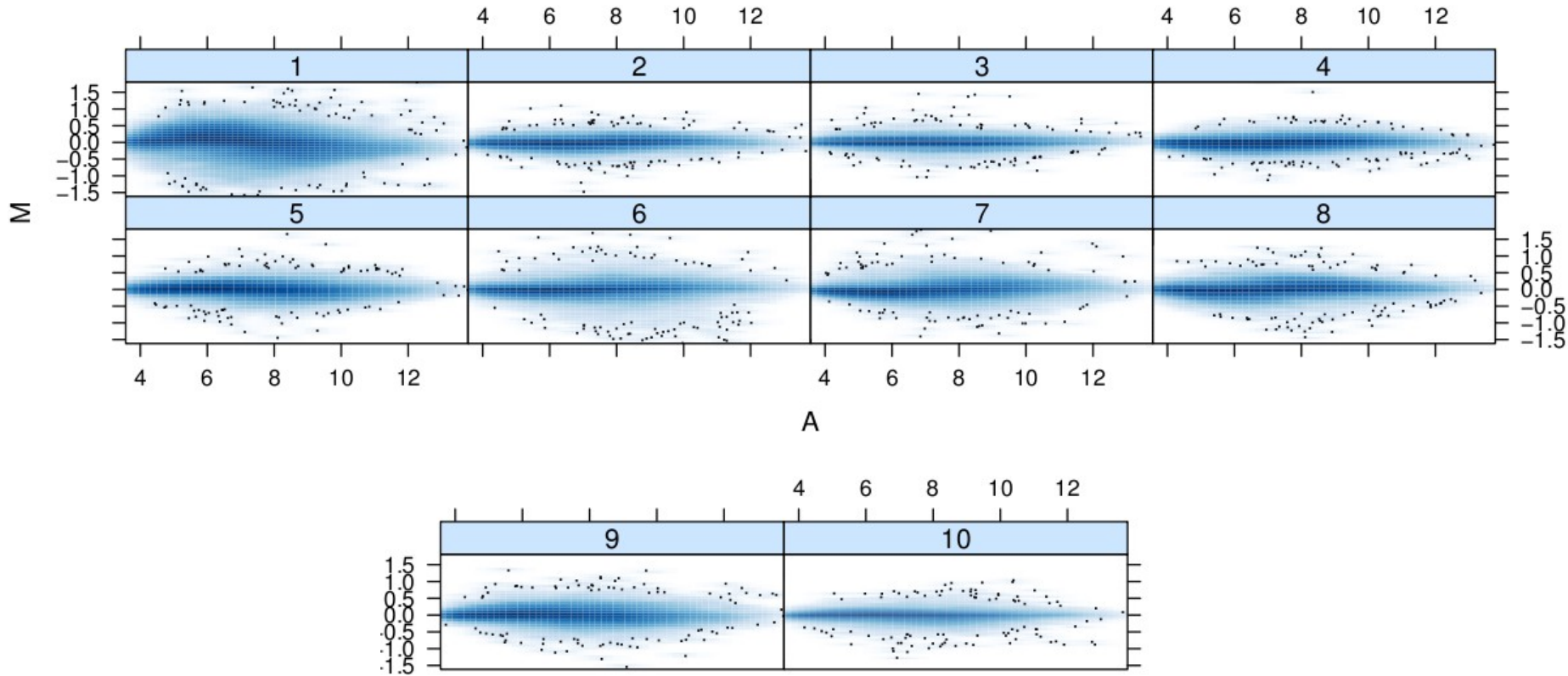
EMEXP886norm quality metrics report

Summary

Array #	Array Name	MA plots	Boxplots	Heatmap
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			

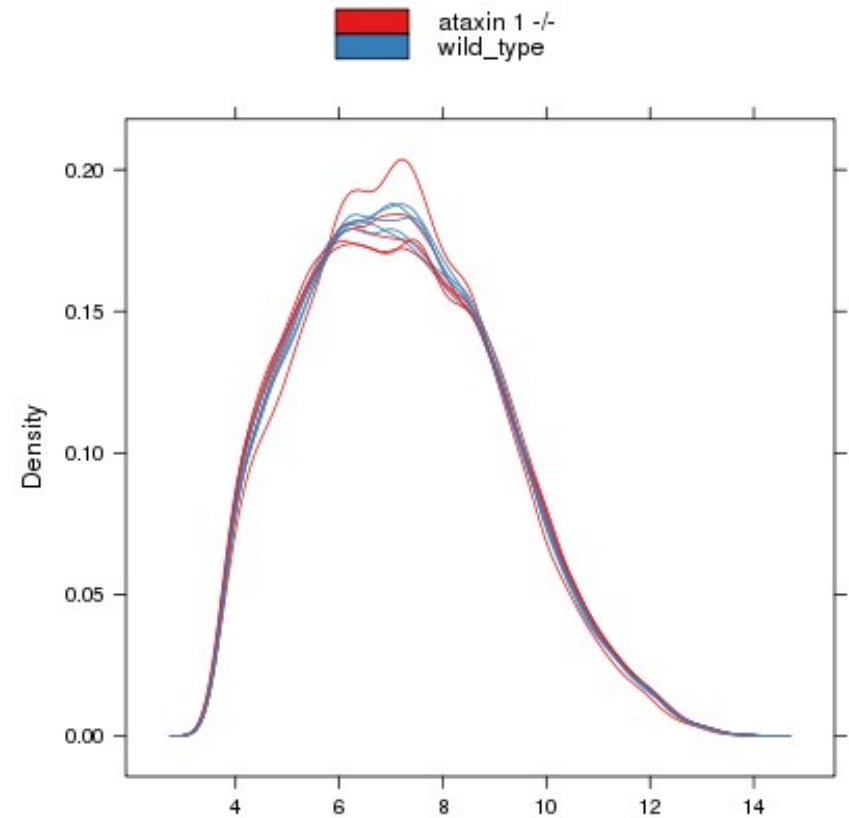
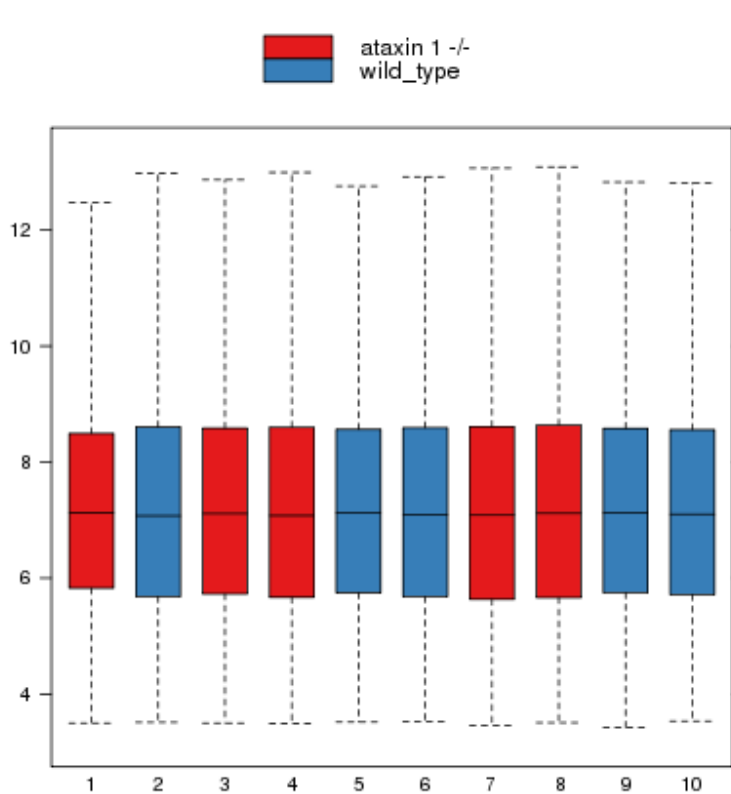
*outlier array

Quality report after normalisation - *per array*



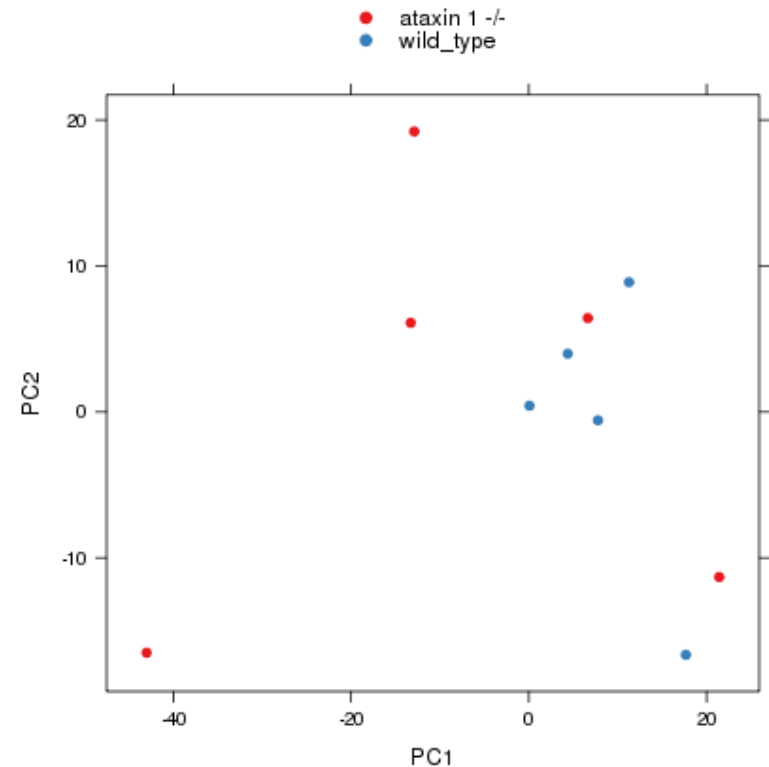
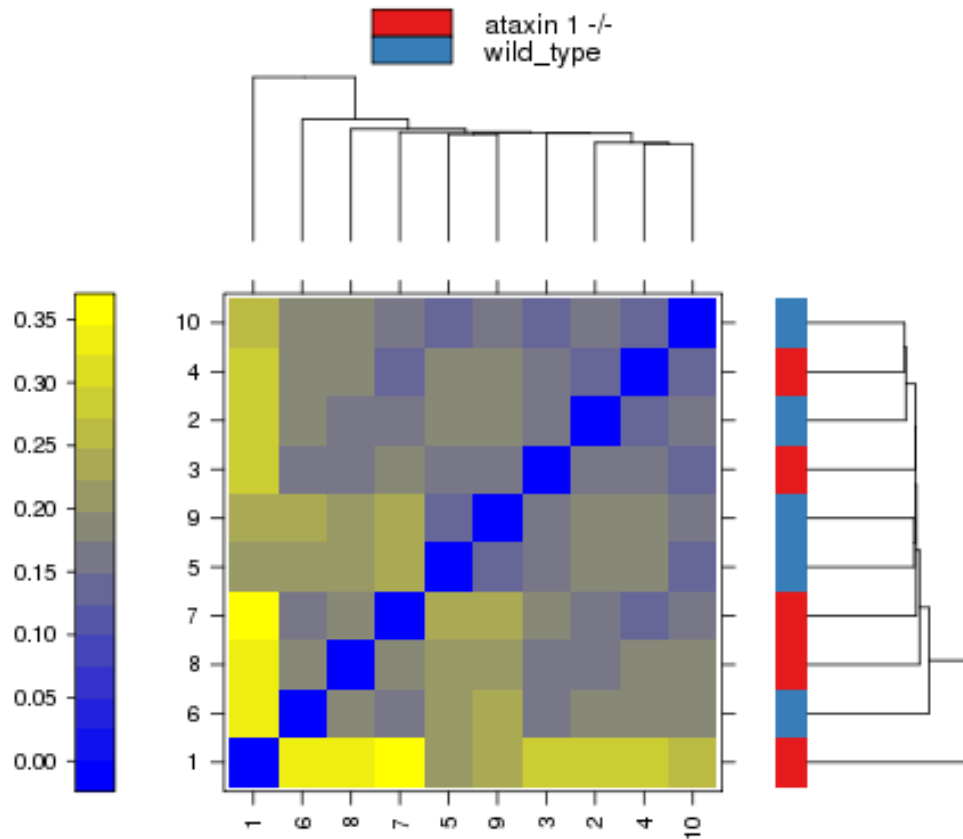
- Excellent M vs A plots except for array #1

Quality report after normalisation - *intensity distributions*



- Very homogeneous intensity distribution
- Smaller wide of the box for array #1

Quality report after normalisation - *array comparison*

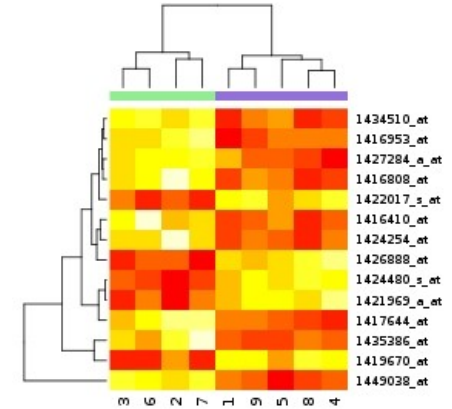


- No clustering of the samples according to biological meaning
- Array #1 further distance from all the other arrays

Outlier array's impact on results

- Moderated t-test

	# 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

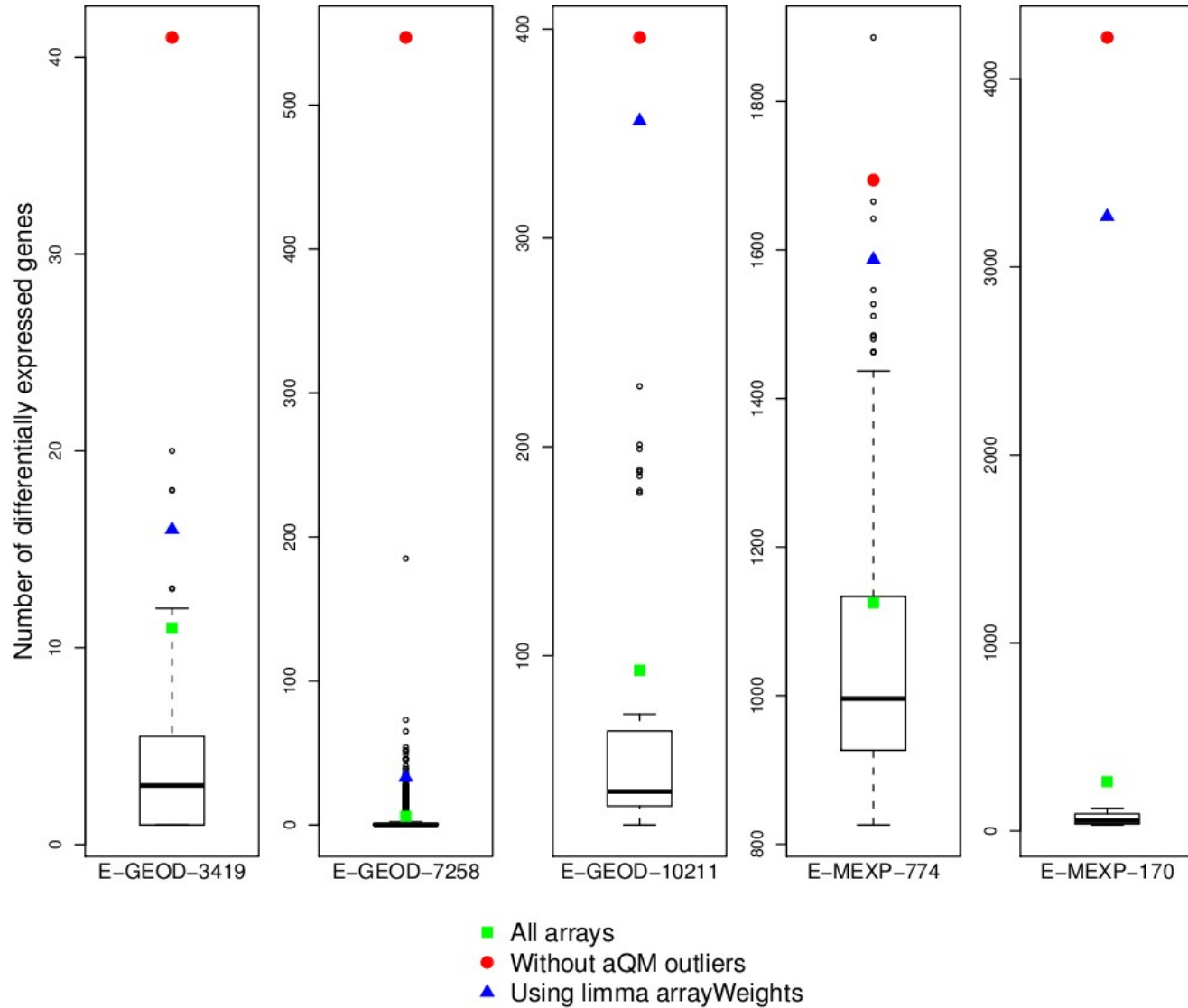
Validation

- Specific effect of array 1?

⇒ Remove one by one each of the arrays then apply a moderated t-test on the remaining samples

	# 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

Importance of outlier detection



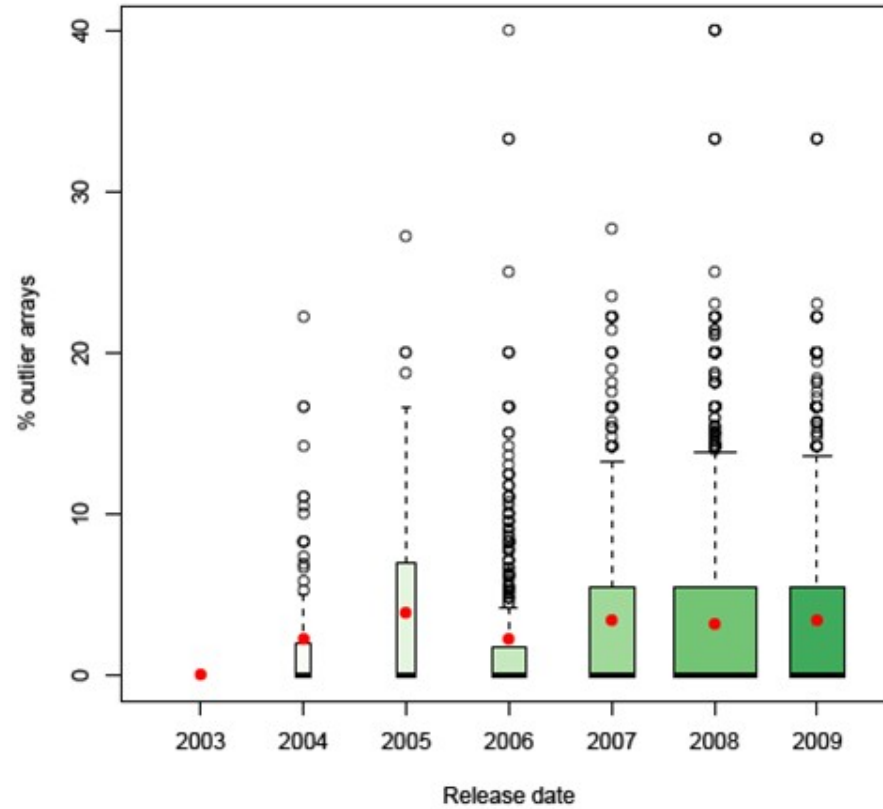
Importance of outlier detection

Pathway name	Genes	p-value when removing outliers	p-value when all arrays
<i>E-GEOD-3419</i>			
Pyrimidine metabolism	37	$<10^{-3}$	0.701
Base excision repair	17	0.001	0.542
DNA replication	19	0.003	0.451
Cell cycle	69	0.009	0.387
TGF-beta signaling pathway	48	0.009	0.558
<i>E-GEOD-7258</i>			
Pentose phosphate pathway	13	0.003	0.588
Fructose and mannose metabolism	28	0.003	0.326
Biosynthesis of steroids	20	0.003	0.012
Oxidative phosphorylation	44	0.003	0.299
Starch and sucrose metabolism	16	0.003	0.317

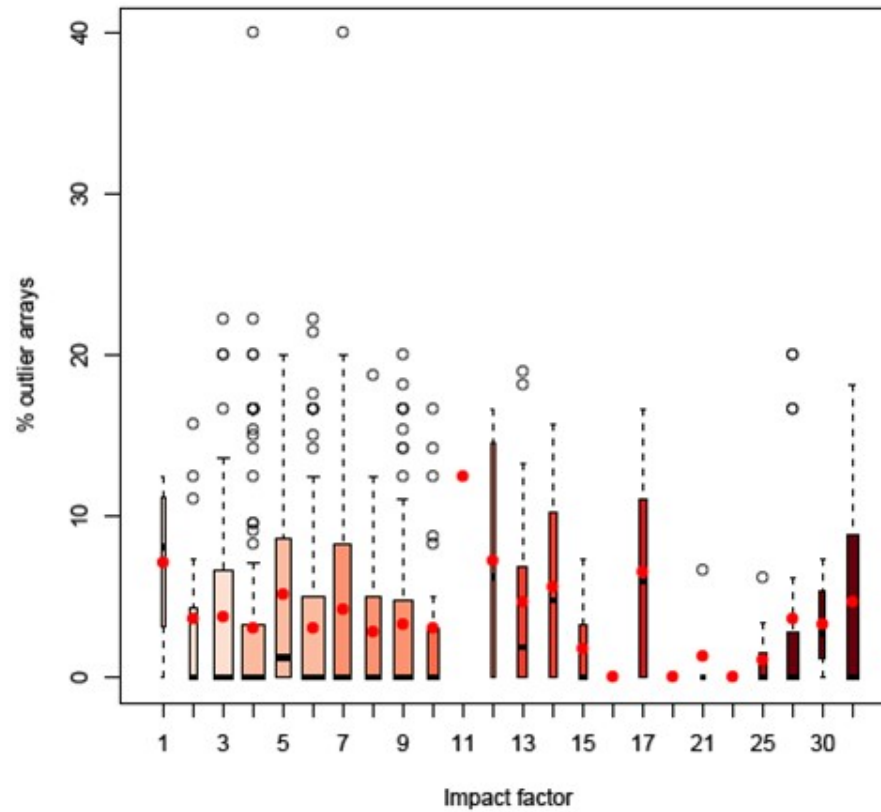
Application to the ArrayExpress database

- ArrayExpress: public repository for microarray data
- Store MIAME-compliant data in accordance with MGED recommendations: MAGE-TAB format
- <http://www.ebi.ac.uk/microarray-as/aer>
- Build R objects from 7000 datasets using ArrayExpress package
- Run arrayQualityMetrics on these datasets
- Study Array-Distance (heatmap) and NUSE outliers

Outliers per year



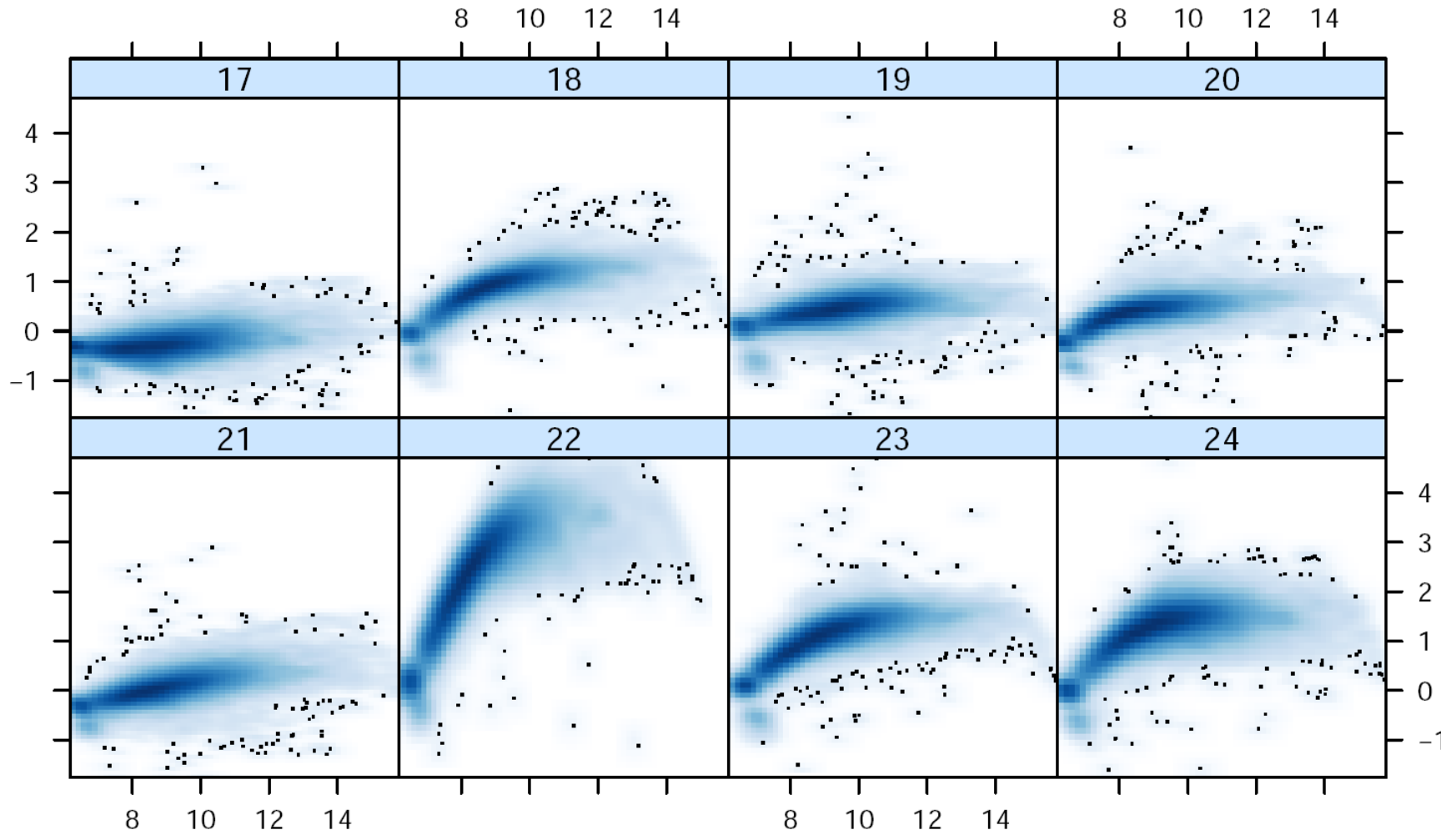
Outliers per impact factor

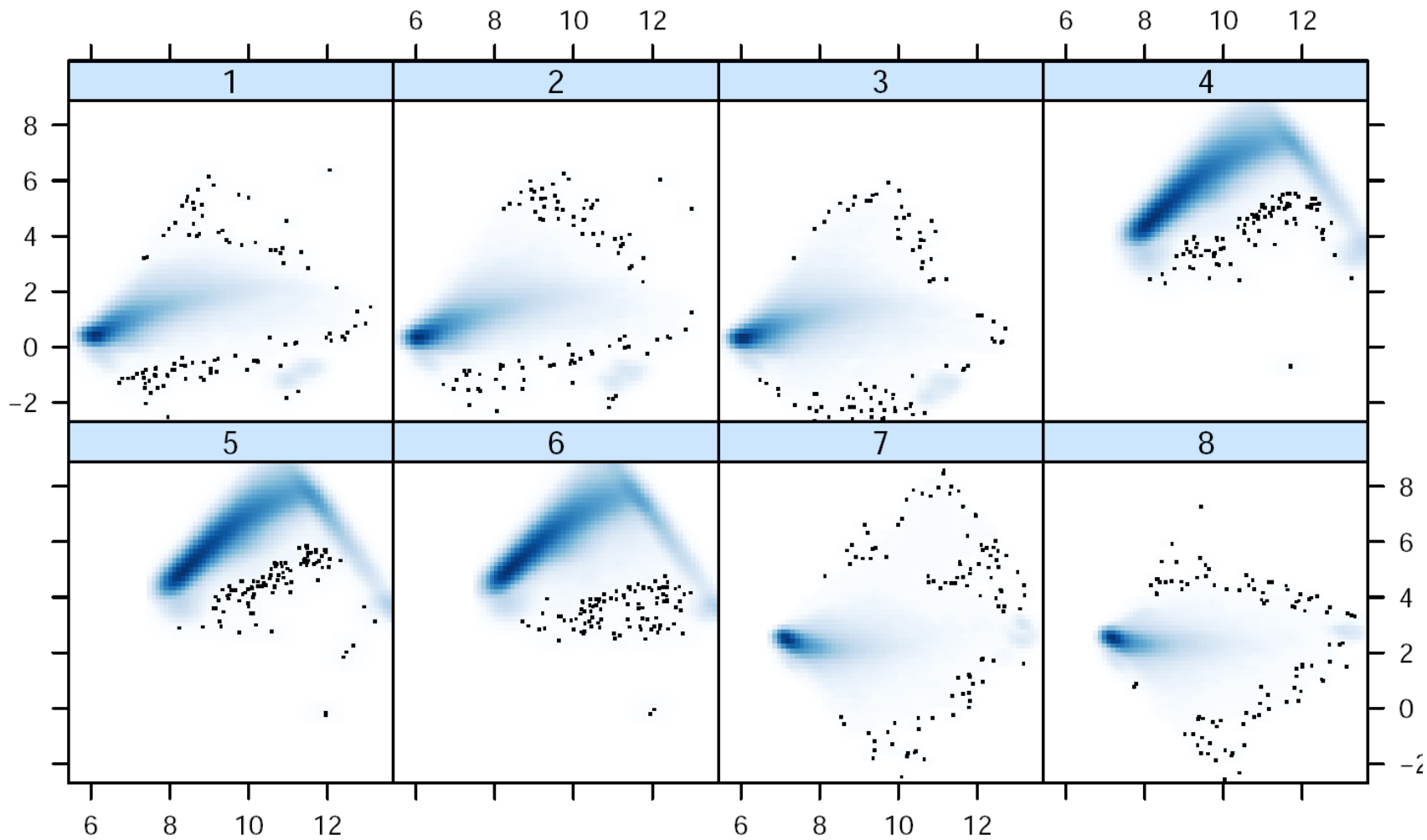


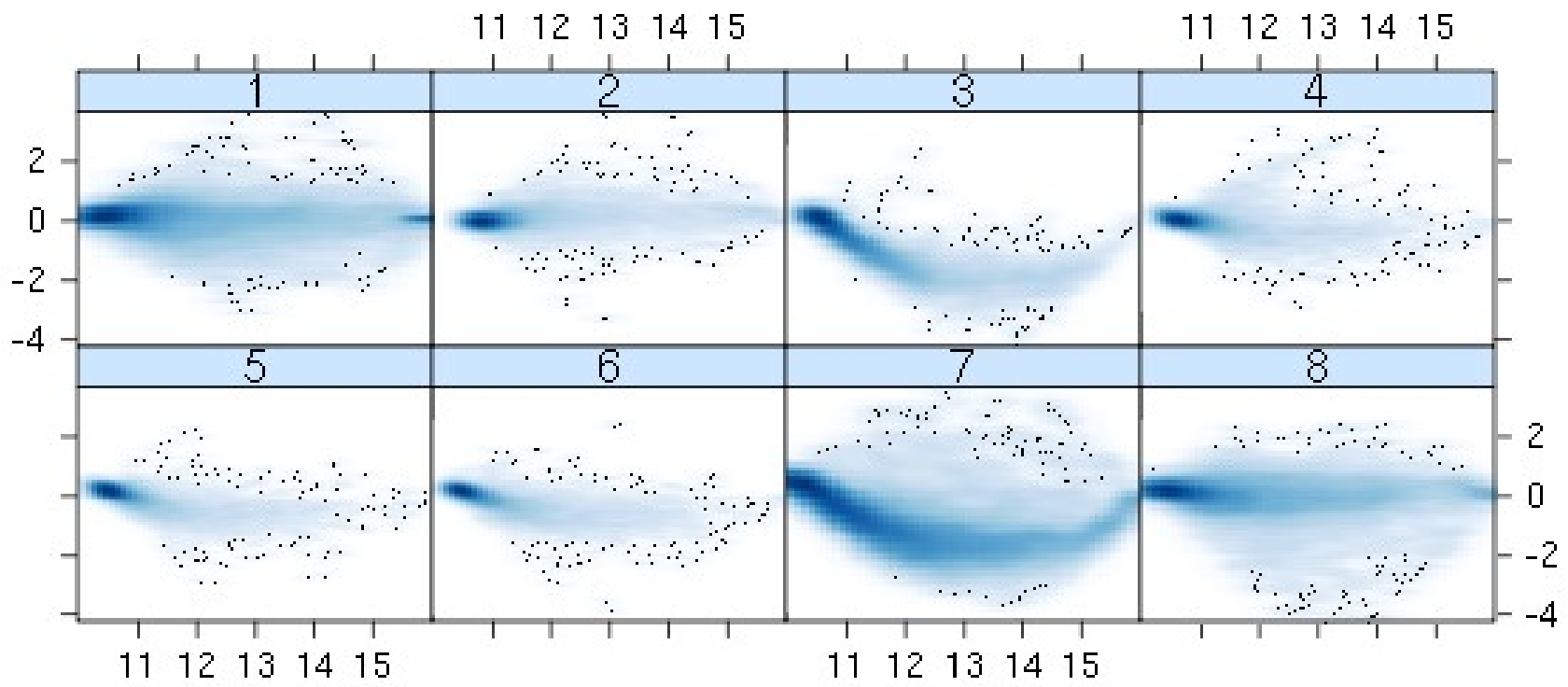
Conclusions

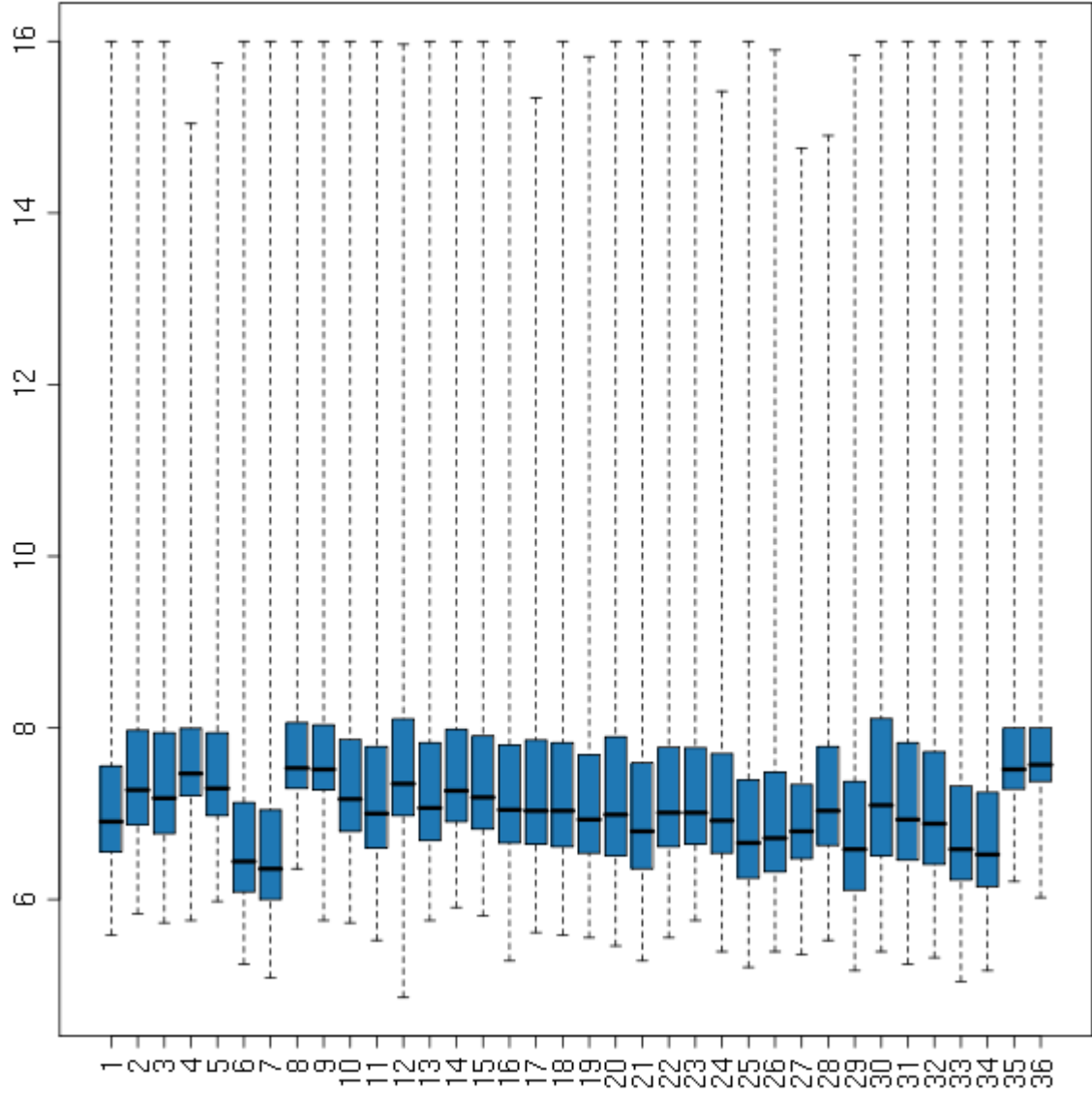
- Quality assessment is important
 - Still needed
 - First “taste” of the data
 - Removing outliers increases statistical power and biological significance
- arrayQualityMetrics
 - Before preprocessing: to choose preprocessing methods
 - After normalisation: to check preprocessing efficiency
 - Comprehensive report
 - Outlier detection
 - Used on any kind of expression array

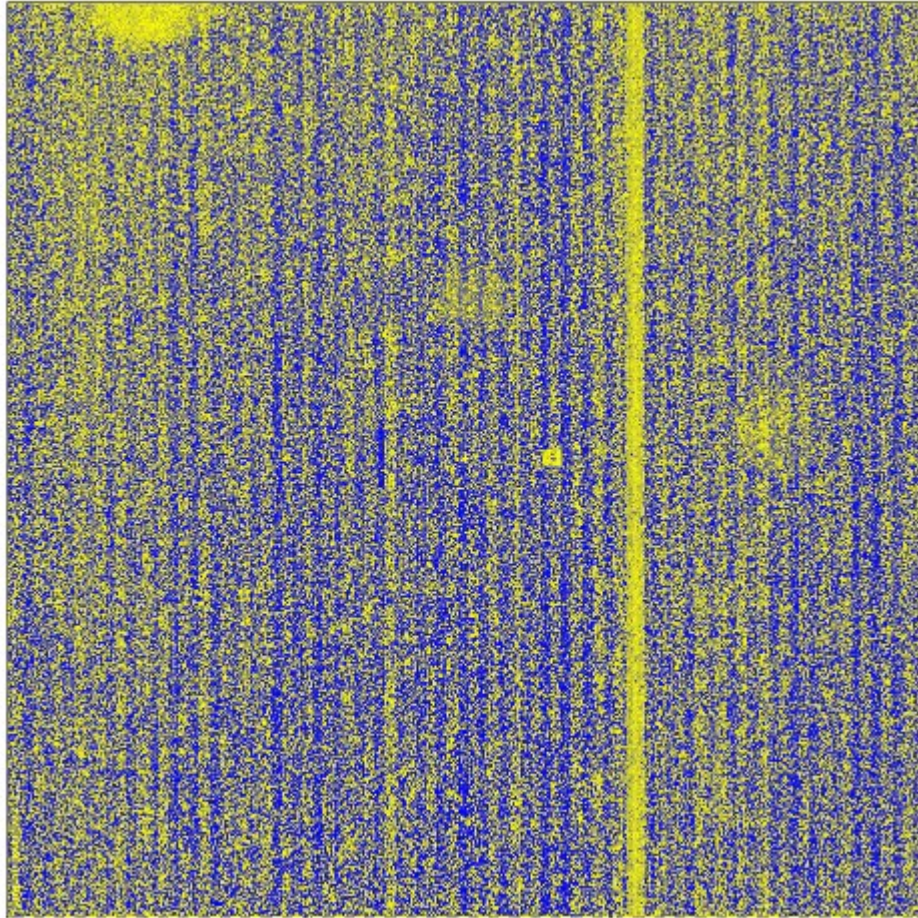
Horror Picture Show



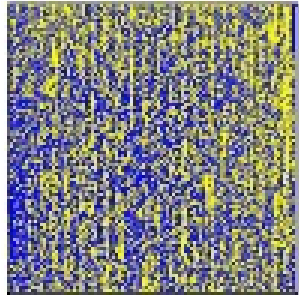




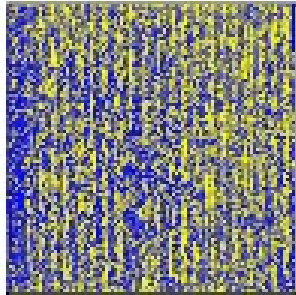




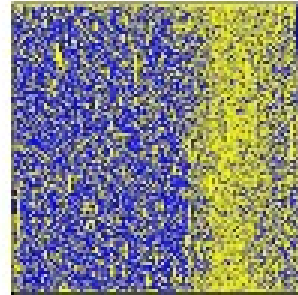
1



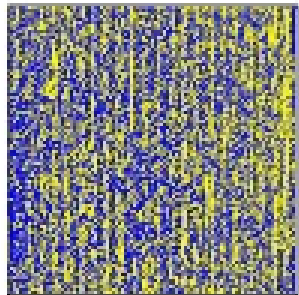
2



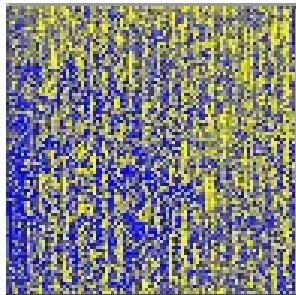
3



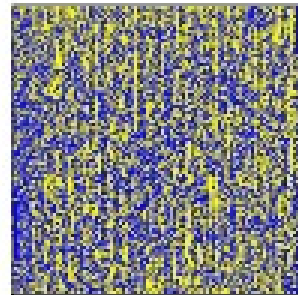
4

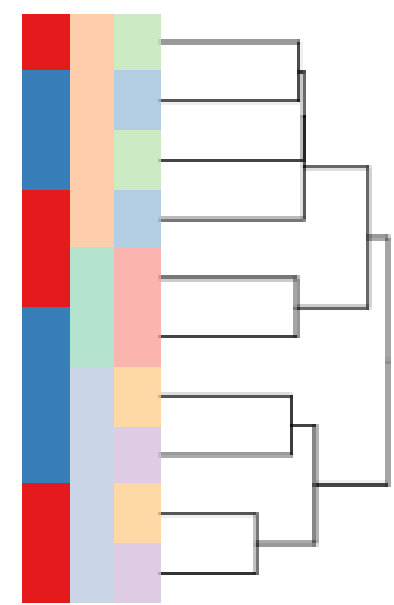
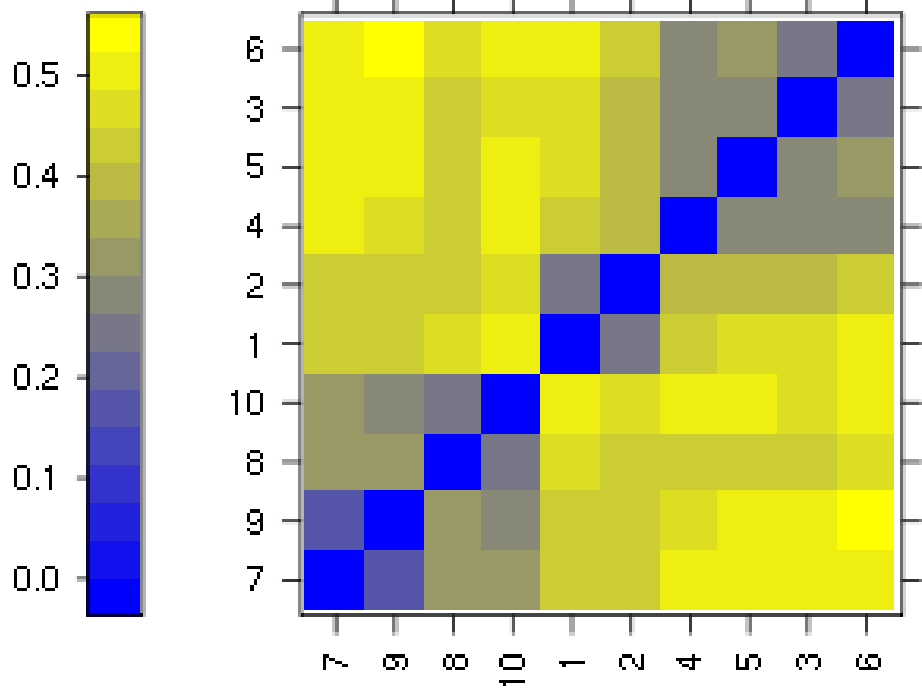
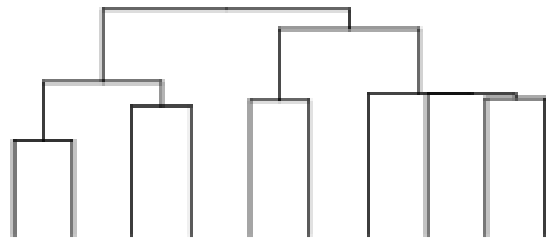
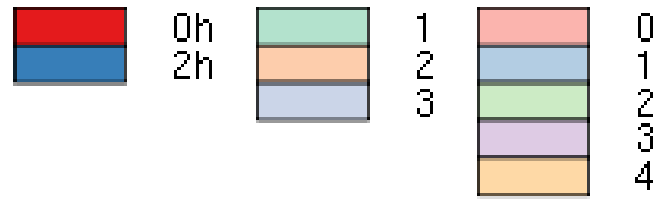


5

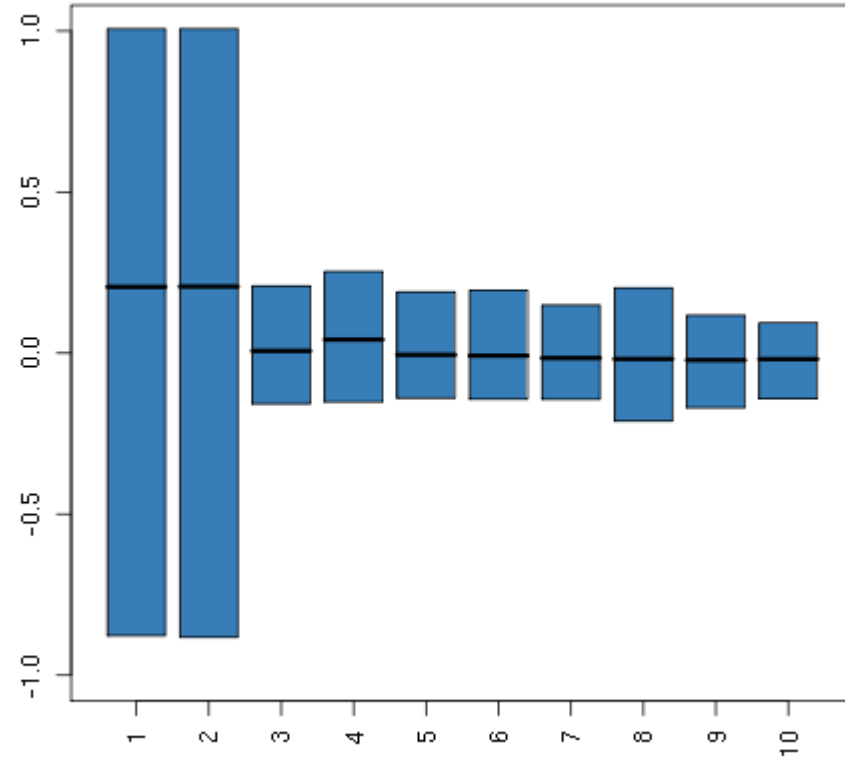


6





RLE



NUSE

