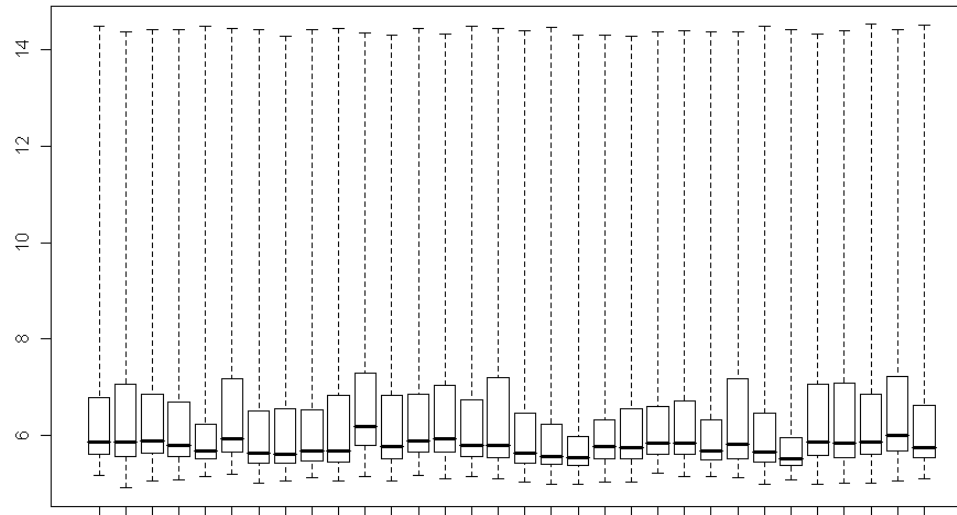




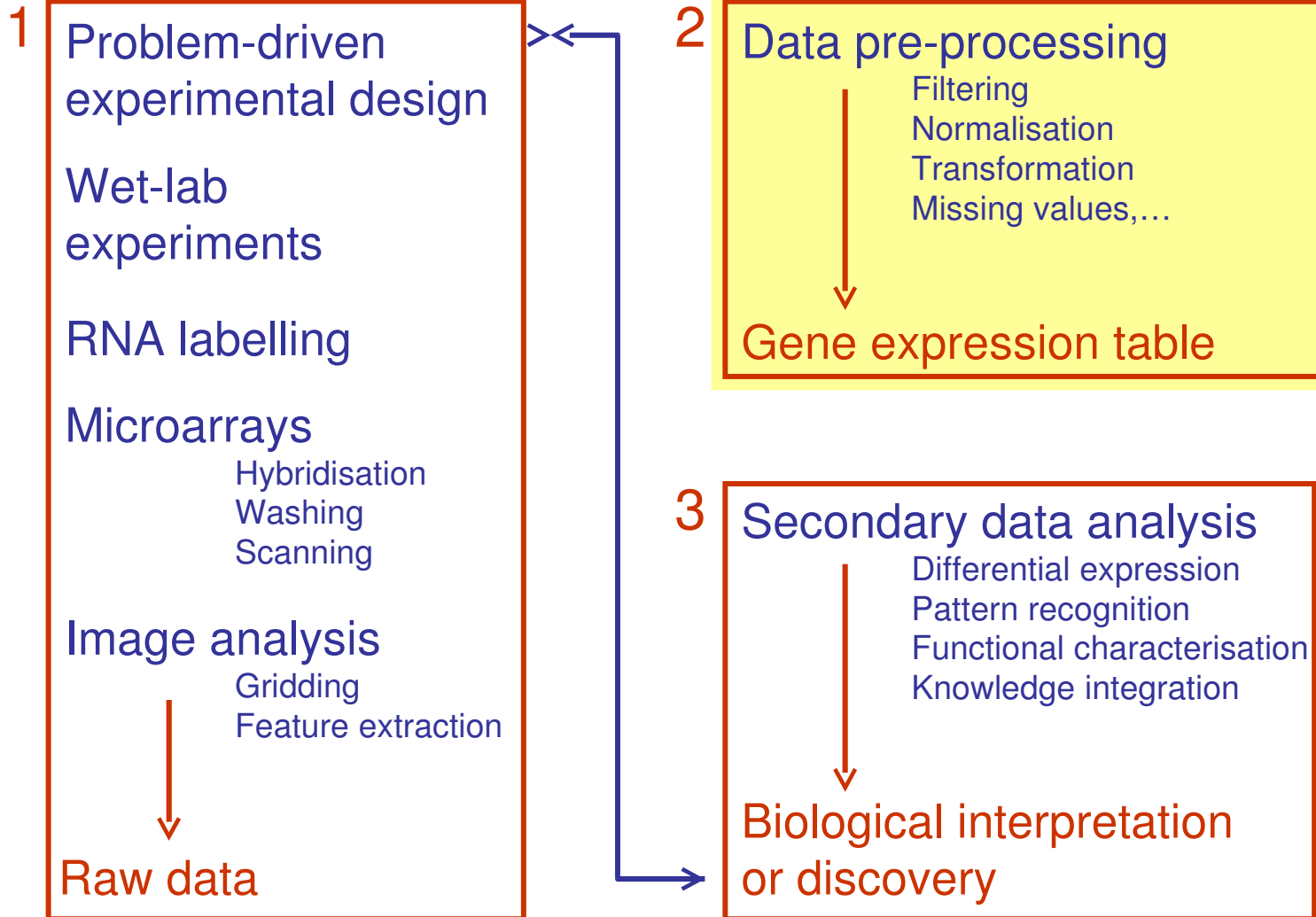
Pre-processing and quality control of microarray data



Christine Stansberg, 20.04.10



Workflow microarray experiment





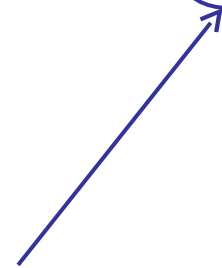
Gene expression table



Tissues / conditions →

Probe ID	Gene ID	Gene Name	Cortex	Hippocampus	Striatum	Cerebellum	Thalamus	Glia
20693868	rCG62444.1	glycine receptor, beta subunit	17,39	17,55	37,65	12,70	26,73	9,79
21370555	rCG44406.1	5-hydroxytryptamine (serotonin) receptor 1A	10,20	26,77	0,88	0,76	1,08	1,96
21027316	rCG52199	adrenergic receptor, alpha 1a	2,85	2,47	1,97	2,85	2,90	2,90
22181482	rCG54200	apolipoprotein E	625,67	748,79	680,45	298,79	402,60	277,70
22013413	rCG56541.1	basic helix-loop-helix domain containing, class B2	23,15	15,65	14,19	7,66	18,25	4,63
20880297	rCG39243.1	bombesin-like receptor 3	2,18	0,47	0,65	1,93	0,46	1,84

Genes ↓



Normalised signal
(i.e. expression level)
of one gene in one tissue



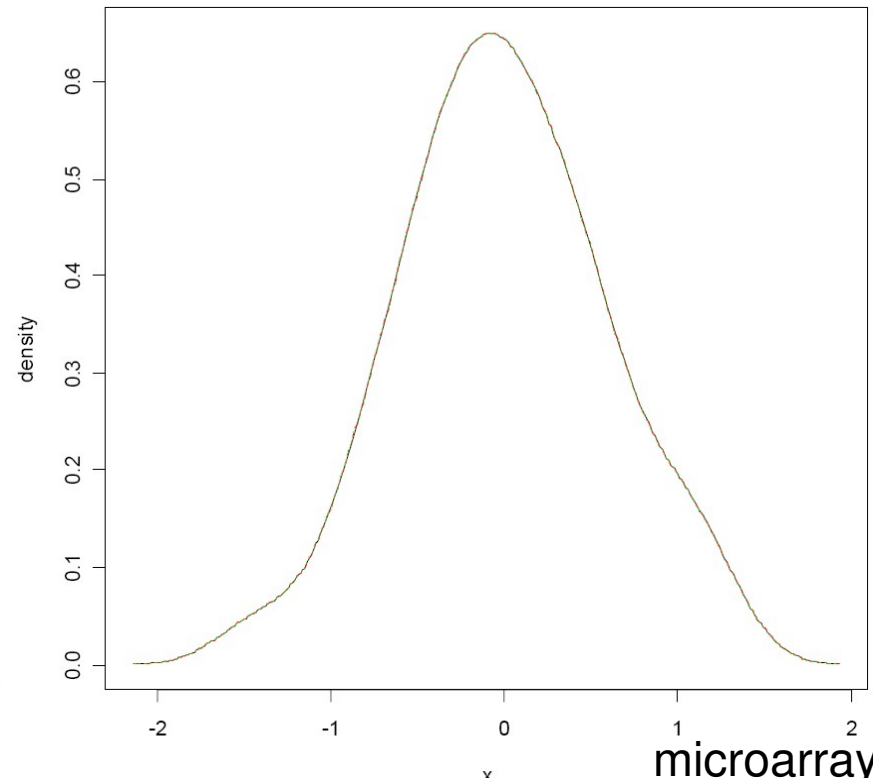
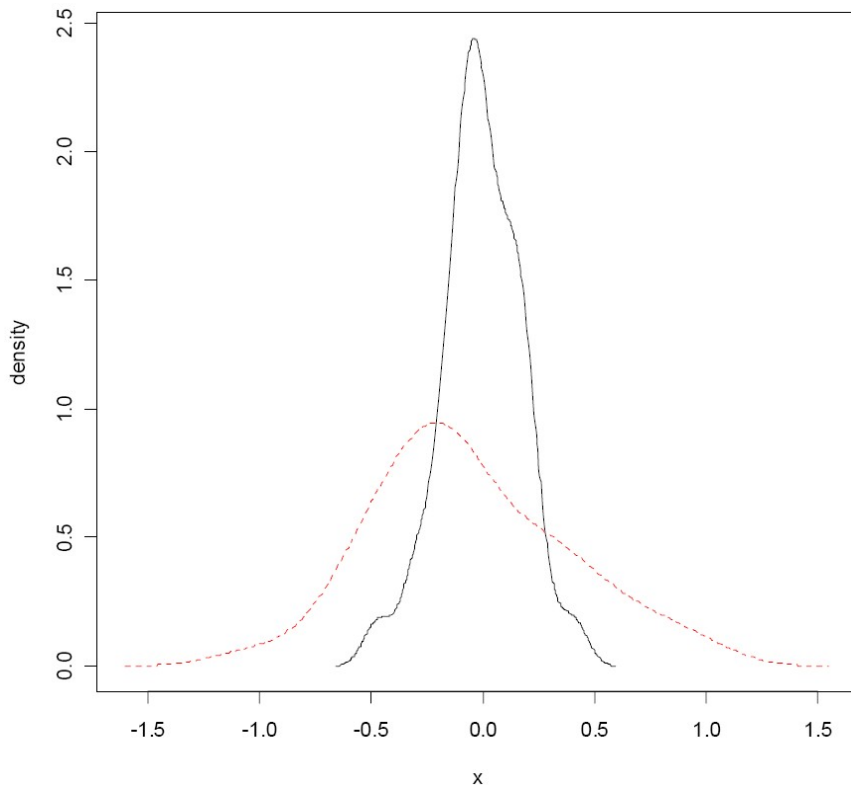
Preprocessing

- Remove effects of the technical process from the data
 - Correct background
 - Filter bad quality spots
 - Missing values must then be replaced
 - Correct for intensity level and between-array variability
 - Arrays need to have a similar signal distribution to be comparable
 - ✂ → Normalise
 - Transform the data
 - Equal variance across intensities
 - Normally distributed



Normalisation

- We assume that the intensity distributions across arrays are the same
- This is not always (never?) the case
- Intensity distributions need to be similar for the arrays to be comparable



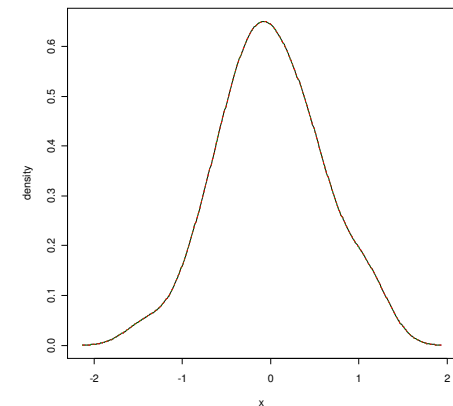
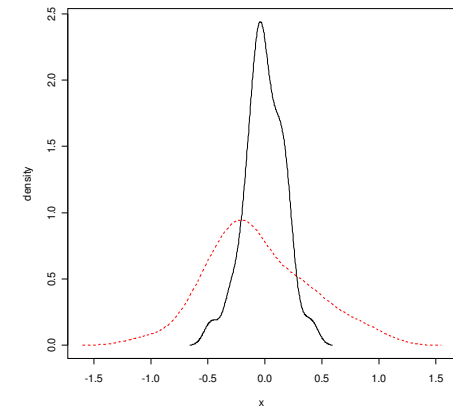


Normalisation

- Remove bias and artefacts from the technical process
 - Intensity differences
 - Reagent batches or age
 - Weather
 - Sample treatment
- After normalisation, data contains primarily biological signal
- **Problem:** Normalisation assumes that
 - Just a few genes are differentially expressed
 - Equal probability for up- and down-regulation

Quantile normalisation

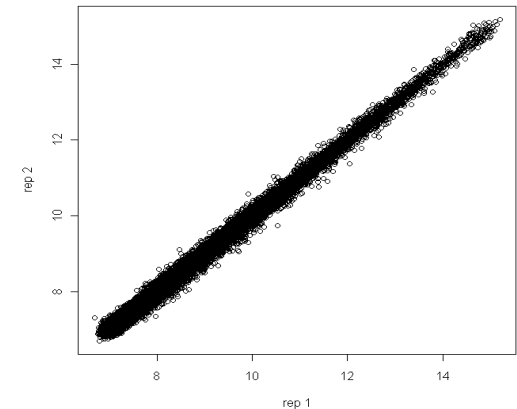
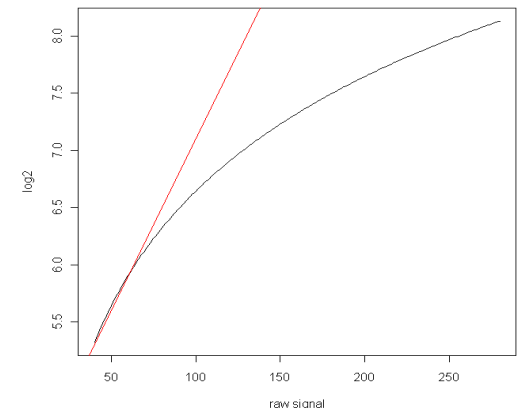
- If array distributions are very dissimilar;
 - Calculate the average distribution function
 - To the probes on both arrays; assign expression level from the average, based on the order of intensity





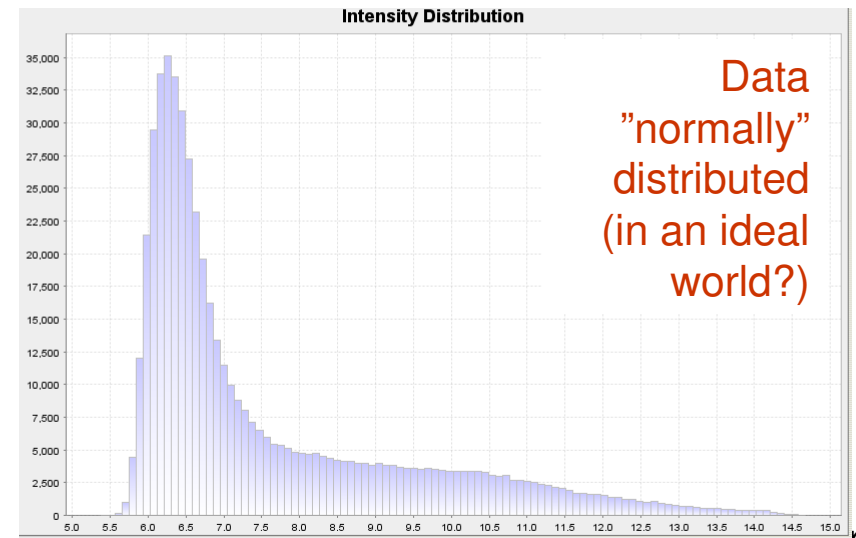
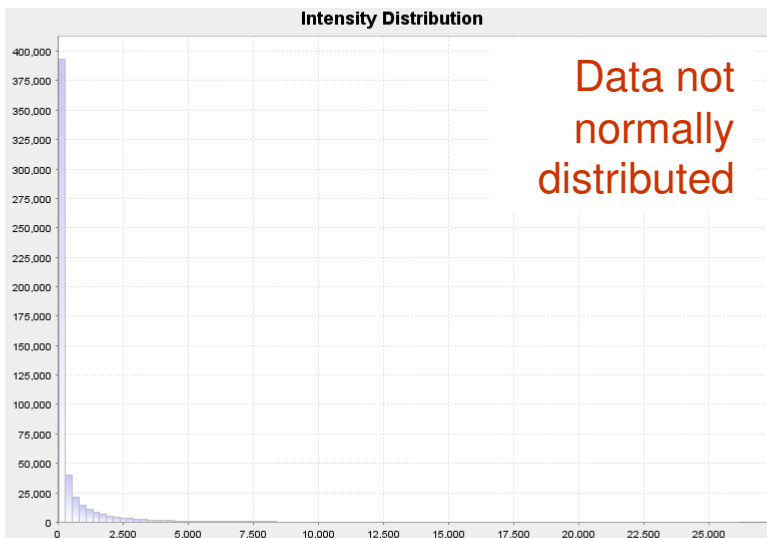
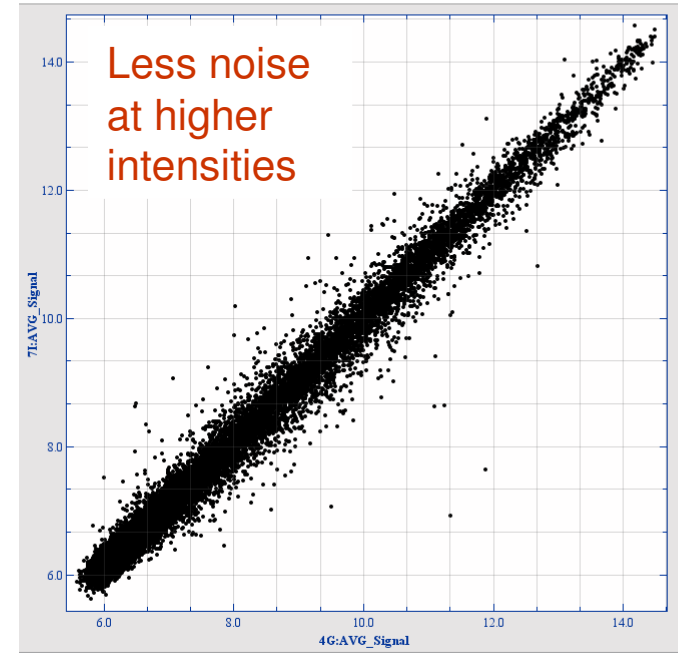
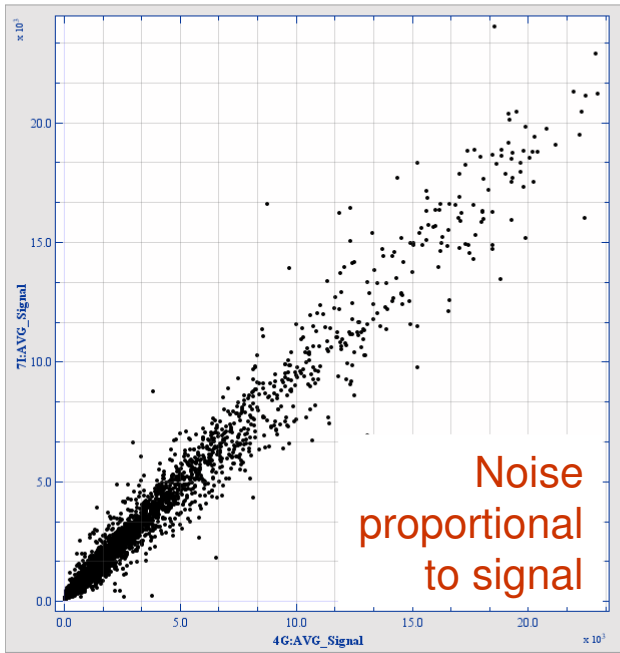
Transformation

- Increasing noise with increasing signal
- Large relative changes in weakly expressed genes hard to detect
- Most of the genes are relatively weakly expressed
- Data not normally distributed



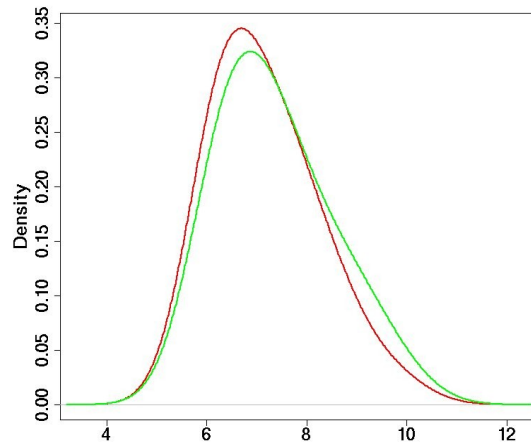


Log transformation

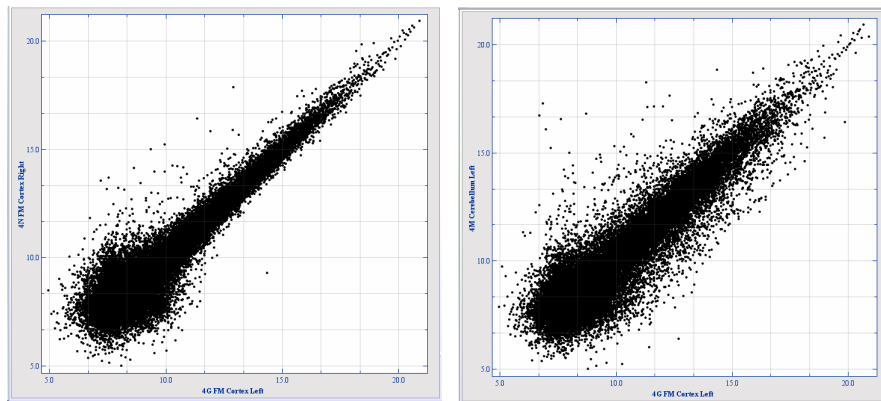




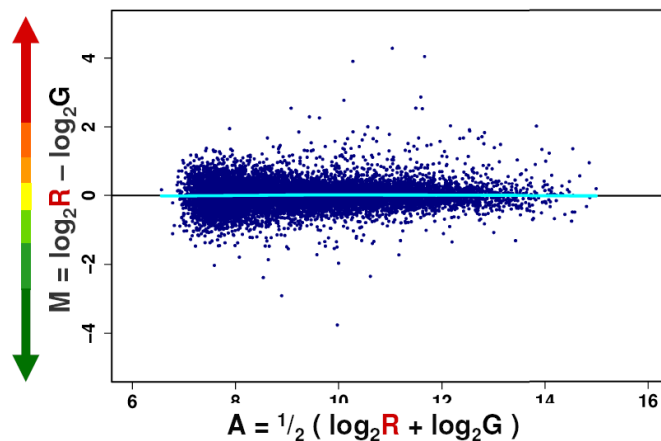
Useful plots



Array plots;
Intensity
distribution
within one array



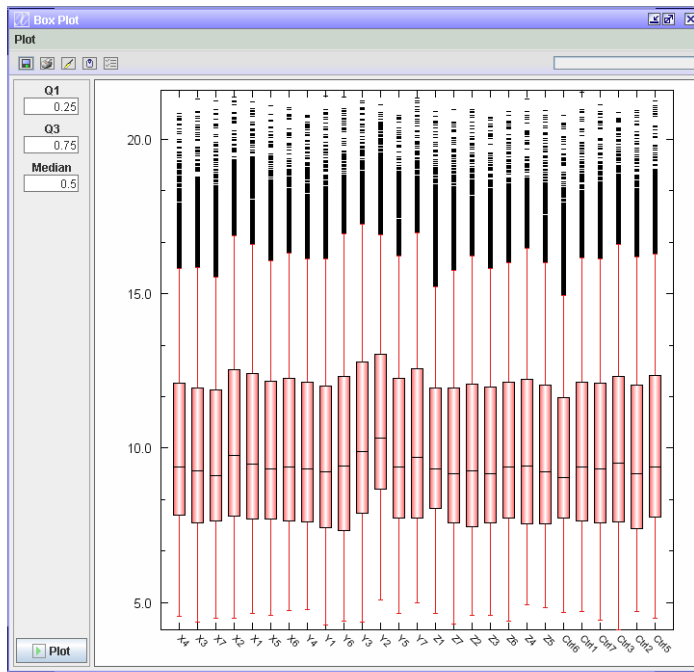
Scatter plots;
Compare intensities
of two arrays or two
channels



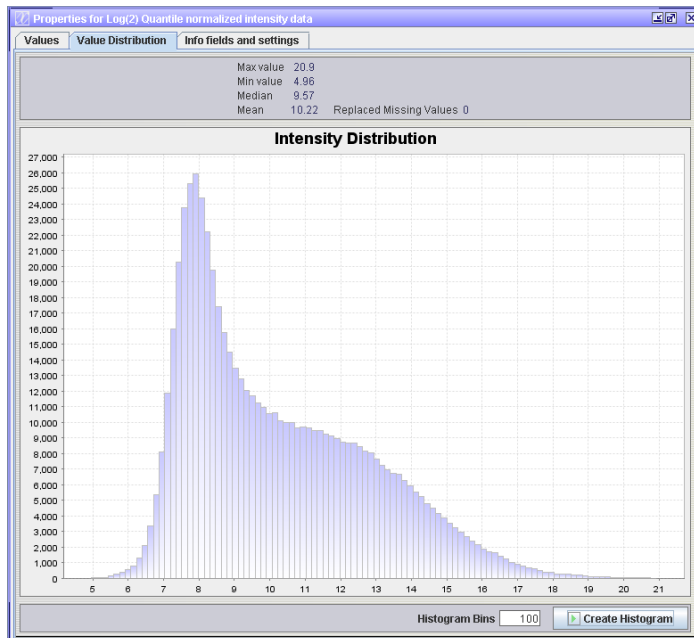
MA plots;
Ratio vs intensity
(2-channel arrays)



More useful plots



Box plots;
Intensity
distributions
across arrays



Density plots;
Intensity
distribution of
experiment





Two-channel normalisation

Reasonable Assumption:

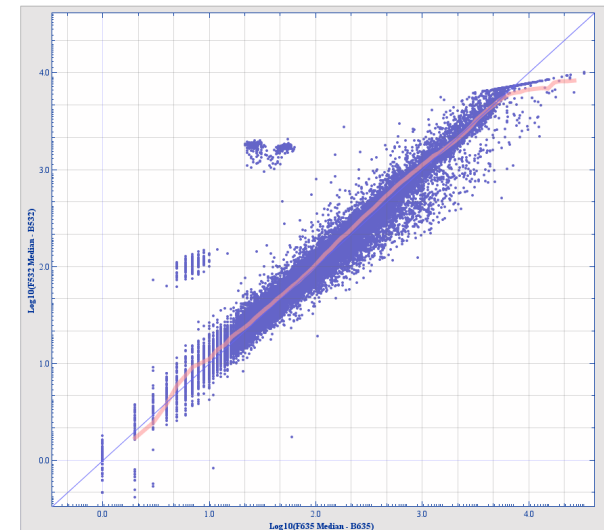
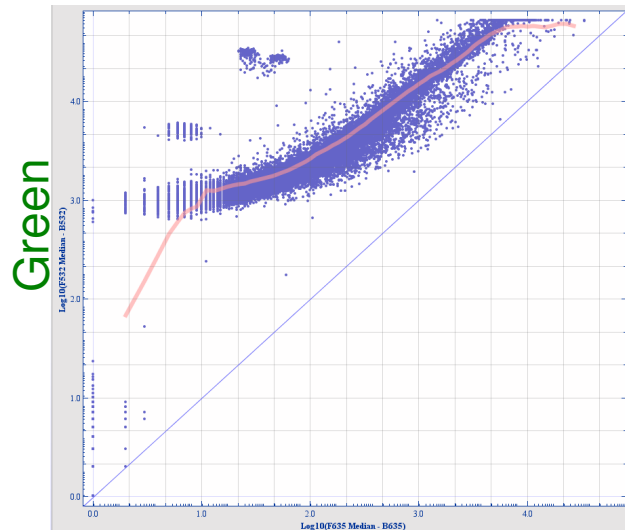
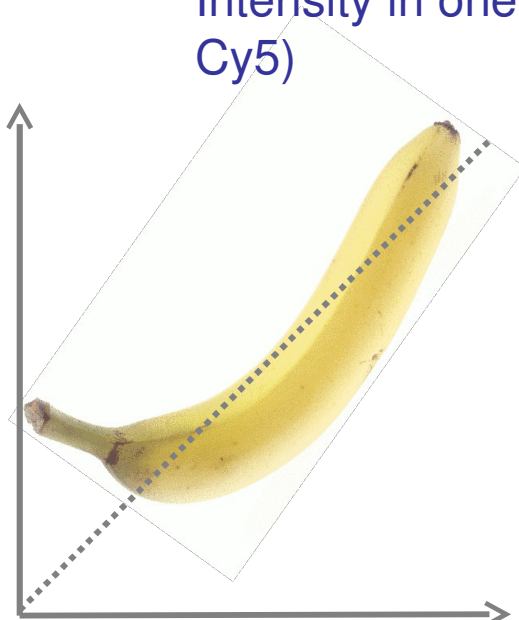
For two colour arrays, in a self self hybridisation, we expect that for each spot, the intensity in the **red** channel = that of the **green** channel

Problem:

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

Dye Bias in Two-channel microarrays:

Intensity in one channel (green; Cy3) may be higher than the other (red; Cy5)





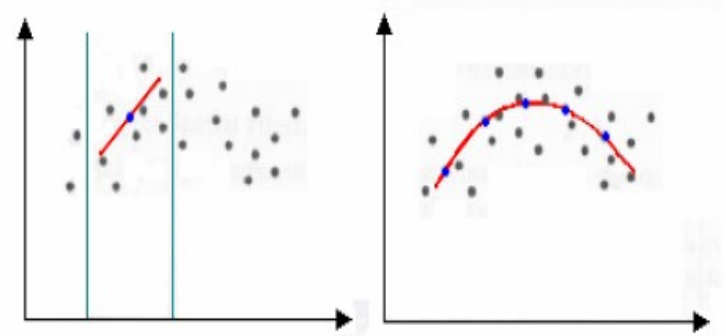
Lowess normalisation

Locally Weighted Least Square Regression

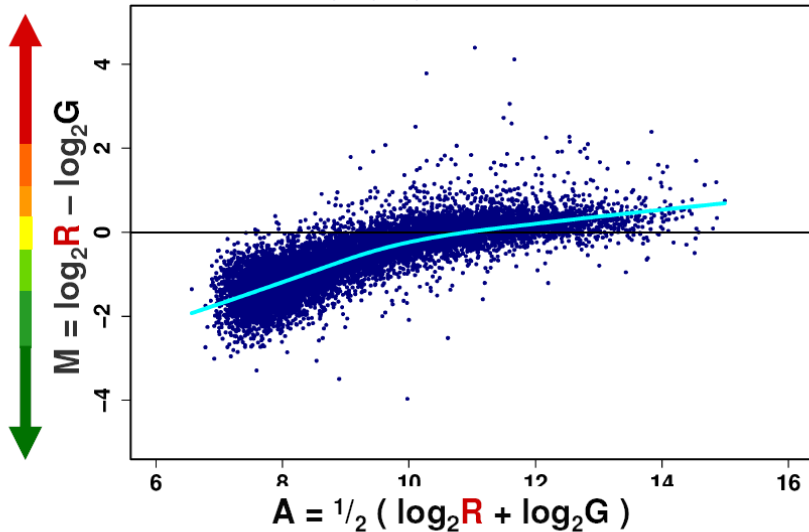
Assumption:
Variation in data is intensity dependent

Smooths the intensity function.

Typically applied to M-A plots



Before



After

