PART II:

Normalisation of microarray data by local regression

Matthias E. Futschik, Dept. of Information Science, Unversity of Otago

29 May 2003

Local Regression

Basic model:

$$Y_i = \mu(x_i) + \epsilon_i \tag{1}$$

Response variable Y is measured with respect to a set of predictor variables **x**. Function μ is locally regressed with independent errors ϵ_i around zero.

Different procedures for local regression exist e.g. LOWESS by Cleveland. Here we used used LOCFIT by C. Loader. The main points are:

- *Evaluation points:* LOCFIT does not perform local regression at every point of the data set, but only at the vertex points of a grid which spans the whole range of variable values $\mathbf{x}_i$.

- *Local polynomial fit:* Quadratic polynomials are locally fitted at the vertex points **z**. In a one-dimensional regression, this leads to the approximation of μ by

$$\mu(z) \approx a_0 + a_1(x - z) + a_2(x - z)^2$$

The neighbouring points x_i are weighted according to the tricube weight function

$$W(x) = (1 - |\frac{x_i - x}{h(x)}|^3)^3$$

with $h(x)$ as the bandwidth which defines the size of the smoothing window. The bandwidth $h(x)$ is the minimal neighbourhood size which includes the fraction α of the total number of points. By choosing α, the user of LOCFIT can determine the smoothness of the fit. The predictor variable **x** might be scaled for multivariate regression.

- *Fitting criteria:* The polynomial coefficients a_i are determined by a local likelihood model. The response variable Y_i is assumed to follow a chosen distribution function. The default distribution in LOCFIT is Gaussian. This leads to a local likelihood criterion that is equivalent to the local least square criterion.

- *Interpolation:* After a local regression is performed for vertex points of the grid, the function μ for a arbitrary point **x** is obtained by interpolation of the function approximations at the vertex points. To ensure that the function μ is globally differentiable, LOCFIT uses a cubic polynomial for interpolation, which includes estimates of the derivatives at the vertices.

Hybridisation Model

To relate the fluorescence signals to changes in gene expression, we introduce a hybridisation model on which we base our normalisation methods. It is especially developed for two-colour arrays consisting commonly of a red (Cy5) and green (Cy3) fluorescence channels. The basic model might, however, be generalised to other types of microarrays.

The fundamental variables in our hybridisation model are the fluorescence intensities of spots in the red (I_r) and the green channel (I_g). These intensities are functions of the abundance of labelled transcripts ($A_{r/g}$). Thus, we have

$$I_{r/g} = f_{r/g}(A_{r/g}, \vartheta) \tag{2}$$

with functions $f_{r/g}$ relating the abundance of the transcripts to the measured intensities and a set of parameters ϑ in the experiment. Note that the functions f_r and f_g might be different.

Under ideal circumstances, this relation of I and A is linear up to an additional experimental error ϵ:

$$I = N(\vartheta)A + \epsilon \tag{3}$$

where N is a normalisation factor and is determined by experimental parameters ϑ such as the laser power or amplification of the scanned signal.

Generally, this simple relation does not hold for microarrays due to effects such as intensity background and saturation. Including an additive background I_b leads to

$$I = NA + I_b + \epsilon = (N + \frac{I_b}{A}) \cdot A + \epsilon = N'A + \epsilon \quad (4)$$

The normalisation factor N' now depends on the intensity. We can obtain the original relation 3 by subtracting the background intensity I_b, so that the background corrected intensity I_{bc} is derived by

$$I_{bc} = I - I_b = NA + I_b + \epsilon - I_b = NA + \epsilon$$

This step is included in most normalisation procedures where the background intensity is estimated by the local background fluorescence surrounding the spot.

Frequently, saturation affects the relation between intensity and RNA abundance. A possible model for these effects is

$$I = \frac{N_1 A}{N_2 A + c} + \epsilon = \frac{N_1}{A(N_2 + c/A)} \cdot A + \epsilon = N'(A)A + \epsilon \quad (5)$$

where N_1, N_2 and c are constants. Although the right-hand side of the equation 5 has the same form as equation 3, the normalisation factor N' is not constant, but varies with the abundance A. Since the saturation is generally of unknown form, the recovery of the original relation between I and A might not be possible.

In a two-colour experiment, ratios of fluorescence intensities are generally used to represent fold changes of gene expression. This procedure has the advantage of controlling for several variations that are inherent to spotted arrays such as size and morphology of the spots, uneven hybridisation and variable amount of spotted DNA. The direct use of fluorescence intensities for analysis is limited, since the rate of incorporation of the dyes in the corresponding cDNA is generally not known. Therefore, fold changes or ratios of gene expressions are the major quantities derived in two-colour experiments. In our hybridisation model, the ratios for labelled transcript abundances (A_r/A_g) are related to the ratios of signal intensities by (I_r/I_g)

$$R = \frac{I_r}{I_g} = \frac{f_r(A_r, \vartheta)}{f_g(A_g, \vartheta)} = \frac{k_r(\vartheta)A_r + \epsilon_r}{k_g(\vartheta)A_g + \epsilon_g} \tag{6}$$

which is based on the equations 2-5. The normalisation factors $k_{r/g}(\vartheta)$ are functions dependent on a set of experimental parameters ϑ. This gives the relation between

the quantities measured (I_r/I_g) and the unknown quantities (A_r/A_g) in which we are interested. Equation 6 can be log_2-transformed to facilitate the computational evaluation and to achieve symmetry between fold changes. This leads to

$$M = \log_2(R) = \log_2(k_r(\vartheta)A_r + \epsilon_r) - \log_2(k_g(\vartheta)A_g + \epsilon_g)$$

To simplify this equation, we use the Taylor expansion

$$\begin{aligned} f(x+\epsilon) &\approx f(x+\epsilon) + \frac{\partial f(x)}{\partial x} \cdot \epsilon \\ \rightarrow \log_2(x+\epsilon) &\approx \log_2(x) + \frac{1}{x \ln(2)} \cdot \epsilon \end{aligned}$$

If the error terms are small and symmetric, we can thus approximate the above equations by

$$\begin{aligned} M &\approx log(k_r(\vartheta)A_r + \epsilon_r) - log(k_g(\vartheta)A_g + \epsilon_g) \\ &\approx (log(k_r(\vartheta)A_r) + \frac{1}{k_r A_r}\epsilon_r) - (log(k_g(\vartheta)A_g) + \frac{1}{k_r A_r}\epsilon_g)) \\ &\approx (log(k_r(\vartheta)) - log(k_g(\vartheta))) + (log(A_r) - log(A_g)) + (\frac{1}{k_r A_r}\epsilon_r - \frac{1}{k_g A_g}\epsilon_g) \\ &\approx \kappa(\vartheta) + D + \tilde{\epsilon} \end{aligned}$$

with $\kappa(\vartheta)$ as additive normalisation factor, D as logged fold changes and $\tilde{\epsilon}$ as the random error. Note that $\kappa(\vartheta)$ can be seen as a term for systematic errors that depends on the experimental settings ϑ.

This results in the final equation:

$$M - \kappa(\vartheta) = D + \tilde{\epsilon} \tag{7}$$

Using this relation, we can derive D from M up to the error term $\tilde{\epsilon}$ once we know the normalisation factor $\kappa(\vartheta)$. The factor $\kappa(\vartheta)$ is generally calibrated by exploiting the relation 7. Depending on the assumptions about the experiment we can proceed with different normalisation methods.

Assuming $\kappa(\theta)$ is constant and the majority of assayed genes are not differentially expressed, the ratios can be linearly scaled to a median value of one*. This leads to *global* or *linear* normalisation. Alternatively, a set of of genes can be selected, which we believe are equally expressed in both samples. The median of these *housekeeping* genes can then be taken to adjust the intensity in both channels by a linear

*The use of median instead of mean is generally preferred for the calculation of a central tendency, since it is less sensitive to outliers.

transformation, so that the intensity medians of the house-keeping genes are the same.

If the factor $\kappa(\theta)$ depends on the fluorescence or laser intensity, it might be derived from the signal ratios assuming a symmetry of the logged fold changes D and error term $\tilde{\epsilon}$. The normalisation factor $\kappa(A)$ can then be locally regressed with respect to the logged signal ratios M. This procedure can be performed using all or a selected subset of genes and is frequently called *intensity-dependent normalisation*. In our hybridisation model, the normalisation factor κ should therefore be a function of A.

$$M_i - \kappa(A_i) = D_i + \epsilon_i$$

If we combine the logged fold change D and error term ϵ_i to the random variable ζ_i which is assumed to be symmetrical distributed around zero, we get

$$M_i = \kappa(A_i) + \zeta_i$$

Since this relation is of the same form as equation 1, we applied a local regression model to capture the intensity dependency of M. The residuals of the regression provided the logged fold changes D up to an error term ϵ_i and were used for the MA- and MXY-plots.

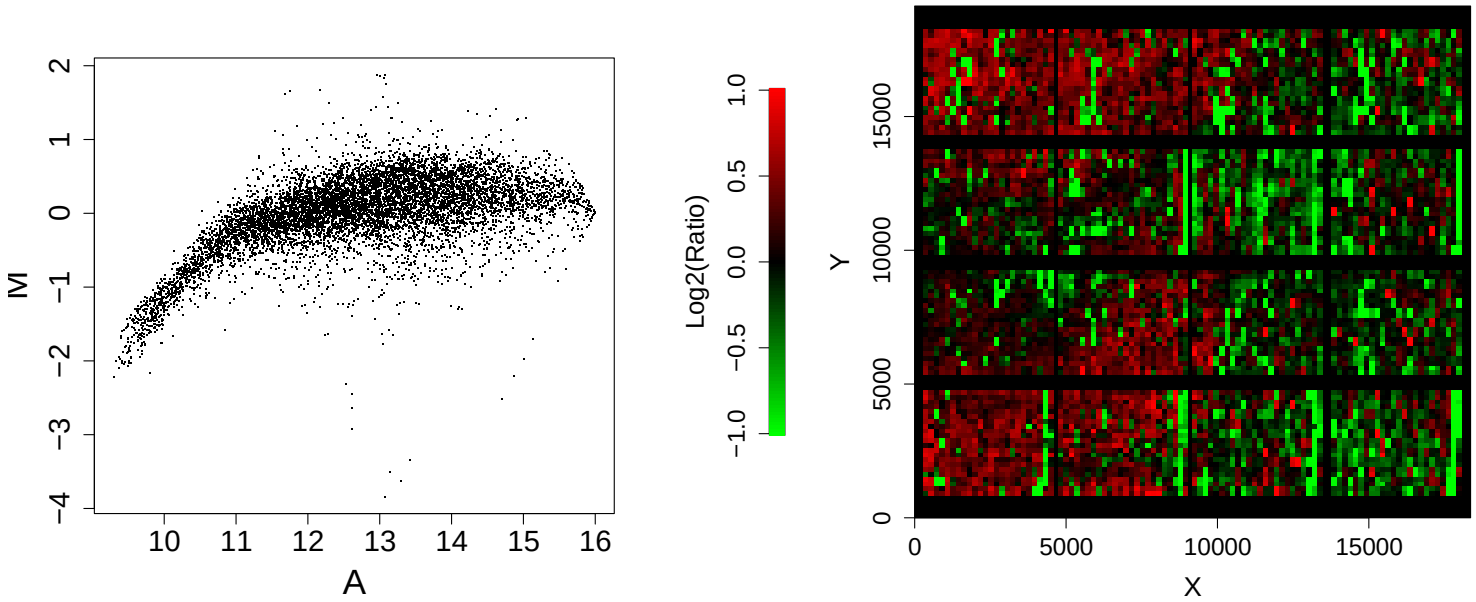
In our studies, we found that the measured spot intensity ratios showed not only an intensity, but also a spatial, bias across the array. We introduce, therefore, a normalisation procedure that simultaneously corrects for bias due to intensity and spatial location. To cope with spatial effect, we included spot location in the regression analysis. The normalisation factor κ is then a function of spot intensity A as well as location (X, Y):

$$M_i - \kappa(A_i, X_i, Y_i) = D_i + \epsilon_i \rightarrow M_i = \kappa(A_i, X_i, Y_i) + \zeta_i$$

Following the same procedure as before, we locally regressed M by the function $\kappa(A, X, Y)$. Since the variables A, X and Y have different ranges of values, we scaled them by their standard deviation. For parameter α, the default value of 0.5 was used.

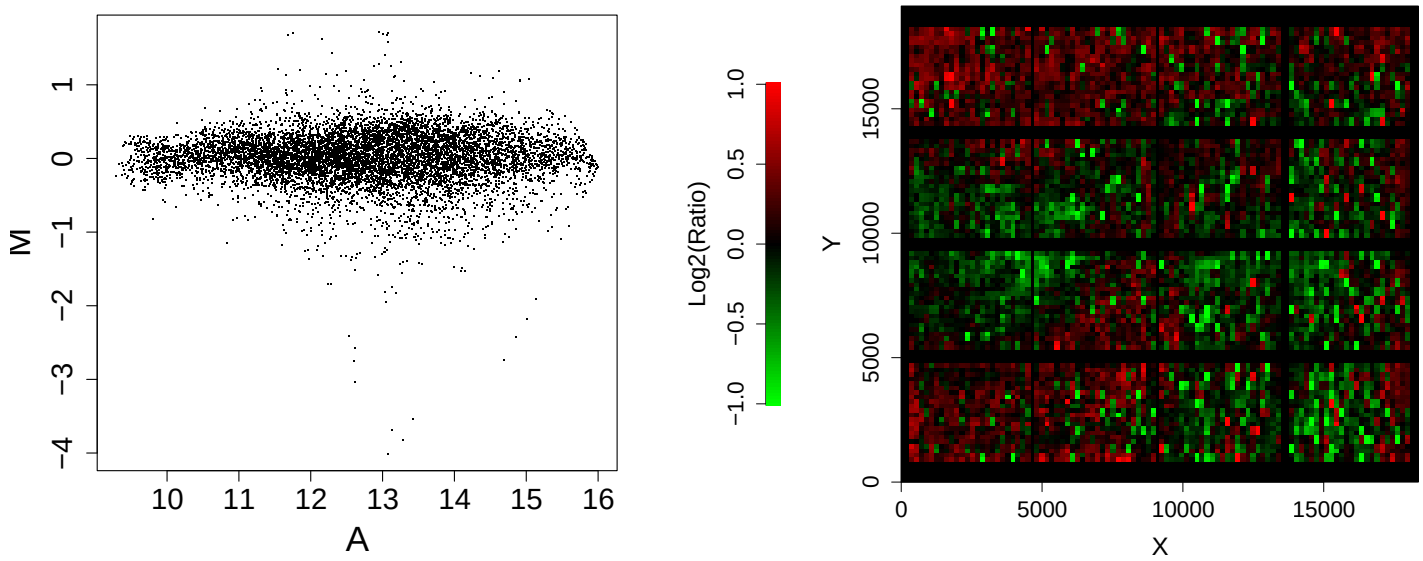
The basic assumptions are overall symmetry of fold changes and random spatial order of spots.

Distribution of ratios for raw data



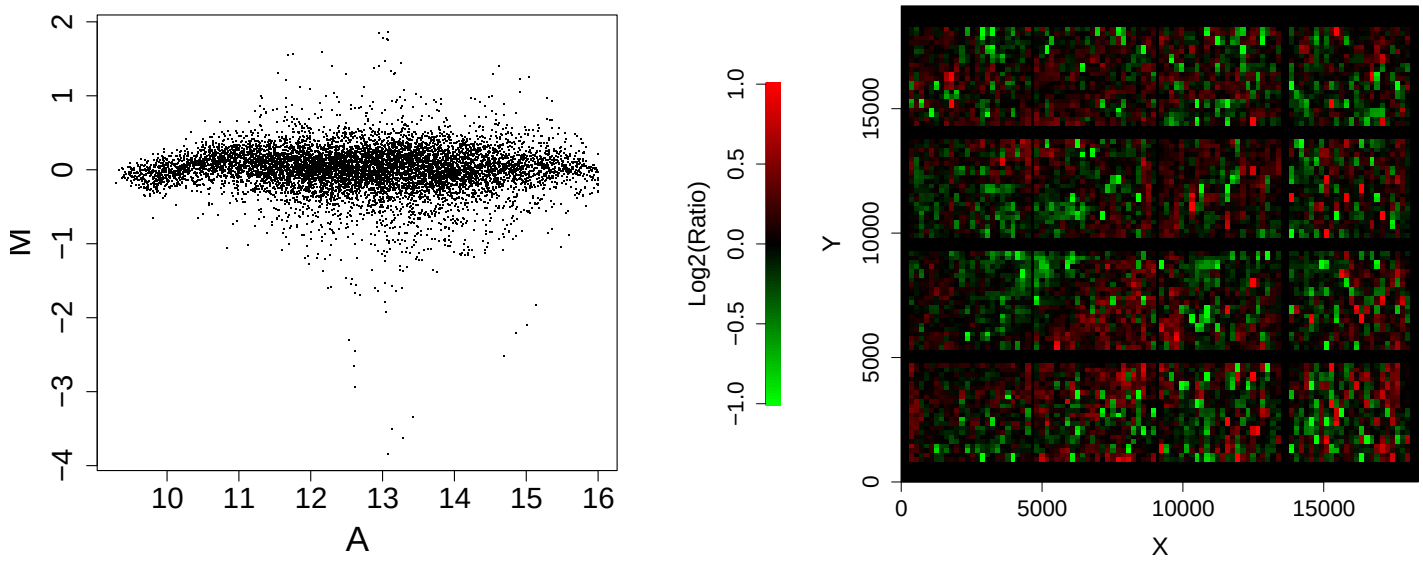
a) The MA-plot indicates intensity bias low intensities and saturation effects.
b) The MXY-plot shows uneven distribution of logged ratios. The rows with consistently negative M are rows of empty control spots.

Distribution of ratios after intensity-dependent normalisation



a) Residuals of the local regression are well balanced around zero in MA-plot. b) Spatial bias is still apparent in the MXY-plot, while the lines of negative M corresponding to empty spots disappeared due to the intensity-dependent normalisation.

Distribution of ratios after local intensity-dependent normalisation



a) The MA-plot is similar to the one for intensity-depend normalisation. b) The MXY-plot shows a reduced spatial bias compared to the MXY-plot for intensity-dependent normalised data.

Statistical Testing of Normalisation

INTENSITY BIAS

- **ANOVA** The observed logged ratios M were divided into 10 groups of equal size based on the ranking of corresponding spot intensities. These intensity intervals defined the factor levels in a single factor ANOVA.

- **Kruskal-Wallis test** To avoid the assumption of data normality, we employed a Kruskal-Wallis test for assessing the equality of the median logged ratio for intensity intervals defined above.

SPATIAL BIAS

- **ANOVA** A natural factor the spatial location on the array is the pin-number, since spots printed by the same pin form a localised block of pins. As 16 pins were used for the spotting, an ANOVA model with 16 factor levels can be defined. It assesses the equality of the average M printed by different pins. Additionally, we performed a one-sample t-test for each block to localise possible pin-effects.

- **Kruskal-Wallis test** The non-parametric Kruskal-Wallis test assesses the equality of median M of spots printed by different pins.

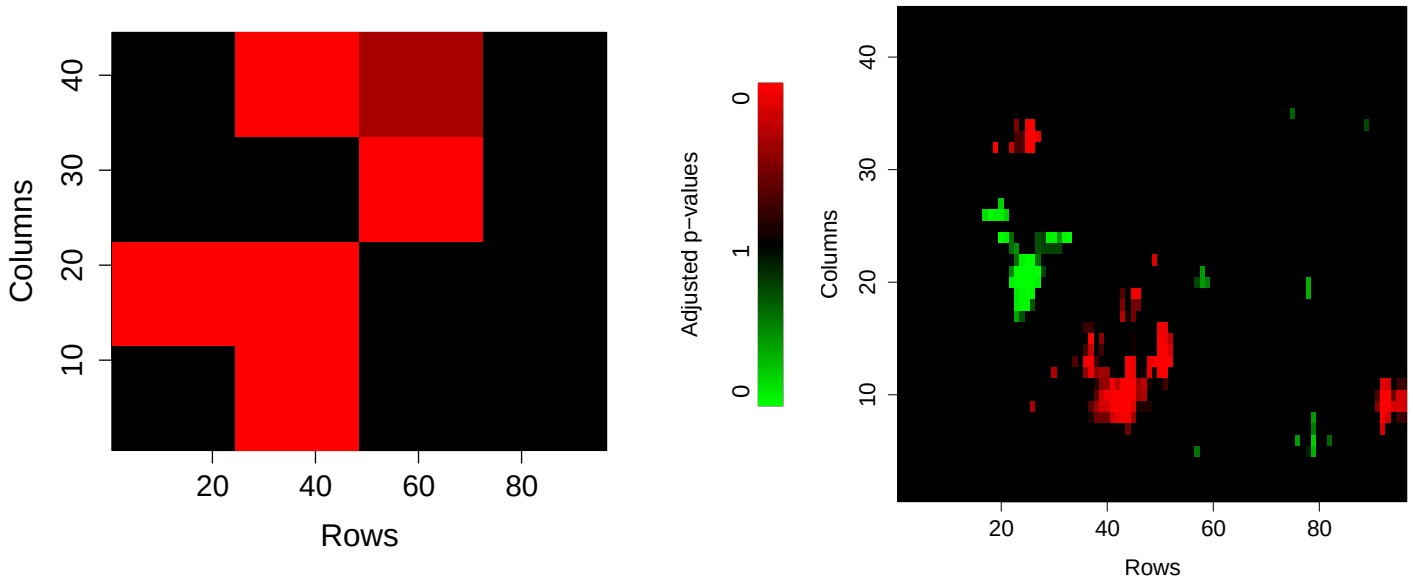
- **Permutation test** The ANOVA and Kruskal-Wallis tests applied here primarily aim to detect bias due to the use of different pins in the spotting procedures. To test for spatial bias independently of the print layout by pins, we employed a permutation test:

  - *Null hypothesis*: H_0 states a random location of the observed logged ratios.

  - *Test statistic* T: Median logged ratio value in the spot's neighbourhood (5x5 spots).

  - *Test distribution:* Empirical distribution was derived from randomly permutated spot locations. The number of permutations was 10^6. To accommodate for the multiple testing procedure, we adjust the p-values using the Bonferroni correction. Note that this leads to rather conservative estimates, since the single tests are not independent. Spots next to each other share the majority of neighbouring spots, so that the median M is likely to be highly correlated.

Statistical significance of spatial ratio distribution



a) The results of the t-test for subarrays (light red: $\alpha = 0.05$, dark red: $\alpha = 0.01$). b) Spots with significant neighbourhoods based on Bonferroni adjusted p-values of the permutation test. The significance for positive and negative M are presented with red and green colouring.

Results of statistical tests

TESTS FOR INTENSITY BIAS

ANOVA				*Kruskal-Wallis test*	
Method	R^2	*p*	*Sig. intervals*	χ^2	*p*
Raw	0.42	$< 10^{-15}$	-/-/-/9	1412.12	$< 10^{-15}$
Linear	0.41	$< 10^{-15}$	-/-/-/8	1389.33	$< 10^{-15}$
Intensity	0.00	0.01	-/1/-/1	35.9	$4.0 \cdot 10^{-5}$
Local int.	0.01	$1.1 \cdot 10^{-8}$	-/1/1/1	84.6	$1.9 \cdot 10^{-14}$
Local int. opt.	0.00	0.62	-/1/-/-	12.84	0.16

ANOVA and Kruskal-Wallis test were based on 10 consecutive intensity intervals. The R^2 value gives the percentage of the total variation which is attributed to the intensity variation by the ANOVA model. The number of significant intervals is derived by the t-test for the significance levels: 0.1, 0.05, 0.01, 0.001.

TESTS FOR SPATIAL BIAS

ANOVA				*Kruskal-Wallis test*		*Corr.*
Method	R^2	*p*	*Sig. blocks*	χ^2	*p*	*r*
Raw	0.11	$< 10^{-15}$	1/2/2/8	740.1	$< 10^{-15}$	0.39
Linear	0.11	$< 10^{-15}$	-/-/1/11	732.8	$< 10^{-15}$	0.39
Intensity	0.10	$< 10^{-15}$	-/1/2/10	675.2	$< 10^{-15}$	0.40
Local int.	0.01	$1.0 \cdot 10^{-5}$	-/1/5/-	97.9	$3.2 \cdot 10^{-14}$	0.24
Local int. opt.	0.00	0.78	-/-/-/-	28.56	0.02	0.13

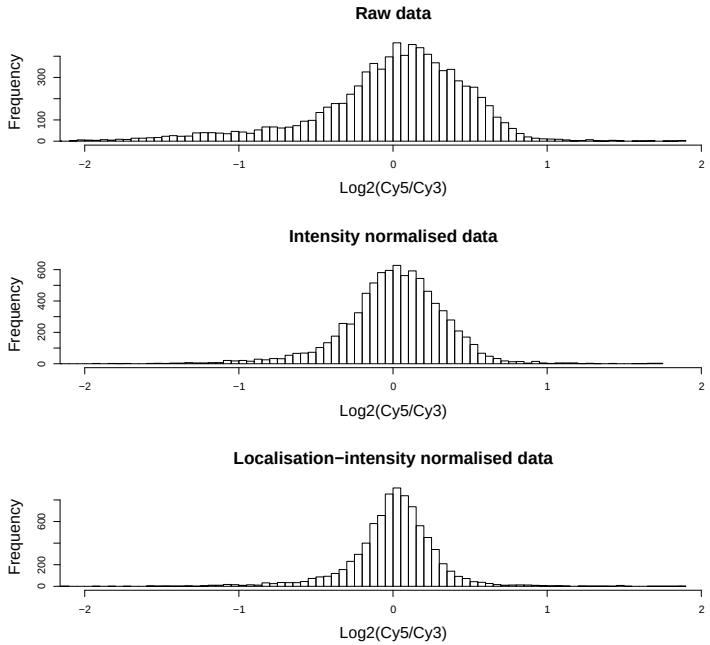
ANOVA and Kruskal-Wallis test are based on blocks of spots printed by the same pin. For the ANOVA test, the number of significant blocks is given for significance levels of 0.1, 0.05, 0.01 and 0.001. For the ANOVA model, R^2 gives the percentage of the total variation attributed to the spatial location. The coefficient r describes the correlation between the logged ratio M of the spot and the median value of M within a neighbourhood window of 5x5 spots.

PERMUATION TEST RESULTS

	Unadjusted p			*Bonferroni adjusted* p_B		
Methods.	< 0.1	< 0.05	< 0.01	< 0.1	< 0.05	< 0.01
Raw	1349	1214	934	247	216	117
Linear	1341	1194	927	235	200	112
Intensity	1248	1071	715	212	185	106
Loc. int.	972	708	377	88	72	38
Loc. int. opt.	729	486	183	7	3	1

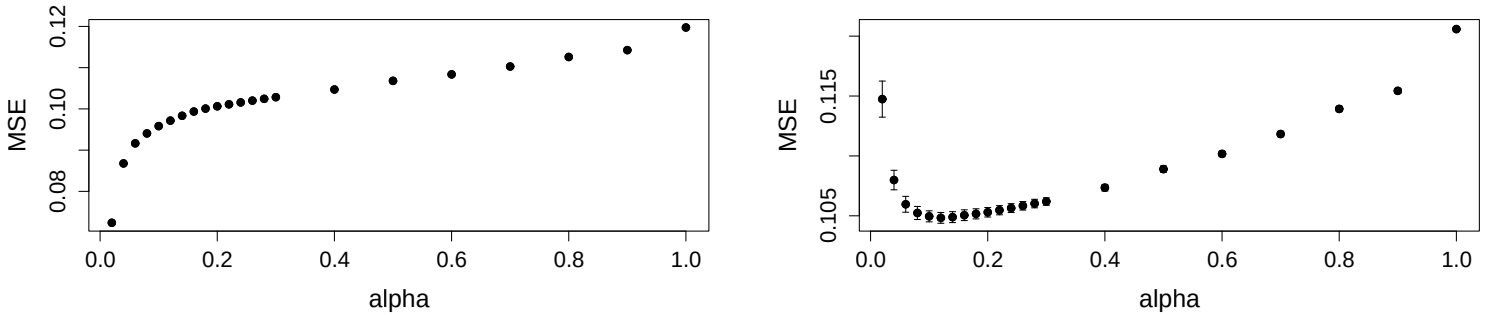
Number of significant spot neighbourhoods in permutation test for different normalisation schemes: Left side gives the number of 5x5 spot neighbourhoods which yielded a significant median for logged ratio M. Numbers on the right side of the table were derived for Bonferroni corrected p-values ($p_B = p/4224$).

Histograms of logged ratio distributions



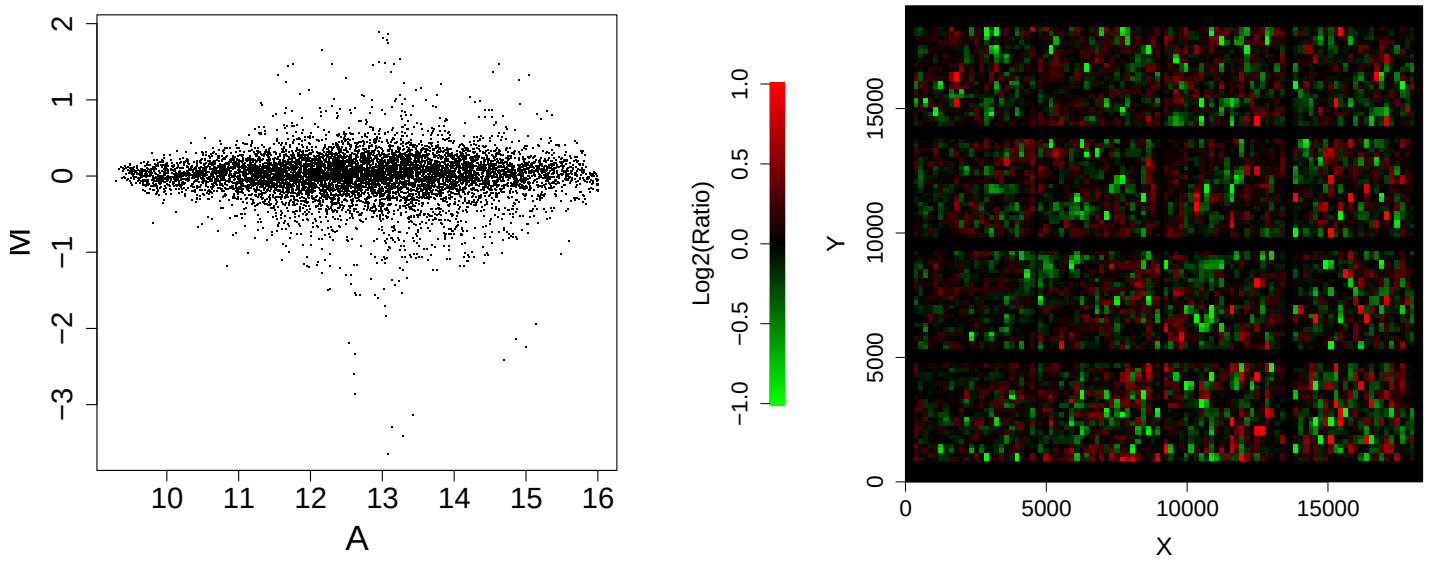
Histograms of distributions of M for raw data, for intensity-dependent normalised data and local intensity-dependent normalised data.

Model selection



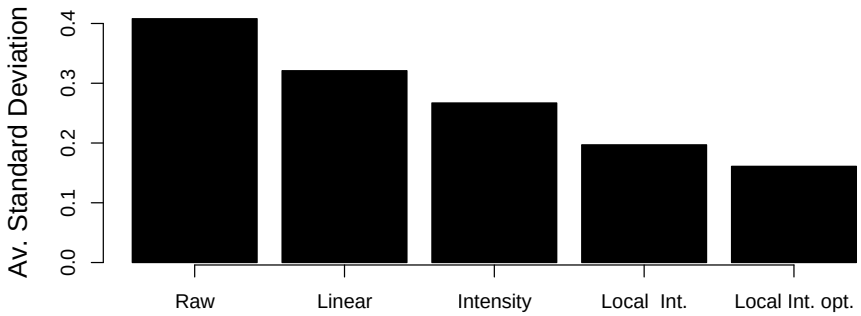
5-fold crossvalidation: Standard error of ratios with respect to α parameter. Top figure presents the average error for the training set, whereas bottom figure shows the average error for the test set. An optimal regression is achieved for $\alpha = 0.12$.

Optimised local intensity-dependent normalisation



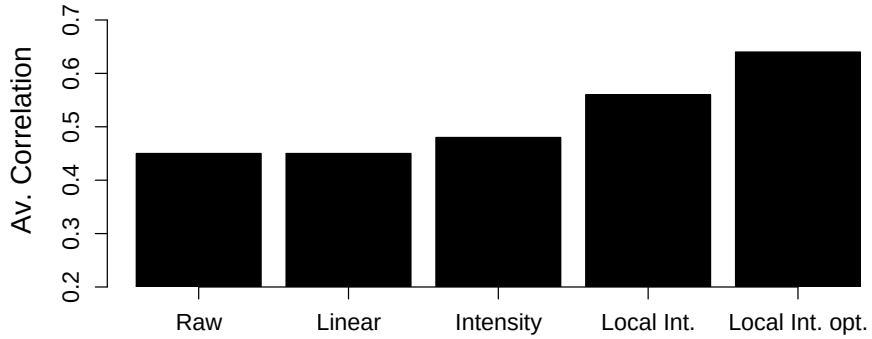
Optimised local intensity-dependent normalisation for parameter $\alpha = 0.12$: Both plots show no apparent bias for logged ratios M with respect to the intensity A or the spot location (X,Y).

Comparison of standard deviations of logged ratios



Average standard deviation σ of logged ratios M across four replicated microarrays for different normalisation methods.

Comparison of correlation between replicate microarrays



Average correlation $\bar{r}$ of logged ratios M between four replicated microarrays for different normalisation methods