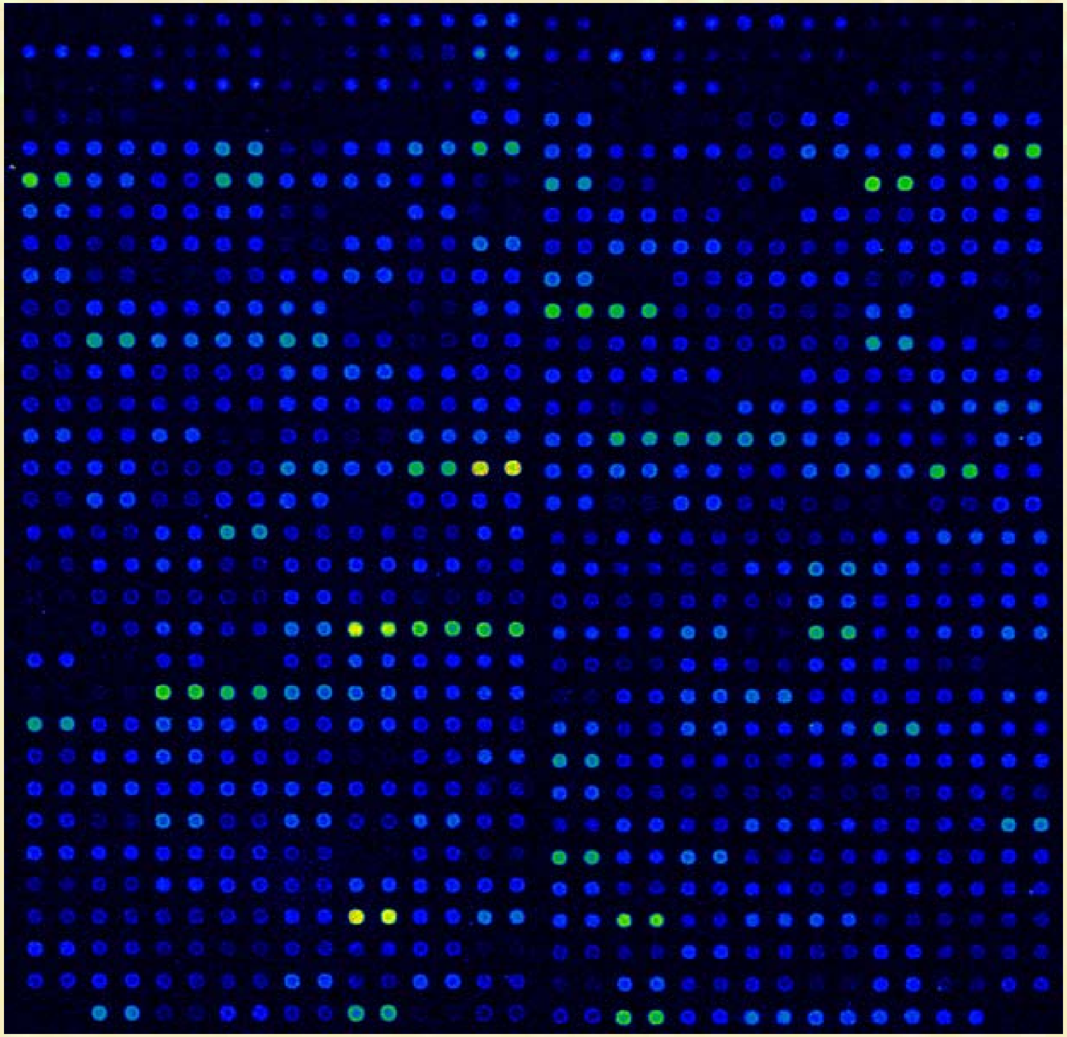
Microarray normalization

Priit Palta

bioinfo Journal Club
19.12.05

Microarrays are nice and good, but…



Can you see papers hidden among these spots?

Microarray data manipulation

1. Image analysis
2. **Normalization**
3. Data analysis

Microarray normalization

- Image segmentation
- PCR yield, contamination
- Signal quantification
- Amplification efficiency
- Spatial effects
- Spotting efficiency
- ‘Background’ correction
- Hybridization efficiency and specificity
- Issues related to array manufacturing

Microarray normalization

- Normalization methods:
  - – Slide normalization
  - – Pin group normalization
- Current methods can deal with:
  - – Intensity (channel, dye) bias
  - – Spatial bias

Methods for intensity and spatial normalization

- Global median normalization – ***gMed*** (Quackenbush 2002):
- Estimate is the median of M values in the array

$$\boldsymbol{M^{*} = M - {}^{\wedge}M}$$

M^{*} is corrected intensity

M is log-intensity ratio ($M = log_2\ R/G$)

$^{\wedge}M$ is constant for one slide (grid), taken as the median of M values in the slide

Methods for intensity and spatial normalization

- Print tip (subarray) loess (lowess, locally weighted scatterplot smoothing) – ***pLo*** (Dudoit *et al.*, 2002):
- Estimate is loess fit *M(A)* within each print tip group

$$\boldsymbol{{}^\wedge M = c_i(A)}$$

A is average log-intensity ($A = (1/2)0.5 \log_2 R*G$)

Methods for intensity and spatial normalization

- Print tip loess with median background correction – ***pLoGS*** (www.biodiscovery.com):
- Estimate is loess fit *M(A)* within each print tip group but the background correction is local group (over a 3 x 3 square)

$$\boldsymbol{\hat{}M = c_i(A)}$$

Methods for intensity and spatial normalization

- Composite of print tip lowess and 2D normalization – ***cPLo2D*** (Yang *et al.*, 2002):
- Estimate is equally weighted sum of loess fit *M(SpotRow, SpotCol)* and loess fit *M(A)* within each print tip group

$$\hat{M} = \alpha * c_i(A) + \beta * c_i(SpotRow, SpotCol)$$

$c_i(SpotRow, SpotCol)$ is the loess estimate of *M* using spot row and spot column coordinates inside the *i*-th print tip as predictors

Methods for intensity and spatial normalization

- Global loess normalization followed by a spatial median filter – ***gLoMedF*** (Wilson *et al.*, 2003):
- Estimate is loess fit M(A) on the whole array plus the median of a 3 x 3 block of spots with the current spot in the center

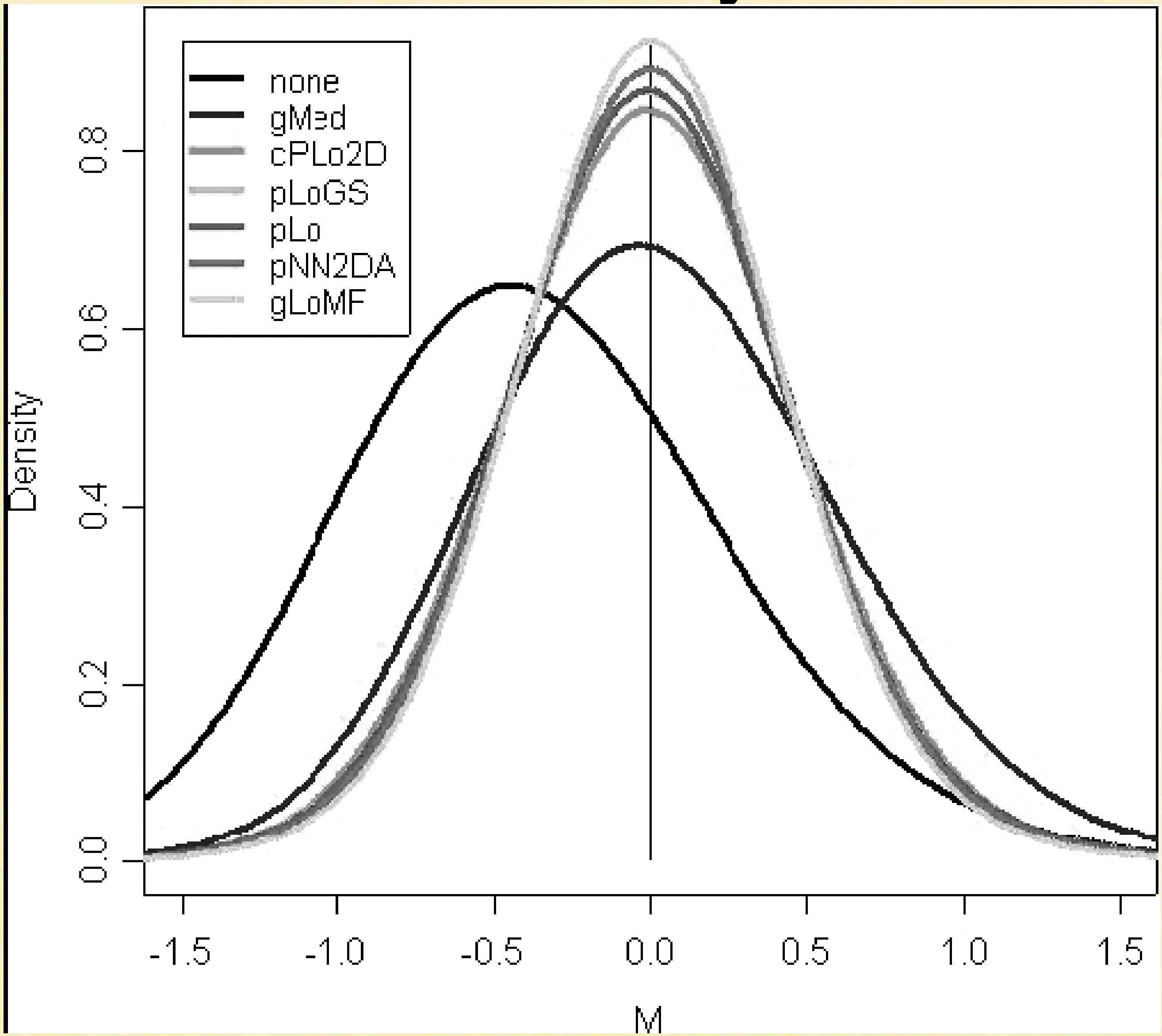
Methods for intensity and spatial normalization

- Neural networks based spatial and intensity normalization - ***pNN2DA*** (Tarca *et al.*, 2005):
- Estimate is the robust neural network fit for *M(A, X,Y)* within each print tip, where X and Y are obtained by binning the print tip space

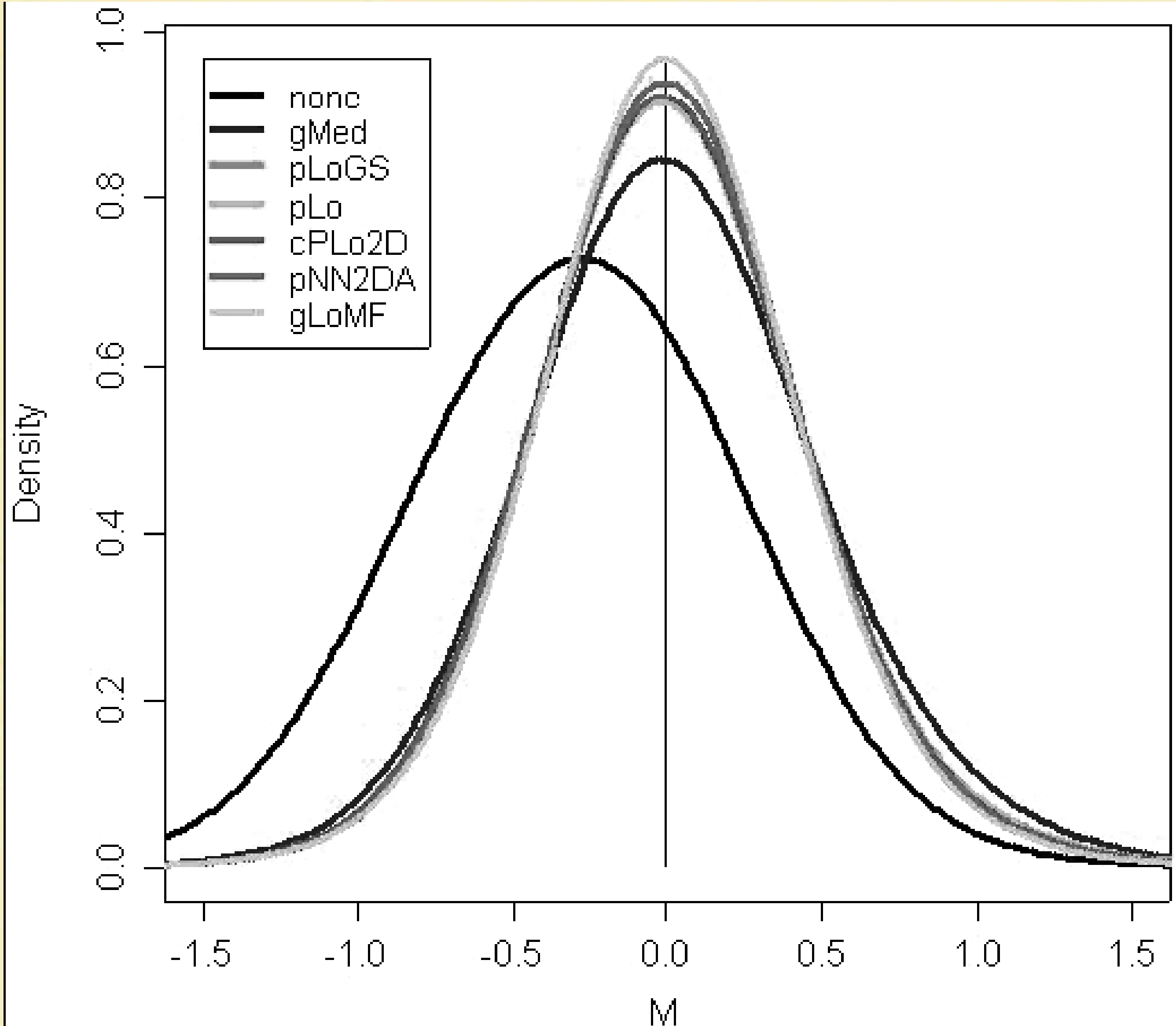
$$\hat{} M = c_i(A, X, Y)$$

c for every print tip *i* in a slide is a trained neural network function

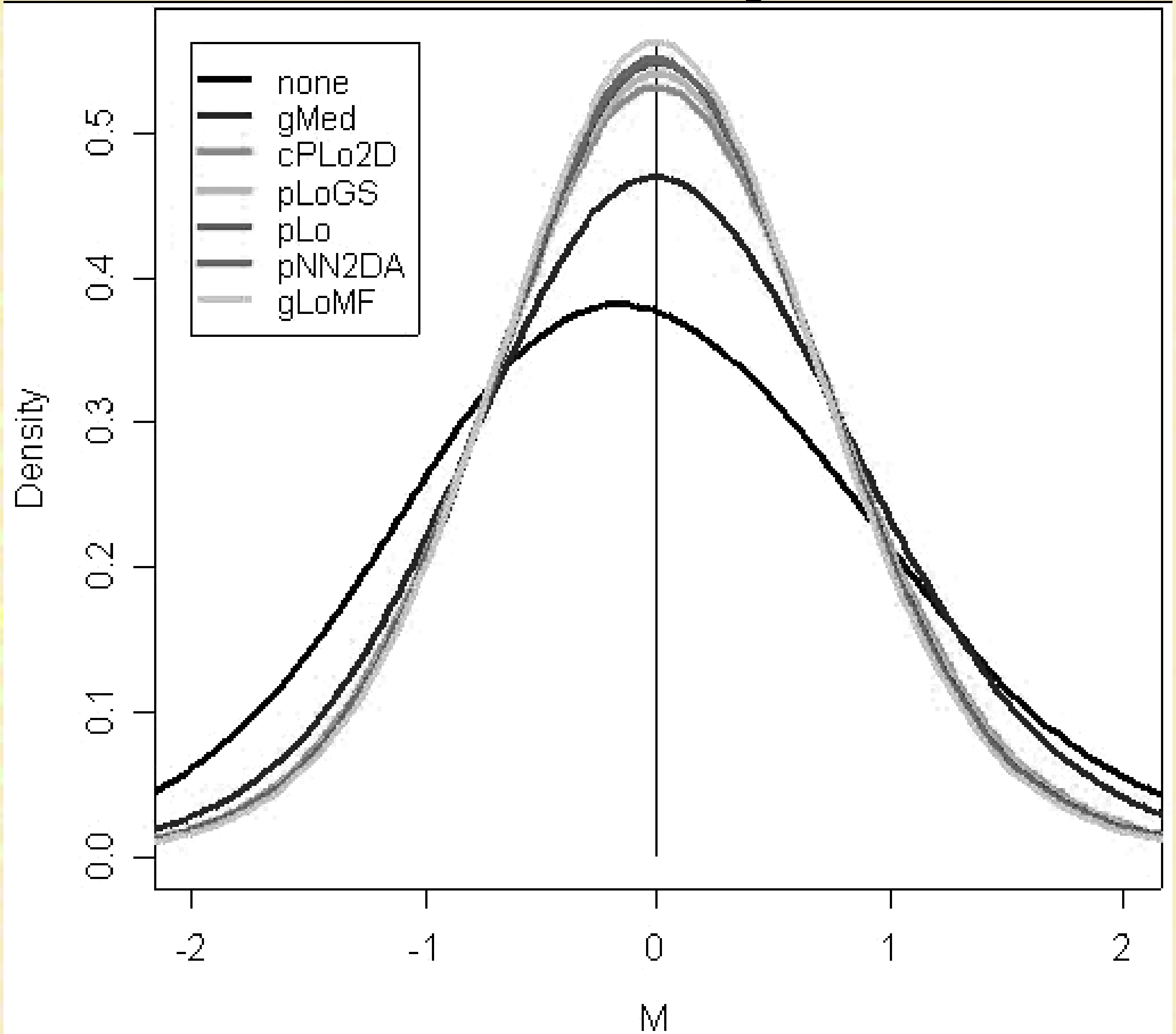
Comparison results: impact on variability



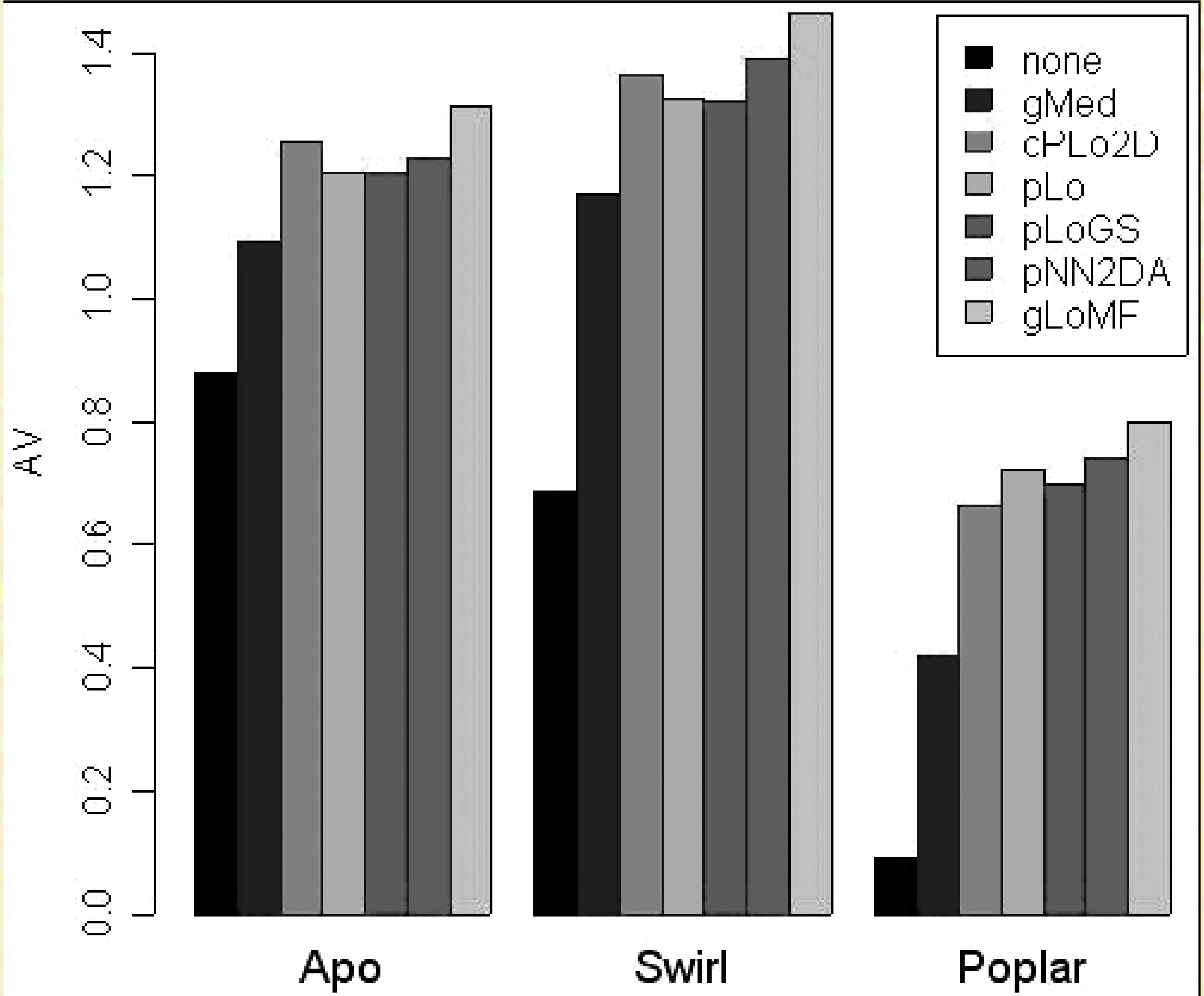
Comparison results: impact on variability



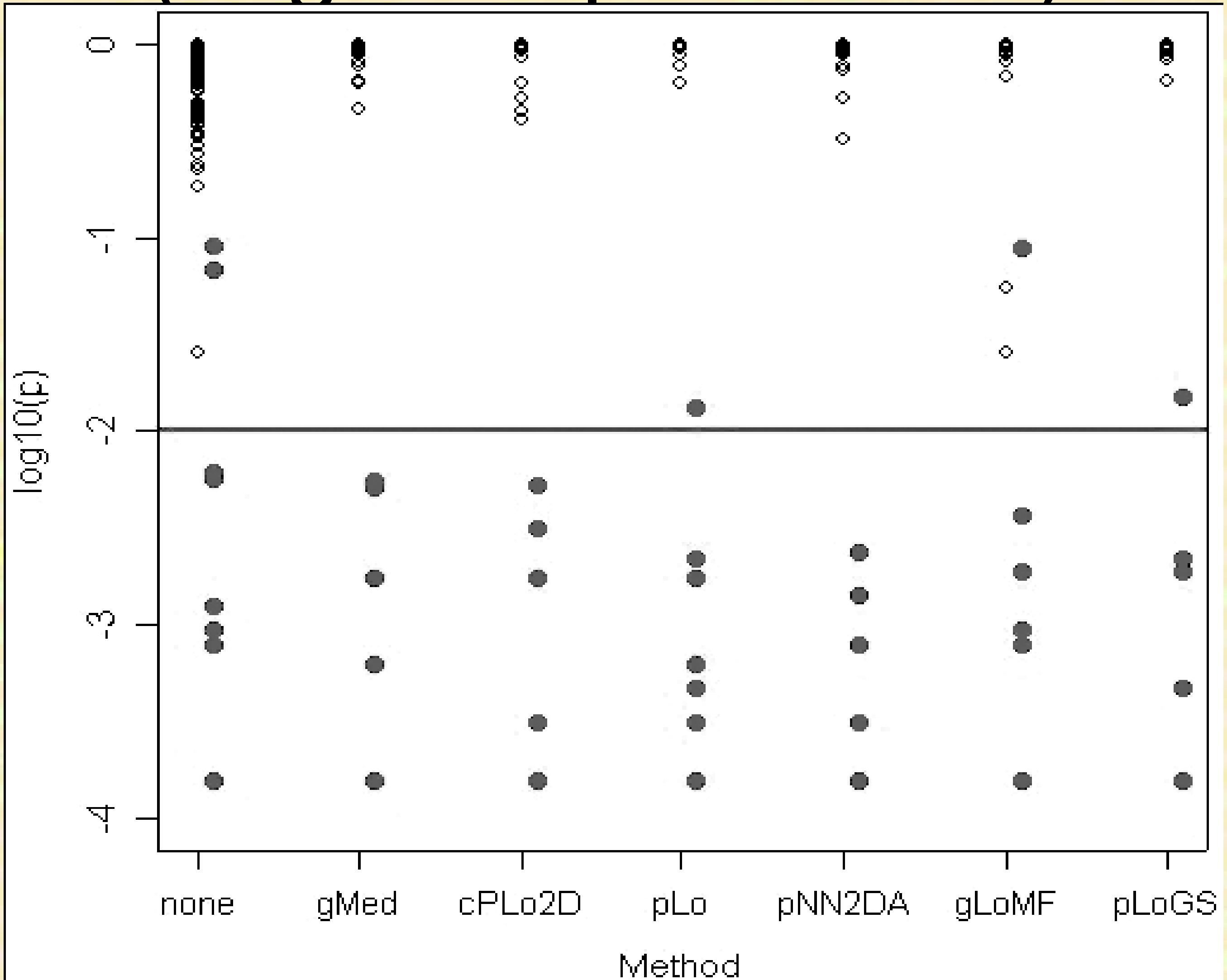
Comparison results: impact on variability



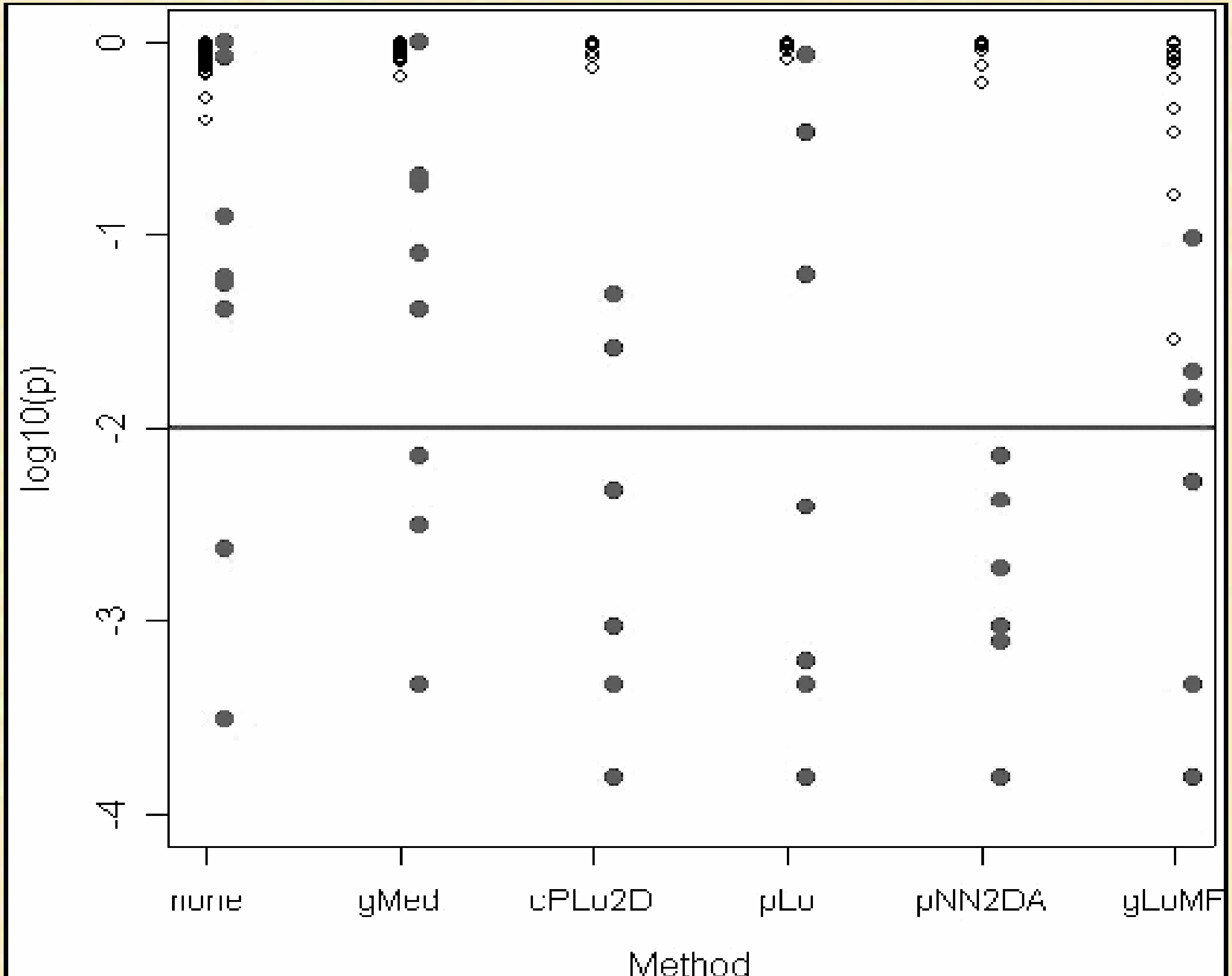
Comparison results: between replicates variability



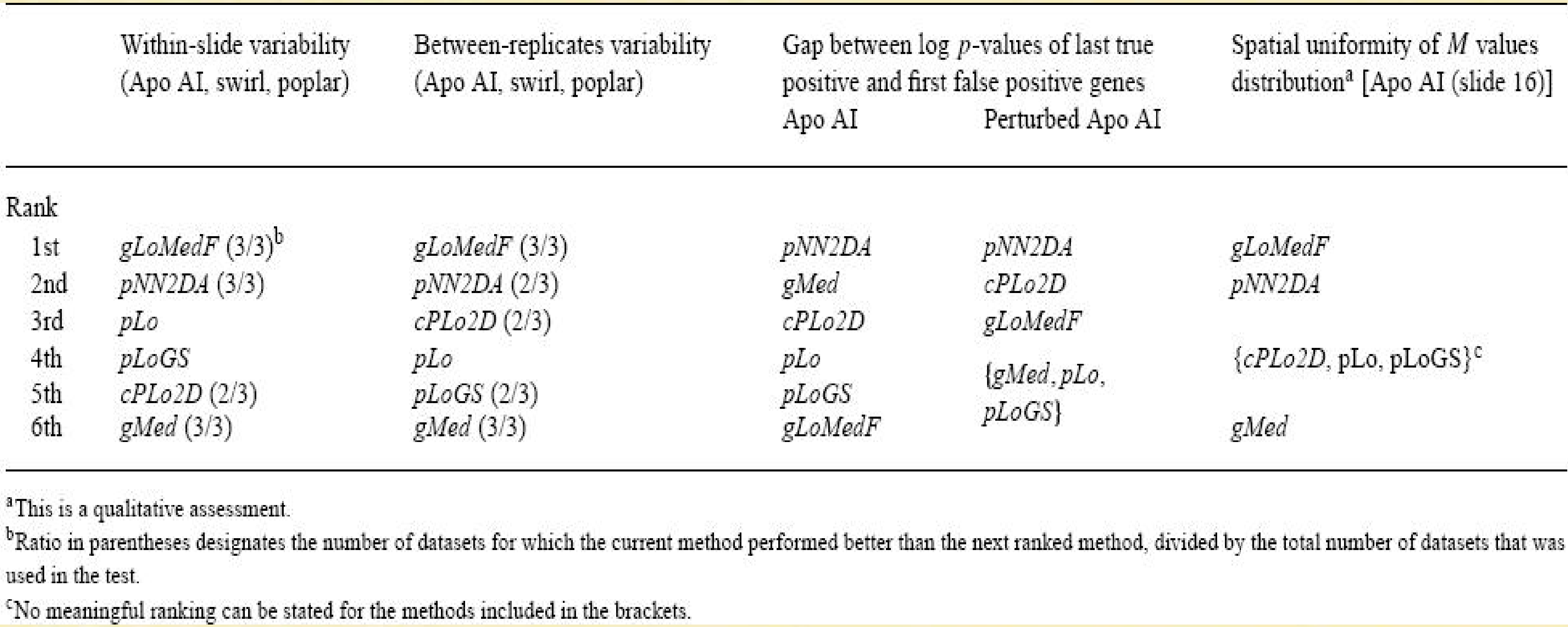
Comparison results: impact on differential expression statistics (original Apo dataset)



Comparison results: impact on differential expression statistics (perturbed Apo dataset)



Comparison results: ranking of methods according to multiple criteria



	Within-slide variability (Apo AI, swirl, poplar)	Between-replicates variability (Apo AI, swirl, poplar)	Gap between log *p*-values of last true positive and first false positive genes		Spatial uniformity of *M* values distribution[a] [Apo AI (slide 16)]
			Apo AI	Perturbed Apo AI	
Rank					
1st	*gLoMedF* (3/3)[b]	*gLoMedF* (3/3)	*pNN2DA*	*pNN2DA*	*gLoMedF*
2nd	*pNN2DA* (3/3)	*pNN2DA* (2/3)	*gMed*	*cPLo2D*	*pNN2DA*
3rd	*pLo*	*cPLo2D* (2/3)	*cPLo2D*	*gLoMedF*	
4th	*pLoGS*	*pLo*	*pLo*	{*gMed*, *pLo*, *pLoGS*}	{*cPLo2D*, pLo, pLoGS}[c]
5th	*cPLo2D* (2/3)	*pLoGS* (2/3)	*pLoGS*		
6th	*gMed* (3/3)	*gMed* (3/3)	*gLoMedF*		*gMed*

[a] This is a qualitative assessment.

[b] Ratio in parentheses designates the number of datasets for which the current method performed better than the next ranked method, divided by the total number of datasets that was used in the test.

[c] No meaningful ranking can be stated for the methods included in the brackets.

Global loess normalization followed by a spatial median filter

(Wilson *et al.*, 2003)

A loess curve *c(A)* is first computed for the whole slide and subtracted from the raw *M* values. Then, a median filter is applied on the residuals to estimate the spatial trend. The median filter simply subtracts from each residual the median of residuals over its spatial neighborhood (a 3 x 3 block of spots with the current spot in the center).

Global loess normalization followed by a spatial median filter

(Wilson *et al.*, 2003)

1. Transform the logarithmic data to the mean vs. difference scale
2. Fit a single loess curve to the transformed data
3. Calculate the residuals from the curve fit

Global loess normalization followed by a spatial median filter

(Wilson *et al.*, 2003)

4. Spatially smooth the residuals with a median filter to estimate the spatial trend
5. Compute the residuals from the spatial trend estimate
6. Rescale the data by dividing through by an estimate of the median absolute deviation (calculated on the final residual mean-difference data)

Neural network based spatial and intensity normalization

(Tarca *et al.*, 2005)

- The objective is to find the best fit of *M* values within a print tip group using the average log-intensity *(A)* as well as the two-dimensional space coordinates of the spots *(X, Y)* as predictors

Neural network based spatial and intensity normalization

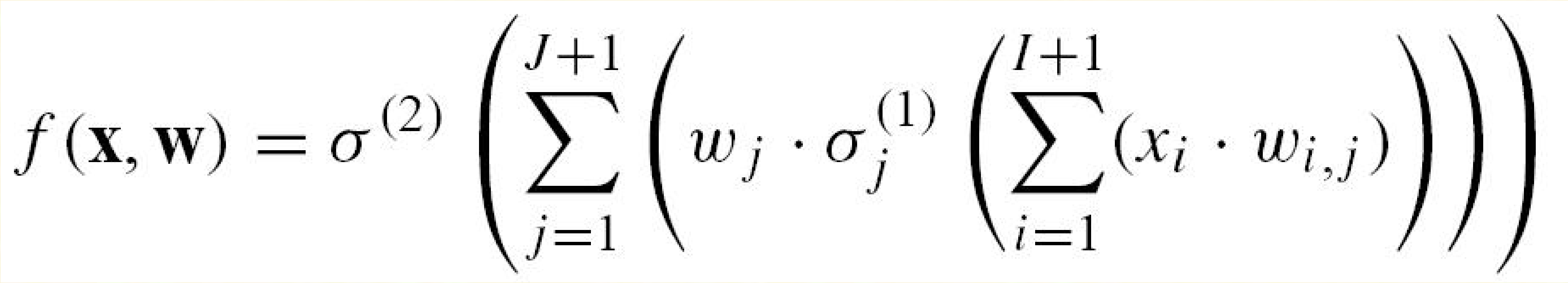
(Tarca *et al.*, 2005)

- Multi-layered feed-forward neural network (multi-layer perceptron, MLP)
- The neural network fitting function approximating the bias on *A, X* and *Y* would read:

$$f\left(\mathbf{x},\mathbf{w}\right)=\sigma^{(2)}\left(\sum_{j=1}^{J+1}\left(w_j\cdot\sigma_j^{(1)}\left(\sum_{i=1}^{I+1}(x_i\cdot w_{i,j})\right)\right)\right)$$

Neural network based spatial and intensity normalization

(Tarca *et al.*, 2005)



$$f(\mathbf{x}, \mathbf{w}) = \sigma^{(2)}\left(\sum_{j=1}^{J+1}\left(w_j \cdot \sigma_j^{(1)}\left(\sum_{i=1}^{I+1}(x_i \cdot w_{i,j})\right)\right)\right)$$

x is a vector having its components as *I*=3 features X, Y, A and constant value of 1 (accounting for the first layer bias)

w are the fitting parameters called weights

w_j and $\sigma_j^{(1)}$ represent the hidden neurons

References

- **Tarca, A. L., Cooke, J. E. & Mackay, J. (2005).** A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. *Bioinformatics* **21**, 2674-2683.
- **Wilson, D. L., Buckley, M. J., Helliwell, C. A. & Wilson, I. W. (2003).** New normalization methods for cDNA microarray data. *Bioinformatics* **19**, 1325-1332.