Estimating the Posterior Probability of Differential Gene Expression from Microarray Data

Marina Sapir and Gary A. Churchill
The Jackson Laboratory, Bar Harbor, ME

Abstract: We present a robust algorithm for estimating the posterior probability of differential expression of genes from microarray data. Our approach is based on an orthogonal linear regression of the signals obtained from the two color channels. Residuals from the regression are modeled as a mixture of a common component and a differentially expressed component and an EM-algorithm is used to deconvolve the mixture. The algorithm provides estimates of the measurement error variance, the proportion of differentially expressed genes, and the probability that each individual gene belongs to the differentially expressed class. We have applied this procedure to both real and simulated microarray data. Our simulation results demonstrate that the algorithm can estimate the key parameters with high precision over a wide range of models. Application of the method to replicated experiments demonstrates that the classification of differentially expressed genes is highly reproducible.

Part I: Modeling Fluorescent Intensities

We used data from self comparisons to identify a scale transformation of the raw fluorescent intensities for which the relationship between the color channels is linear with additive errors that are independent of the absolute signal intensity.

We found that a logarithm transform of intensities offset by a color dependent constant provided linearity and homogeneous variance for a wide range of data sets. We fit the relationship to fluorescent intensities y_1 and y_2 obtained from the "red" and "green" color channels:

$$\log(y_1 - \gamma_1) = \alpha + \beta \log(y_2 - \gamma_2)$$

where
g1 and g are the offsets and
a and b are the parameters of the regression line.

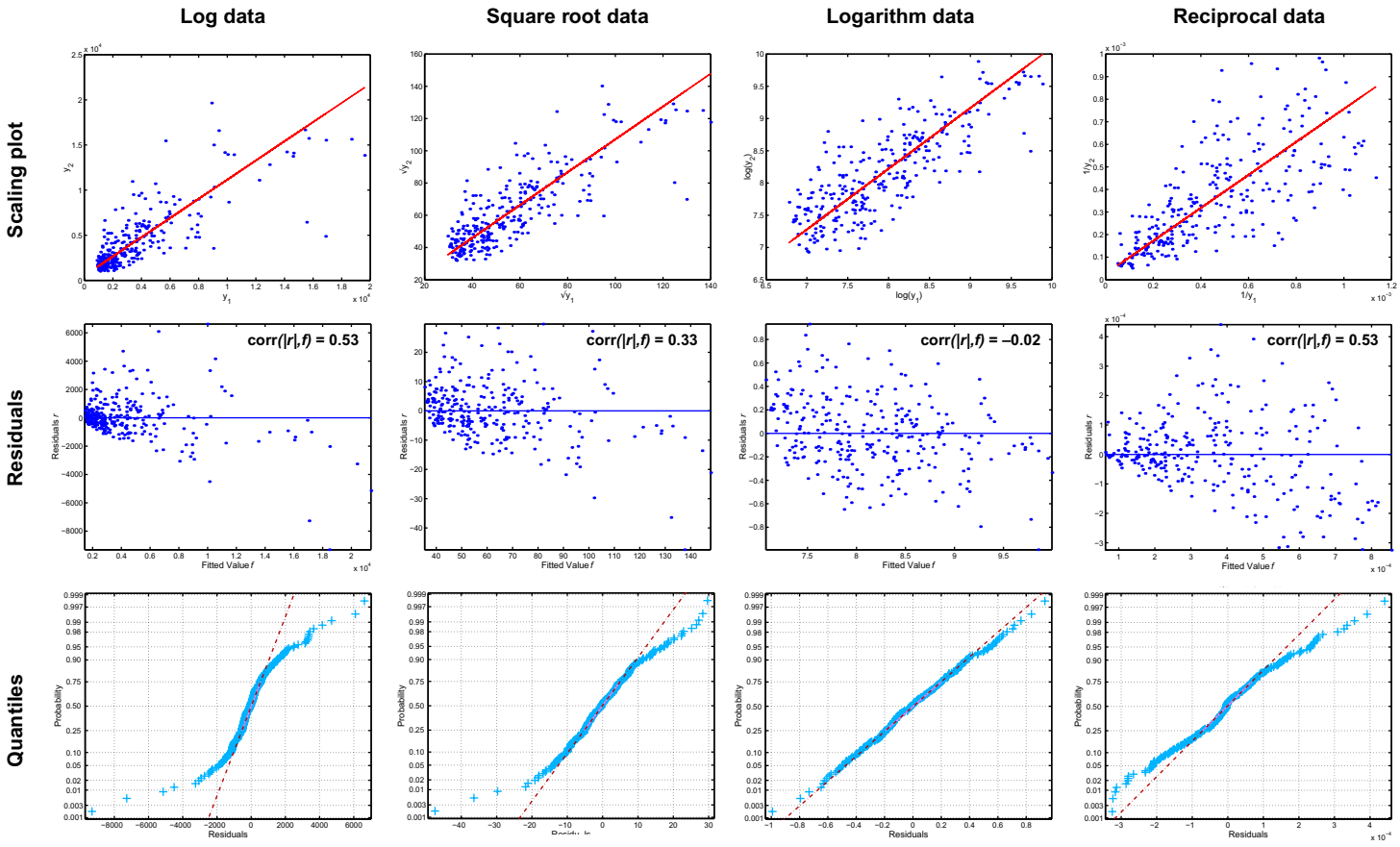
This empirically derived relationship between the fluorescent signals in self comparisons suggests a plausible model for the relationship between signal intensity y_{ig} and mRNA concentration x_{ig} for a gene g in sample i for non-self comparisons

$$y_{ig} = a_i x_{ig}^{bi} + c_i.$$

The follow panels show

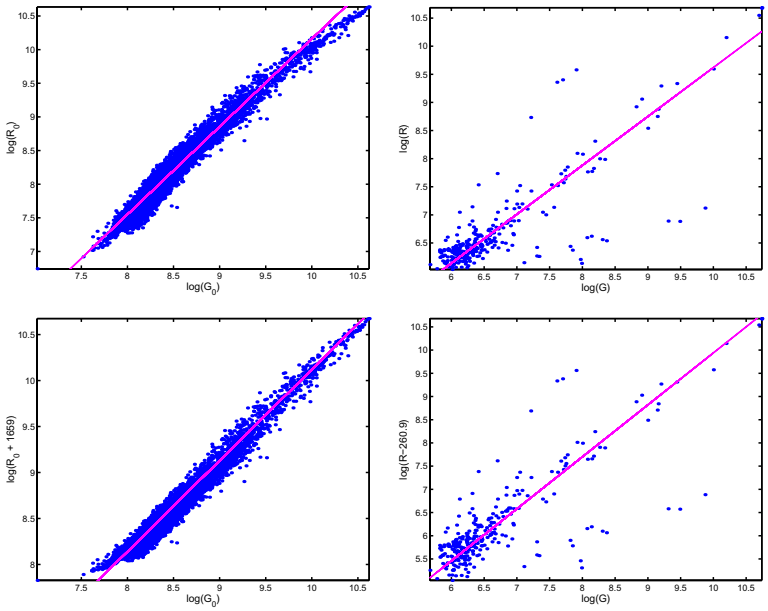
- Details of the method for fitting a linear function to a scatterplot of (transformed) intensities.
- The effects of different scaling functions (identity, square root, logarithm and reciprocal) on the distribution of residuals.
- The effect of the offset parameter on the linearity of the log-log scatter plot of signal intensities. We believe that the offset is directly related to the background signals in the two color channels.

Selection of the Scaling Function: Placenta-Placenta Self Comparison on GEM Array



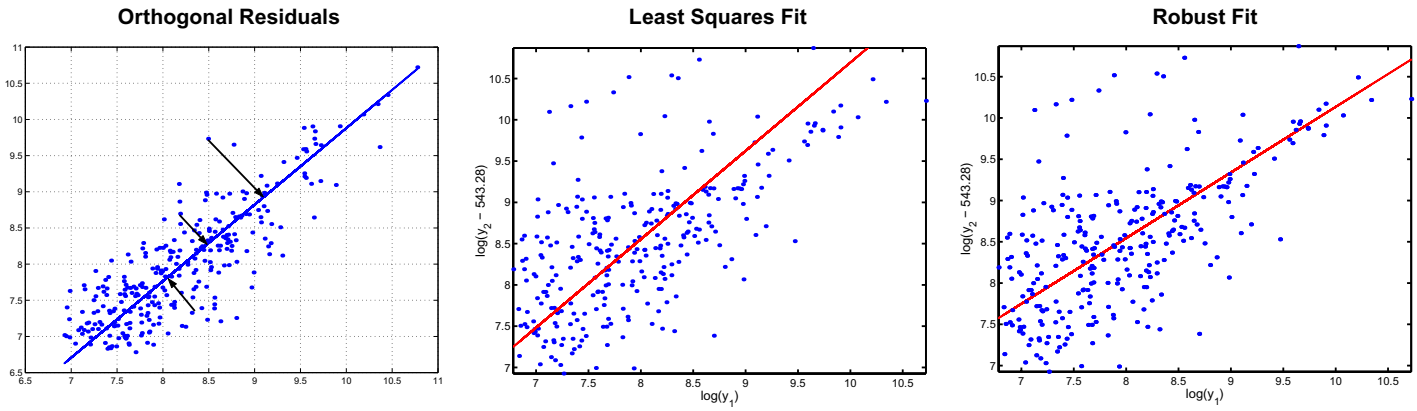
Conclusion: The residuals from a logarithm scaling transformation are approximately normal and have minimal correlation with fitted values.

The Shifted Logarithm Scaling Function



Conclusion: Shifting the raw measurement improves linearity on the logarithm scale.

Fitting a Regression Line to the Data



Orthogonal residuals provide a symmetric treatment of the two fluorescent intensities y1 and y2.
Robust regression reduces the influence of differentially expressed genes of the Fitted line.

Part II: Modeling the Residuals

In a non-self comparison, the orthogonal residuals, r are modeled as a two component mixture. The first component, C, represents the class of common genes that expressed equally in the two mRNA populations. The second component, D, represents differentially expressed genes.

$$\Pr(r) = (1 - \pi)\Pr(r|C) + \pi \Pr(r|D)$$

where
$1 - \pi$ = proportion of common genes
π = proportion of differentially expressed genes

Residuals from common genes are modeled as a Normal distribution with mean 0 and variance s2. Residuals from differentially expressed genes are modeled as a Uniform distribution.

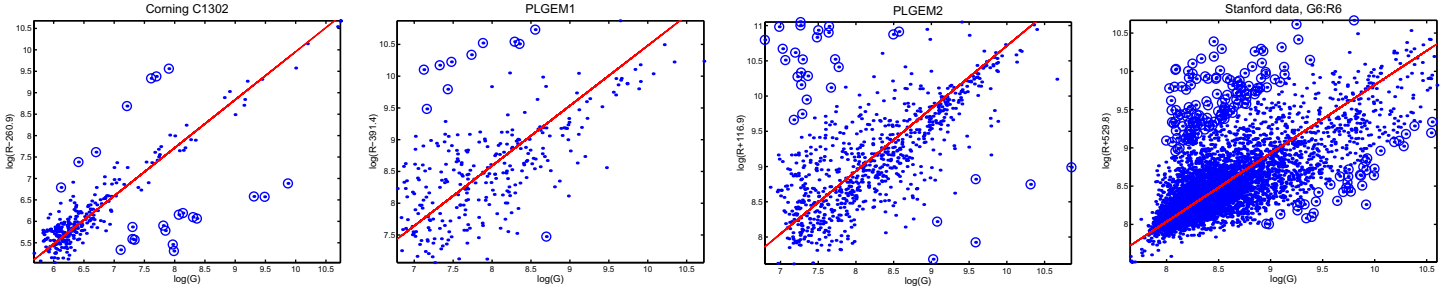
Deconvolution of the Residuals

An EM algorithm is used to obtain estimates of the proportion of differentially expressed genes, π, and the residual error variance, s2. The E-step of the EM algorithm applies Bayes' rule to obtain the conditional probabilities

$$\Pr(D|r) = \frac{\pi \Pr(r|D)}{\pi \Pr(r|D) + (1 - \pi)\Pr(r|C)}$$

Thus we can obtain estimates of the posterior probability of differential expression based on information in the regression residuals.

Examples of Differentially Expressed Genes



Consistency of Results

Replication across arrays:
Placenta vs Liver From GEM2

Array 2 \ Array 1	P_1<-0.5	-0.5<P_1<0.5	P_1>0.5
P_2 < -0.5	3	8	0
-0.5 < P_2 < 0.5	6	720	5
P_2 > 0.5	0	17	20

Array 2 \ Array 1	P_1<-0.95	-0.95<P_1<0.95	P_1>0.95
P_2 < -0.95	2	3	0
-0.95 < P_2 < 0.95	0	748	2
P_2 > 0.95	0	11	13

For spots replicated with an array: Corning data

	Posterior probability	
	Chip 1300	Chip 1302
1 / lgfbp1	1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00
2 / Gh	0.22 1.00 0.44 0.09	0.29 1.00 1.00 1.00
46 / Mt2	1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00
58 / Ax3	-0.02 0.01 -0.01 0.01	-0.98 -0.03 -0.02 -0.02
94 / igf2r	-1.00 -0.99 -1.00 -1.00	-0.35 -0.92 -1.00 -1.00
97 / ppara	-1.00 -1.00 -1.00 -1.00	-1.00 -1.00 -1.00 -1.00
98 / Igf2	-1.00 -1.00 -1.00 -1.00	1.00 1.00 1.00 1.00

Conclusions

- Raw ratios are not reliable measures of differential expression
- The two color channels in self comparisons are related by a linear model on the logarithm scale with an additive offset
- Robust orthogonal regression can be used to fit a linear model to data from non-self comparisons
- Residuals from the orthogonal regression can be separated into common and differentially expressed components using an EM algorithm
- Classification of differentially expressed genes is consistent on both within array and across array replications