

Supplementary materials: Molecular analysis of positional identity in the developing mouse retina

Elva Diaz, Yee Hwa Yang, Todd Ferreira, Kenneth C. Loh, Yasushi Okazaki,
Yoshihide Hayashizaki, Marc Tessier-Lavigne, Terence P. Speed and John Ngai.

September 15, 2002

A1: Preprocessing - Normalization

The raw data from a cDNA microarray experiment consist of pairs of image files. Pre-processing steps such as image analysis and normalization are required to extract reliable measures of the fluorescence intensities and allowed comparisons within and between slides. All images were processed by `Spot` with foreground seeds set to 5 pixels square. In addition, within and between slide normalization was performed on all data. For within slide normalization, we used the "print-tip group scale normalization" method followed by multiple-slide scale normalization allowing comparison of different experiments. Further details of our normalization methods can be found in Yang et al. (2002).

A2: Analysis: Identification of differentially expressed genes

After careful pre-processing, the multiple-slide normalized gene expression data considered here can be summarized in a gene-by-slide matrix X of log intensity ratios $\log_2 R/G$, with k rows corresponding to the genes being studied and n columns corresponding to the n different hybridizations. However, this does not take into consideration the fact that the experimental design imposes a very specific structure on the columns of this matrix. An important statistical question is to determine how to combine data from different slides while taking into account the design of the experiment.

Description of the motivation and development of our statistical analysis can be found in the supplement of the accompanying paper. The details specific to our experiments are presented below.

The following tables summarises our experiments:

	Cy5	Cy3		Cy5	Cy3
y1	N	W	y11	V	W
y2	T	W	y12	D	N
y3	V	W	y13	N	T
y4	D	W	y14	T	V
y5	N	W	y15	V	D
y6	T	W	y16	D	N
y7	V	W	y17	N	T
y8	D	W	y18	T	V
y9	N	W	y19	V	D
y10	T	W			

For a typical gene i , let us denote the gene's mean intensity value corresponding to the four different regions (nasal, temporal, dorsal and ventral) of the retina and the whole retina by N_i , T_i , D_i , V_i and W_i respectively. Furthermore, define the log transformation of these values to be $n_i = \log_2 N_i$, $t_i = \log_2 T_i$, $d_i = \log_2 D_i$, $v_i = \log_2 V_i$ and $w_i = \log_2 W_i$. To estimate the spatial gene expression for gene i , we fit the following linear model:

$$y_i = X\beta_i + \epsilon_i,$$

where y_i is a vector of log-ratios from the different slides; X is the design matrix and β_i is a vector of parameters. The four estimable parameters we choose are given by $n_i^{(w)} = n_i - w_i$, $t_i^{(w)} = t_i - w_i$, $d_i^{(w)} = d_i - w_i$, and $v_i^{(w)} = v_i - w_i$.

The values associated with the experiments presented in this paper are given below.

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \\ y_{9i} \\ y_{10i} \\ y_{11i} \\ y_{12i} \\ y_{13i} \\ y_{14i} \\ y_{15i} \\ y_{16i} \\ y_{17i} \\ y_{18i} \\ y_{19i} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \cdot \begin{pmatrix} n_i^{(w)} \\ t_i^{(w)} \\ d_i^{(w)} \\ v_i^{(w)} \end{pmatrix} + \epsilon_i \quad (1)$$

The parameter β_i can be estimated by $\hat{\beta}_i = (X'X)^{-1}X'y_i$. In practice, we used a robust linear model, so that our estimates are less affected by outliers. The data are fitted to the linear model described above by iteratively reweighted least squares (IWLS) procedure using the function `rlm` provided in the library `MASS` in the statistical software package `R`. Details of the theory and implementation of the “robust linear model” can be found in Venables & Ripley (1999).

After obtaining

$$\hat{\beta}_i = \begin{pmatrix} \hat{n}_i^{(w)} \\ \hat{t}_i^{(w)} \\ \hat{d}_i^{(w)} \\ \hat{v}_i^{(w)} \end{pmatrix},$$

the six contrasts can then be calculated by taking appropriate linear combination of the vector of parameter estimates. That is, for a typical gene, the six pairwise contrasts can be calculated by:

$$\begin{pmatrix} (\hat{NT})_i \\ (\hat{DT})_i \\ (\hat{VT})_i \\ (\hat{DN})_i \\ (\hat{VN})_i \\ (\hat{DV})_i \end{pmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{pmatrix} \hat{n}_i^{(w)} \\ \hat{t}_i^{(w)} \\ \hat{d}_i^{(w)} \\ \hat{v}_i^{(w)} \end{pmatrix} \quad (2)$$

Four-way estimate of expression.

For easier visualization, we would like to estimate the value of a single effect against an average of all four. The parameterization here is thus $n_i^* = n_i - \frac{1}{4}(n_i + t_i + d_i + v_i)$, whose unbiased estimate is $\hat{n}_i^* = \tilde{n}_i = \hat{n}_i^{(w)} - \frac{1}{4}(\hat{n}_i^{(w)} + \hat{t}_i^{(w)} + \hat{d}_i^{(w)} + \hat{v}_i^{(w)})$. In effect, such estimation recreates the hypothetical pooled retina reference ($n_i + t_i + d_i + v_i$) in silico. Notice that this does not represent the absolute expression profile, but rather the relative expression between a portion of the retina and the pooled reference. The following equation provides the matrix multiplication that leads to this four-way representation:

$$\begin{pmatrix} \tilde{n}_i \\ \tilde{t}_i \\ \tilde{d}_i \\ \tilde{v}_i \end{pmatrix} = \begin{bmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{bmatrix} \cdot \begin{pmatrix} \hat{n}_i^{(w)} \\ \hat{t}_i^{(w)} \\ \hat{d}_i^{(w)} \\ \hat{v}_i^{(w)} \end{pmatrix} \quad (3)$$

The above procedures is applied to all 19,000 genes on the array and produces different sets of estimated contrasts. By applying equation 2 to all 19,000 genes, we can create a gene by 6-pair matrix (*C6*) of retina profiles. In addition, a gene by 4-way matrix (*C4*) of retina profiles can be obtained from equation 3. These two matrices contain the same fundamental information, and are specific linear transformations of one other.

A3: Cluster analysis.

We selected for clustering the genes having the largest and smallest 50 values for each of the 6 pairwise contrasts, this being 362 unique genes in all. The names of these genes are given in supplement B. An hierarchical clustering of these genes was performed using a modified Mahalanobis distance and Ward agglomeration. The metric used to measure the similarity between the retina expression profiles for gene i and gene j is defined as:

$$\hat{\delta}_{ij} = (\hat{\beta}_i - \hat{\beta}_j)'(X'X)(\hat{\beta}_i - \hat{\beta}_j),$$

where X is the design matrix of our experiment defined in equation 1. This metric was chosen instead of Pearson correlation or Euclidean distance because our input here consists of correlated parameter estimates with different variances. It can be thought of as a weighted Euclidean distance which, when measuring the similarity between two retina expression profiles, places more emphasis on components which are more precise, and which also adjusts for correlation.

Rather than identifying interesting profiles by visual inspection of clusters arising by cutting the resulting dendrogram at one place, we now consider all 362 clusters that consist of two or more genes which arise when we cut the dendrogram at all possible places. We measure the heterogeneity h of a cluster by taking the largest of the 6 standard deviations of the pairwise contrasts. For any score s , we can find the maximal set of disjoint clusters having $h < s$. This is equivalent to cutting the dendrogram at different places down different branches, with the value s determining placement of the cut. Using $s = 0.3$ we obtained 9 clusters, which we regard as a reasonable balance of heterogeneity versus number of clusters.

In each of the 9 clusters, we ranked the genes based on the Mahalanobis distance between the retina expression profile for that gene and the origin, written as $\hat{\delta}_i = \hat{\beta}_i'(X'X)\hat{\beta}_i$. Genes from each cluster with a high $\hat{\delta}_i$ score were selected for verification. As another way of depicting the relationship between these clusters, we carried out a second hierarchical clustering using the average profiles of the 9 disjoint clusters obtained above. A dendrogram based on correlation metric and Ward agglomeration was then built.

We also define a coefficient of variation (CV) for each cluster as the minimum across contrasts of the CV across genes within a cluster. That is, we calculate the standard deviation divided by the absolute value of mean across genes of each contrast, and then taking the minimum value across contrasts. A small CV implies that for genes within this cluster, the contrast values are likely to be non-zero, i.e. unlikely to be random variation.

All of the analyses just described were performed in the statistical software **R** based mainly on the “*Statistics for Microarray Analysis*” (SMA) package. **R** scripts will be provided upon request.

References

- [1] Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistical Sinica.
- [2] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. 95: 14863-14868.

- [3] Ihaka, R., Gentleman, R. (1996) A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5, 299-314.
- [4] Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* 29:389-95.
- [5] Kerr, M. K., Martin, M., Churchill, G. A. (2000). Analysis of variance for gene expression microarrays. *Journal of Computational Biology* 7:819-837.
- [6] Venables, W. N., Ripley, B. D. (1999) *Modern applied statistics with S-PLUS* (ed. Springer).
- [7] Wolfinger, R. D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 8:625-637
- [8] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30, e15.