Reproducibility, Variance Stabilization, and Normalization in CodeLink^{TM} Data with Application to Cancer in Rats

Sue Geller[1], David M. Rocke[2], Danh Nguyen[3], Raymond Carroll[4].

Keywords: microarray, variance, transformation, CodeLink^{TM}

Although the experimental aspects of microarray methods are maturing rapidly, the analysis of the array data is still a difficult exercise with many issues that have not been fully resolved in the literature. One of the issues is that, in order to use parametric model-based analysis and even most non-parametric methods, the variance of the replicates of the data needs to be constant across expressions. It was determined by Rocke and Durbin ([4]), that the variance was not constant for spotted microarray data and instead conformed to a model, which had been used in the analytical chemistry and the environmental science literature. This is a two error component model, $y = \alpha + \mu e^{\eta} + \epsilon$, in which there is an additive error that dominates when μ is small and the proportional error that dominates when μ is large. In a subsequent manuscript ([1]), they developed a transformation to stabilize the variance. The applicability of the two component model to Affymetrix data as well as an algorithm for simultaneously finding the constant of the transformation and normalizing the data was given in Geller, Gregg, Hagerman, and Rocke ([2]). In all these (and other) cases, a data set was used that consisted of technical replicates, so the effect of the transformation on variance stabilization is unstudied in sets with multiple technical replicates.

The effect, of lack thereof, of diet on cancer has been a long-standing question. In addition to the usual usual observational trials, the effect of diet in the presence of carcinogens was studied on the genetic level in rats using the CodeLink^{TM} microarray platform ([3]). Three diets (corn oil, fish oil, olive oil) were given with either saline or the carcinogen AOM, and the rats killed at 12 hours or 10 weeks. Twenty two of the 59 rats had two technical replicates, one three technical replicates, and three had four technical replicates, making this data set a useful one for studies of reproducibility and variance stabilization. While many questions remain unanswered, the following are some of the conclusions of our analysis to date.

- The CodeLink^{TM} data from these studies appear to conform in broad terms with the two error component model, $y = \alpha + \mu e^{\eta} + \epsilon$.
- CodeLink^{TM} data has a great deal of variability among technical replicates.

[1]correspondence to Sue Geller, Department of Mathematics – MS 3368, Texas A&M University, College Station, TX 77843-3368 or email to `geller@math.tamu.edu`.

[2]CIPIC, 2343 Academic Surge, University of California, One Shields Ave, Davis, CA 95616

[3]Department of Epidemiology and Preventive Medicine, University of California, Davis, CA 95616

[4]Department of Statistics – MS 3143, Texas A&M University, College Station, TX 77843-3143

- It is possible to transform the gene expressions so that the variance within genes across arrays is approximately the same regardless of the level of expression of the genes. This transformation resembles the logarithm at high expression levels and resembles a linear transformation at low expression levels.
- After transformation, the data for highly expressed genes appear approximately normally distributed, but for low expression levels there are frequent outliers. These may be caused by dust or scratches that have only a small effect on highly expressed genes, but larger proportional effects on genes with low expression.
- The choice of constant for the transformation can be in a broad range and still produce transformed technical replicates with constant variance and symmetric errors. Thus, the same constant can be used for all the data to produce a data set with approximately constant variance regardless of the level of expression of the genes for each set of technical replicates.
- Normalization in conjunction with stabilization of variance is more effective than normalization and then stabilization of variance, i.e., normalization may be incomplete or inaccurate if it is done before transformation.

References

[1] Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. 2002. A Variance Stabilizing Transformation for Gene Expression Microarray Data, *Bioinformatics* 18:S105-S110.

[2] Geller, S. C., Gregg, J. P, Hagerman, P., and Rocke, D. M. 2003. Transformation and Normalization of Oligonucleotide Microarray Data, *Bioinformatics* 19:1817-1823.

[3] Ramakrishnan, R., Dorris, D., Lublinsky, A., Nguyen, A., Domanus, M., Prokhorova, A., Gieser, L., Touma, E., Lockner, R., Tata, M., Zhu, X., Patterson, M., Shippy, R., Sendera, R. J., and Mazumder, A. 2002. An Assessment of Motorola CodeLink Microarray Performance for Gene Expreseesion Profiling Applications. *Nucleic Acids Res.* 30 (7):e30.

[4] Rocke, D. M. and Durbin, B. P. 2001. A Model for Measurement Error for Gene Expression Arrays. *J. Comp. Bio* 8:557-569.