Varistran: Anscombe's variance stabilizing transformation for RNA-seq gene expression data

Paul Francis Harrison[1]

[1]Monash Bioinformatics Platform, Monash University

8 March 2017

Paper DOI: http://dx.doi.org/10.21105/joss.00257
Software Repository: https://github.com/MonashBioinformaticsPlatform/varistran
Software Archive: http://dx.doi.org/10.5281/zenodo.846925

Summary

RNA-seq measures RNA expression levels in a biological sample using high-throughput cDNA sequencing, producing counts of the number of reads aligning to each gene. Noise in RNA-seq read count data is commonly modelled as following a negative binomial distribution, where the variance is a quadratic function of the expression level. However many statistical, machine learning, and visualization methods work best when the noise in a data set has equal variance. Varistran is an R package that uses Anscombe's (1948) variance stabilizing transformation for the negative binomial distribution to transform RNA-seq count data, so that the noise has equal variance across all measured gene expression levels. The transformed data may be treated as $\log_2$ transformed gene expression levels, but with variability reduced at low read counts. Varistran also includes a function to open a Shiny report with simple diagnostic visualizations, including a plot to assess how effective the variance stabilization was, a biplot of samples and genes, and a heatmap. This allows defective samples, sample mislabling, and batch effects to be easily identified.

References

Anscombe, Francis J. 1948. "The Transformation of Poisson, Binomial and Negative-Binomial Data." *Biometrika* 35 (3/4): 246–54.