Anscombe's 1948 variance stabilizing transformation for the negative binomial distribution is well suited to RNA-Seq expression data

Paul Harrison paul.harrison@monash.edu Monash Bioinformatics Platform, Monash University, Australia

Objective

To apply the wide range of statistical and machine learning methods and data visualization techniques to RNA-Seq expression data which have not been designed specifically with this data in mind, we would like a transformation of RNA-Seq count data that:

- Is as close as possible to a logarithmic transformation, allowing results to be interpreted in terms of fold changes.
- Has noise in each gene of roughly equal variance.

Logarithmic transformation is not finite for zero count, and tends to inflate the variance for genes with low expression levels, so we would like some manner of **moderated log transformation**.

Conclusion

edgeR provides a function `cpm` that can compute moderated $\log_2$ read counts. Using edgeR's estimate of the common dispersion ϕ, a reasonable choice for the `prior.count` parameter in the `cpm` function is $0.5/\phi$. The more dispersion (biological noise) that is present, the less moderation that is needed.

An improvement on this is possible by using Anscombe's [3] variance stabilizing transformation for the negative binomial distribution, and by choosing the dispersion parameter with the explicit aim of stabilizing variance. We have implemented this in an R package called Varistran.

`https://github.com/MonashBioinformaticsPlatform/varistran`

The negative binomial distribution

The negative binomial distribution has been proposed as a model of variation in RNA-Seq expression data [1]. The negative binomial distribution has two parameters, a mean μ and a "dispersion" parameter ϕ. The dispersion is sometimes alternately given as $k=1/\phi$, for example as the `size` parameter in the `d/p/q/rnbinom` family of functions in R.
The variance of the negative binomial distribution captures the essence of why RNA-Seq expression data is difficult to work with:

$$\sigma^2=\mu+\phi\mu^2$$

The first term represents Poisson noise, which is amplified by log transformation. The second term represents biological variation; if we had this term alone, log transformation would precisely stabilize the variance.

Variance stabilizing transformation

Naive transformation: If the variance does not change too rapidly, a reasonable variance stabilizing transformation can be obtained by integrating $\int \frac{1}{\sigma} d\mu$ to obtain, omitting a scaling factor:

$$y=\sinh^{-1}\sqrt{\phi x} \qquad (\sinh^{-1}x=\ln\left(x+\sqrt{1+x^2}\right))$$

R package DESeq2 [2] offers this transformation.

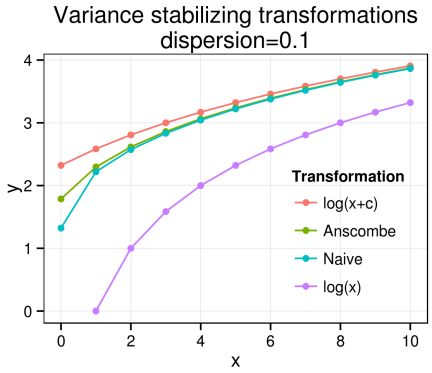
Anscombe's transformation: Anscombe [3] suggests a small correction to our naive transformation:

$$y=\sinh^{-1}\sqrt{\frac{x+3/8}{1/\phi-3/4}}$$

log(x+c) transformation: Anscombe further suggests a simpler transformation suitable where the mean is large:

$$y=\ln\left(x+\frac{1}{2\phi}\right)$$

This resembles the $\log_2(x+c)$ transformation with "prior count" c offered by the `cpm` function in R package edgeR, suggesting a reasonable value for this prior count parameter.



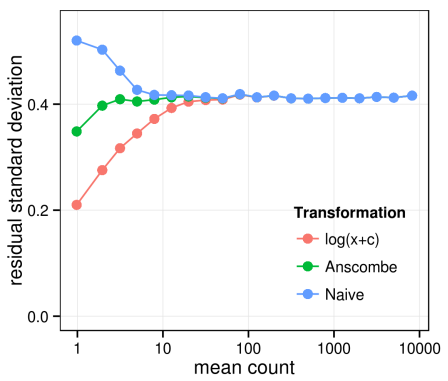
Implementation

We have implemented these transformations in an R package Varistran.

Counts are scaled before transformation so all samples have the mean library size, library sizes adjusted using the TMM method. Results are scaled and offset so as to behave like $\log_2$ Reads Per Million for large x.

Results

Simulation

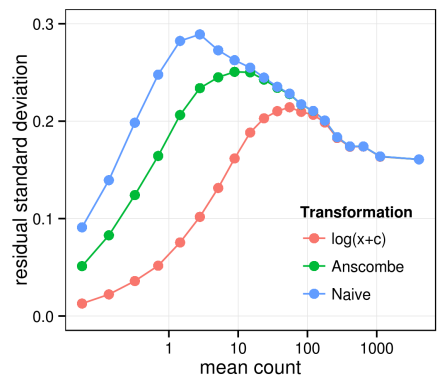


We simulate 100,000 "genes", 4 samples, with a dispersion of 0.1 and a range of mean expression levels. Here and below, we sort genes by average count then partition the genes into 20 groups, and plot the mean residual standard deviation of each group.

Anscombe's transformation performs excellently with simulated data, the other two transformations performed less well.

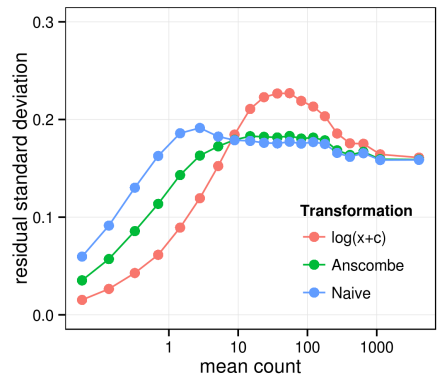
Real RNA-Seq data

Using edgeR's dispersion estimate:



We use Bottomly's [4] RNA-Seq comparison of the brains of two inbred strains of mice. There are 21 striatum samples, sequenced using three flowcells. A batch effect is evident from these flowcells, this is included in the linear model when estimating the dispersion and calculating residuals. Using edgeR [5], we estimate a common dispersion of 0.02.

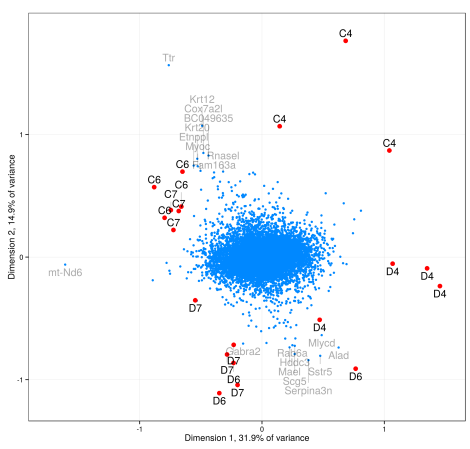
Using a dispersion estimate optimized to equalize noise:



The authors of DESeq2 [2] suggest that the Poisson component of noise in RNA-Seq data may be over-dispersed (a different kind of dispersion!)*. We therefore tried choosing the dispersion parameter to minimize the Coefficient of Variation (standard deviation divided by mean) of the residual standard deviation after fitting a linear model. Only genes with a mean count of at least 5 were used.

The dispersion estimate for the log(x+c) method was largely unchanged, but the dispersion estimate for Naive and Anscombe methods was reduced to around 0.01. The dispersion parameter here is really acting as the balance between over-dispersed Poisson noise and biological noise.

The transformed counts are now ready for further analysis, for example the biplot shown to the right. This figure was produced using the `plot_biplot` function in Varistran. C/D indicates strain and 4/6/7 indicates flowcell number. Blue points are genes and red points are mice. We can see a subset of genes differentially expressed between the strains, and a broad spread of expression perhaps due to the batch effect.



* DESeq2 finds 2.5-fold over-dispersion of the variance of the Poisson component even in our simulated negative binomial data, so we have chosen not to trust DESeq2's estimate of the over-dispersion.

References

[1] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, vol. 23, no. 21, pp. 2881–2887, Nov. 2007.
[2] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014.
[3] F. J. Anscombe, "The Transformation of Poisson, Binomial and Negative-Binomial Data," *Biometrika*, vol. 35, no. 3–4, pp. 246–254, Dec. 1948.
[4] D. Bottomly, N. A. R. Walter, J. E. Hunter, P. Darakjian, S. Kawane, K. J. Buck, R. P. Searles, M. Mooney, S. K. McWeeney, and R. Hitzemann, "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays," *PLoS ONE*, vol. 6, no. 3, p. e17820, Mar. 2011.
[5] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Res.*, vol. 40, no. 10, pp. 4288–4297, May 2012.