Differential proportionality – an alternative to differential gene expression not requiring sample normalization

I. Erb[1], T. Quinn[2], D. Lovell[3], and C. Notredame[1]
[1]Centre for Genomic Regulation (CRG), C\Dr Aiguader 88, 08003 Barcelona, Spain; *ionas.erb@crg.eu*
[2]Deakin University, Geelong, Victoria, Australia
[3]Queensland University of Technology, Brisbane, Queensland, Australia

Abstract

In gene expression, the emergence of large aggregated data sets along with new single-cell technologies have led to a heterogeneity of samples that makes normalization extremely difficult. The few existing log-ratio applications to gene expression analysis (Fernandes and others, 2013; Lovell and others, 2015) do not fully overcome the problem of sample heterogeneity as their results depend crucially on the choice of a reference in the form of a gene or gene set (Erb and Notredame, 2016).

Here we propose a differential analysis of all possible gene ratios. More precisely, considering n samples coming from two different conditions, we propose a statistic to detect proportionality (i.e. log-ratio variance close to zero) between genes $\vec{x}$ and $\vec{y}$ in one condition that differs in the proportionality factor in the other condition:

$$\vartheta(\vec{x},\vec{y}) = \frac{k \cdot \mathrm{var}\, \mathrm{L}^{\vec{x},\vec{y}}_{\{1,\ldots,k\}} + (n-k) \cdot \mathrm{var}\, \mathrm{L}^{\vec{x},\vec{y}}_{\{k+1,\ldots,n\}}}{n \cdot \mathrm{var}\, \mathrm{L}^{\vec{x},\vec{y}}_{\{1,\ldots,n\}}}, \tag{1}$$

where by $\mathrm{L}^{\vec{x},\vec{y}}_{\{1,\ldots,k\}}$ we denote the log ratio of $\vec{x}$ and $\vec{y}$ over the indices $\{1,\ldots,k\}$. ϑ can be obtained from a decomposition of log-ratio variance into between and within group variance. (The denominator corresponds to the latter, and ϑ values fall between 0 and 1, with smaller values indicating better separation.) Note that ϑ is related to the statistic F underlying one-way ANOVA by $F = (1-\vartheta)/\vartheta$. In fact, a standard differential expression framework can now be applied (applied, however, on *ratios*) using false discovery rates from permutation tests to detect significant values of ϑ.

As an example, we apply this framework to a data set of 98 post-mortem brain samples (Lonsdale, J. and others, 2013) from cortex and cerebellum. Unlike in classical differential expression studies, where the main result is a list of genes whose read counts differ between conditions, here we obtain a list of gene pairs whose ratio of co-expression differs between conditions. This allows for a subsequent network analysis, cf. (Tesson and others, 2010) for the classical equivalent called *differential correlation.*

We also derive an alternative to ϑ that can handle zeroes and compares with the use of pseudo counts. For this statistic, the three terms of the form $k \cdot \mathrm{var}\, \mathrm{L}^{\vec{x},\vec{y}}_{\{1,\ldots,k\}}$ in (1) are replaced respectively by

$$\sum_{i=1}^{k} \left(\frac{x_i^\alpha}{\frac{1}{k}\sum_{j=1}^{k} x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{k}\sum_{j=1}^{k} y_j^\alpha} \right)^2 . \tag{2}$$

This is inspired by the observation that chi-square distances converge to log-ratio variances when applying a Box-Cox transformation with the parameter $\alpha \to 0$ (Greenacre, 2009). We supplement this work with an R package that provides a fast and efficient implementation of these analyses.

References

Fernandes, A. et al. (2013). ANOVA-Like Differential Gene Expression Analysis of Single-Organism and Meta-RNA-Seq. *PLoS one 8*(7), e67019.

Lovell, D. et al. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comp Biol 11*, e1004075.

Erb, I. and Notredame, C (2016). How should we measure proportionality on relative gene expression data? *Theory Biosci 135*(1-2), pp. 21–36.

Lonsdale, J. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat Genet 45*, 580585.

Tesson, B.M. et al. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics 11*, pp. 497.

Greenacre, M. (2009). Power transformations in correspondence analysis. *Comput Statist Data Anal 53*, pp. 3107–3116.