Differential proportionality – a normalization-free approach to differential gene expression

Ionas Erb[1,*], Thomas Quinn[2], David Lovell[3], Cedric Notredame[1]

[1]Centre for Genomic Regulation (CRG), Barcelona, Spain; [*]ionas.erb@crg.eu
[2]Deakin University, Geelong, Victoria, Australia
[3]Queensland University of Technology, Brisbane, Queensland, Australia

Abstract. Gene expression data generated by next generation sequencing technologies (RNA-seq) are relative: the total number of sequenced reads has no biological meaning. Additional knowledge, in the form of an unchanged reference, is necessary to “normalize” the data; however, this reference can usually only be estimated. Here we propose to base gene expression analysis entirely on ratios, so normalization factors cancel by default. Although the differential expression of individual genes cannot be recovered this way, the ratios themselves can be differentially expressed (even when their constituents are not). Specifically, we show how the differential expression of gene ratios can be formalized by decomposing log-ratio variance (LRV) and deriving intuitive statistics from it. Here we focus on the change in proportionality factors between two groups of samples. For this, we propose a statistic that is equivalent to the squared t-statistic of one-way ANOVA, but for gene ratios. In doing so, we show how precision weights can be incorporated to account for the peculiarities of count data, and, moreover, how a moderated statistic can be derived in the same way as the one following from a hierarchical model for individual genes. We also show how to deal with zero counts, deriving expressions of our statistics that are able to incorporate them. The proposed framework is applied to a data set from the GTEx consortium [1] consisting of 98 samples from the cerebellum and cortex, with selected examples shown. An R package containing a computationally efficient implementation of the approach was released as an addendum to the propr package.

Statistics for differential ratio expression

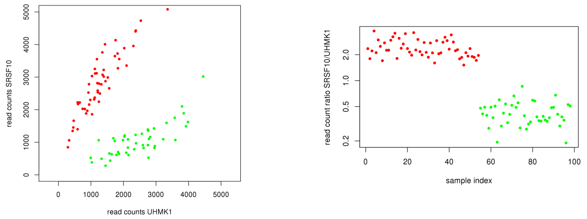
Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be vectors representing genes whose components are read counts in different samples. Using the notation

$$\mathrm{L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k} := \left(\log\frac{x_1}{y_1},\ldots,\log\frac{x_k}{y_k}\right),$$

for k components, we now consider n samples falling into two groups (conditions) comprising k and n-k samples. A decomposition of log-ratio variance (LRV) into within-group and between-group variance gives

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

This is within-group over total LRV, a number between 0 and 1. If ϑ is small, $\boldsymbol{x}$ and $\boldsymbol{y}$ are *differentially proportional.* They are proportional within each group but differ in their proportionality factors:



The well-known squared t-statistic from one-way ANOVA is related to ϑ via

$$F = (n-2)\frac{(1-\vartheta)}{\vartheta}.$$

For both ϑ and F we can determine a false discovery rate (e.g. by permutation tests).

Refinements: We can now also weight these statistics using per-observation precision weights for the read counts coming from the genes (“voom” [2]). A moderated statistic can be derived for ratios in equivalence to the one following from a hierarchical Bayesian model for the genes (“limma” [3]).

Handling zeros

The Box-Cox family of data transformations with parameters α returns the logarithm for α tending to zero:

$$\log(x) = \lim_{\alpha\to 0}\frac{x^\alpha - 1}{\alpha}.$$

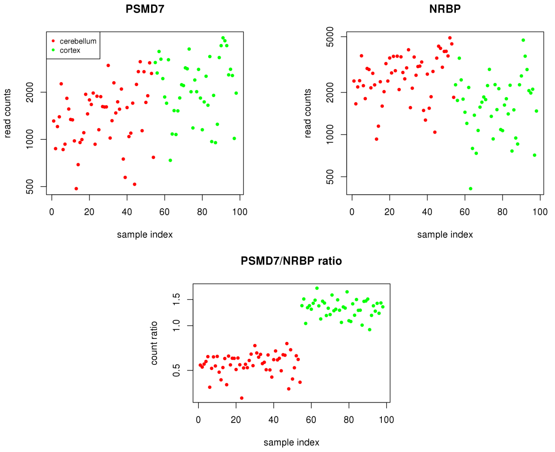
It has been shown [4] that this establishes a connection between log-ratio analysis and Correspondence Analysis (which can handle zeros naturally). We use this connection to approximate LRV by the (squared) χ-square distance between vectors taken to the power of α

$$d_\alpha(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^\alpha}{\frac{1}{n}\sum_{j=1}^n x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{n}\sum_{j=1}^n y_j^\alpha}\right)^2.$$

Replacing LRVs by the corresponding expressions involving α makes all our statistics compatible with zero counts. The smaller α, the better the approximation, but larger α will allow for higher significance of pairs involving zero counts.

Relation to differential gene expression

If we have a reference $\boldsymbol{z}$ (e.g. a gene) that is known to be unchanged across all samples, then $\vartheta(\boldsymbol{x},\boldsymbol{z})$ can be used as a measure for the differential expression of $\boldsymbol{x}$. Thus differential proportionality of gene pairs involving $\boldsymbol{z}$ is just differential expression of the partner.
Additionally, it can be shown formally that differentially expressed genes with opposite direction of change form differentially proportional pairs.
On the other hand, differentially proportional pairs need not contain differentially expressed genes:



More formally, one can show that the within-group LRV of the pair has to be sufficiently large for the pair to contain a differentially expressed gene.

Bibliography

[1] Lonsdale, J et al. (2013) The genotype-tissue expression (GTEx) project. *Nat Genet 45,* 580585.
[2] Law, CW et al. (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology 15,* R29.
[3] Smyth, GK (2005) Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 397-420.
[4] Greenacre, M (2009) Power transformations in correspondence analysis. *Comput Statist Data Anal 53,* 3107-3116.