The freely available R (R Development Core Team, 2009) package R/EBcoexpress implements the algorithm of Dawson and Kendziorski (2011), an empirical Bayesian approach for identifying differentially co-expressed (DC) gene pairs. Microarray and related high-throughput genomic experiments seek to identify genes that vary across biological conditions. This is often accomplished by identifying genes with changes in mean expression level, so-called differentially expressed (DE) genes. Although useful, major biological insights have resulted far less frequently than originally expected (Pollack, 2007; Zilliox et al., 2008). This is in part because diseases can manifest due to a de- or re-regulation of genes that does not significantly affect each gene’s average expression. Thus, identifying other types of differential regulation may increase our ability to distinguish between groups and provide insight into their distinct etiologies (for a discussion, see de la Fuente, 2010). In our setting, gene pairs are either equivalently co-expressed (EC) or DC. When there are three or more conditions, there are many ways to be DC and hence there will be multiple DC ‘classes’. While the approach implemented in EBcoexpress provides a false discovery rate (FDR) controlled list of significant DC gene pairs without sacrificing power. It is applicable within a single study as well as across multiple studies. For more information on the underlying theory, simulations and an application, please see our original paper in Biometrics. For a fully worked example with details at each step of the analysis, please see the vignette that accompanies the R/EBcoexpress package.

2 FEATURES

R/EBcoexpress calculates posterior probabilities for all EC/DC classes by assuming a Bayesian framework for the generation of correlations across conditions for all pairs and estimating the hyperparameters of that framework using an Expectation-Maximization (EM) algorithm. We highlight a few aspects of this process:

- **Customizable correlation computations:** The analysis requires correlations for some set of gene pairs; we outsource the computations to C for efficiency. Although Pearson’s correlation can be used, R/EBcoexpress defaults to the biweight midcorrelation, which is similar to Pearson’s statistic but is robust to outliers.

- **Customizable FDR control:** R/EBcoexpress outputs a (no. of gene pairs)-by-(no. of classes) matrix of posterior probabilities for all EC/DC classes. The EC posterior probabilities may be used to generate a ‘hard threshold’ version of FDR-control; however, as this approach is somewhat more conservative than necessary and hence less powerful, the package provides a function that provides ‘soft threshold’ FDR-control which controls the posterior expected FDR.

- **Visualization:** R/EBcoexpress provides graphical representations of co-expression exhibited by the data. The user may call up expression data for a given pair and superimpose a ‘robust'
A single-study analysis requires a (no. of genes)-by-(no. of samples) matrix of expression values. These values should be normalized in the log2 scale. Associations on the log scale are instead important, remain on the log scale, and are often returned on the log-scale after normalization. Whether or not this is acceptable depends on the investigator: if associations between raw measurements are of interest, anti-log the data; if associatios on the log scale are instead important, remain on the log scale.

After EM computations are complete, as aforementioned every analysis produces a (no. of gene pairs)-by-(no. of classes) matrix of posterior probabilities for all EC/DC classes which may be used for FDR-control and, in analyses where there are three or more conditions, the availability of posterior probabilities for each DC class allows further assortment and inference among the DC gene pairs. This output may inform visualization choices as previously described. Additionally, a function is provided that returns the number of times each gene is included in a DC pair, given a threshold; this information may be used to identify genes that exhibit ‘differential hubbing’.

4 SUMMARY

R/EBCoexpress provides a simple interface inside the R statistical programming language for the identification and exhibition of DC gene pairs.

Conflict of Interest: none declared.

REFERENCES


