The Journal of Open Source Software

Compare Expression Profiles for Pre-defined Gene Groups with C-REx

Mingze He[1, 2], **Kokulapalan Wimalanathan**[1, 2], **Peng Liu**[1, 3], **and Carolyn J. Lawrence-Dill**[1, 2, 4]

1 Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA, 50011 **2** Department of Agronomy, Iowa State University, Ames, Iowa, USA 50011 **3** Department of Agronomy, Iowa State University, Ames, Iowa, USA 50011 **4** Department of Agronomy, Iowa State University, Ames, Iowa, USA 50011

DOI: 10.21105/joss.01255

Software

- Review
- Repository
- Archive

Submitted: 28 January 2019
Published: 01 May 2019

Summary

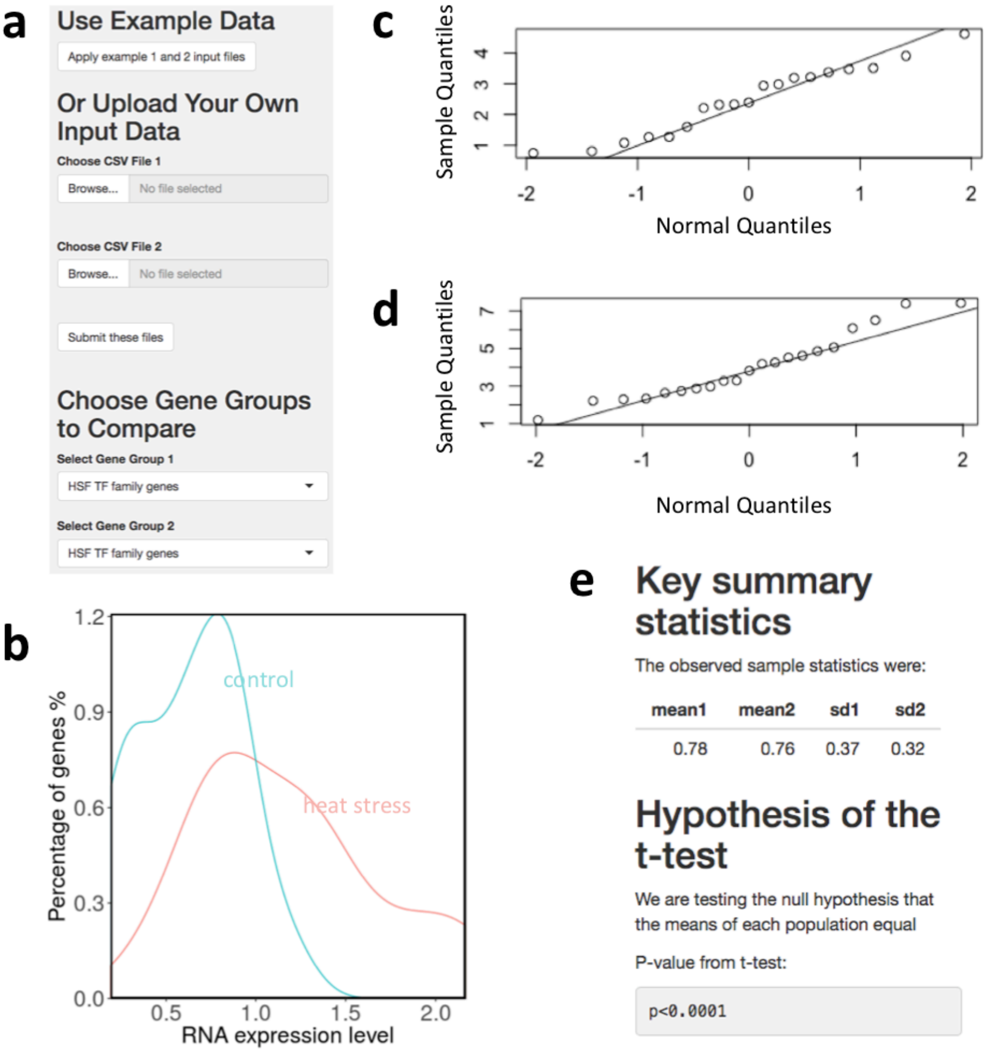
Most gene expression analysis methods discover groups of genes that are co-expressed, rather than testing whether a specified gene group behaves in a concerted manner. We implemented a novel statistical method designed to assess significance of differences in RNA expression levels among specified groups of genes. Our Shiny web application C-REx (Comparison of RNA Expression) enables researchers to readily test hypotheses about whether specific gene groups share expression profiles and whether those profiles differ from those of other groups of genes. We implemented data transformation, a normality visualizer, and both parametric and non-parametric tests for determining whether gene groups are functioning in concert or in contrast both within and between conditions. Here, we demonstrate that the C-REx application recovers well-known biological phenomena (e.g., response to heat stress).

Statement of need

RNA-seq-based gene expression levels can be variable to the extreme (Conesa et al., 2016). Many sophisticated methods have been developed and implemented to reduce noise in datasets (Ding et al., 2015; Leek, Johnson, Parker, Jaffe, & Storey, 2012; Stegle, Parts, Piipari, Winn, & Durbin, 2012) and to compare and define sets of differentially expressed genes (DEGs)(Anders & Huber, 2010; Jelle J Goeman & Bühlmann, 2007; Jelle J. Goeman, Oosting, Cleton-Jansen, Anninga, & Houwelingen, 2005; B. Li & Dewey, 2011; Patro, Duggal, Love, Irizarry, & Kingsford, 2017; Robinson & Oshlack, 2010; Trapnell et al., 2012). For the most common types of expression analyses, DEG sets are identified as those genes for which log2 (treatment/control) differences are > 1. Next, the identified set of DEGs is analyzed to find shared functional characteristics, often via gene ontology (GO) enrichment (Thomas et al., 2003). This results in discovery of genes that share expression profiles alongside shared functional annotations for that gene group. While this method helps to form gene groups and to figure out what functional characteristics the genes identified have in common, it does not enable specific hypothesis testing. For example, if a group of twelve genes are all involved in a particular biochemical pathway, a researcher cannot use enrichment to determine whether that gene group's expression changes are unique from other genes or gene groups. To enable this kind of assessment, we developed a method that determines whether specified groups of genes are similar (or different) in their expression patterns (He, Liu, & Lawrence-Dill, 2018). To do this, we changed the dimension of comparison and treated each gene as a variable and compare groups rather than relying on defining DEGs individually. Our method uses log-transformed gene expression values, which are nearly Gaussian in distribution. If a normality requirement is met, Student's t-test can be applied to assess the significance of differences among groups of genes between samples or treatments. If not, Wilcoxon signed-rank test can be applied. Here, we describe C-REx, an application that implements this method

Example usage case: expression differences for specified gene groups in heat stress

To use C-REx, first choose to carry out a 'within sample' or a 'between sample' comparison. Next, upload or select expression input file(s) from the examples provided (see format details in Supplementary Materials). For within sample comparisons, a single file is uploaded whereas between-sample comparisons require two input files. In Figure 1, panel a, two preloaded example datasets from maize (heat stressed and non-stressed; (Makarevitch et al., 2015)) are analyzed via the between-sample comparison (described in detail by (He et al., 2018)). Expression input files specify gene sets by name, including a set of designated housekeeping genes, which are used for sample normalization. Once input files are specified, dropdowns entitled "Choose Gene Groups to Compare" are populated by the gene group names specified in the input files. In panel b, differences in expression of heat shock factor transcription factor (HSF TF) genes between stress and non-stress conditions are shown. The curves are calculated as follows. The C-REx tool generates log-transformed gene expression values, normalizes gene expression values based on housekeeping gene expression means, and graphs the normalized and transformed expression value distributions. As shown in panels c and d, a Q-Q plot is created to allow the user to assess whether the normality assumption has been met for each generated distribution (therefore indicating that parametric statistics can be used). If the data satisfies the normality requirement, the user clicks the "Student's t-test" tab to generate panel e, a "Key summary statistics" report including mean, standard deviation, and a p-value. For cases where the normality assumption is not met, the Wilcoxon signed-rank test is implemented and available. The analysis in Figure 1 indicates that the HSF TF gene group is significantly up-regulated in maize under heat stress. Mathematical details of the method are outlined in (He et al., 2018).



Key summary statistics

The observed sample statistics were:

mean1	mean2	sd1	sd2
0.78	0.76	0.37	0.32

Hypothesis of the t-test

We are testing the null hypothesis that the means of each population equal

P-value from t-test:

```
p<0.0001
```

Figure 1 : C-REx analysis between heat stress and non-stress conditions for maize HSF TF. **(a)** Click "Apply example 1 and 2 input files" and choose "HSF TF family genes" for each sample. Click the "Gene distribution" tab to produce **(b)** the expression level density plot (non-stress-green, heat stress-pink). Inspect whether the transformed data satisfy the normality requirement by selecting the "normality test" tab. Heat stress shown in **(c)** and non-stress shown in **(d)** with each log-transformed expression value shown as a black circle. Diagonal indicates perfect concordance between the normal distribution and transformed expression values. Click "Student's t-test" tab for **(e)** a statistical summary with means, standard deviation, and p-values.

Acknowledgements

We thank Carla Mann and Ian Braun for helpful discussion and suggestions as well as Scott Zarecor and Darwin Campbell for technical assistance.

Funding

This work has been supported by the Iowa State University Plant Sciences Institute Faculty Scholars program.

Conflict of Interest: none declared.

References

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106. doi:10.1186/gb-2010-11-10-r106

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., et al. (2016). A survey of best practices for RNA-seq data analysis. doi:10.1186/s13059-016-0881-8

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., et al. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, *31*(13), 2225–2227. doi:10.1093/bioinformatics/btv122

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, *23*(8), 980–987. doi:10.1093/bioinformatics/btm051

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K., & Houwelingen, H. C. van. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, *21*(9), 1950–1957. doi:10.1093/bioinformatics/bti267

He, M., Liu, P., & Lawrence-Dill, C. J. (2018). A hypothesis-driven approach to assessing significance of differences in RNA expression levels among specific groups of genes. doi:10.1016/j.cpb.2017.12.003

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883. doi:10.1093/bioinformatics/bts034

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, *12*(1), 323. doi:10.1186/1471-2105-12-323

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., & Springer, N. M. (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genetics*, *11*(1). doi:10.1371/journal.pgen.1004915

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*, R25. doi:10.1186/gb-2010-11-3-r25

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507. doi:10.1038/nprot.2011.457

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., et al. (2003). PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. doi:10.1093/nar/gkg115

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562–578. doi:10.1038/nprot.2012.016