Bioinformatics, 34(9), 2018, 1589–1590
doi: 10.1093/bioinformatics/btx835
Advance Access Publication Date: 22 December 2017
Applications Note

OXFORD

Gene expression

myTAI: evolutionary transcriptomics with R

Hajk-Georg Drost[1,2,*], Alexander Gabel[2], Jialin Liu[3], Marcel Quint[4] and Ivo Grosse[2,5]

[1]Sainsbury Laboratory Cambridge, University of Cambridge, Cambridge CB2 1LR, UK, [2]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany, [3]Université de Lausanne, Département d'Ecologie et d'Evolution, Quartier Sorge, 1015 Lausanne, Switzerland, [4]Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany and [5]German Center for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on August 28, 2017; revised on December 21, 2017; editorial decision on December 17, 2017; accepted on December 21, 2017

Abstract

Motivation: Next Generation Sequencing (NGS) technologies generate a large amount of high quality transcriptome datasets enabling the investigation of molecular processes on a genomic and metagenomic scale. These transcriptomics studies aim to quantify and compare the molecular phenotypes of the biological processes at hand. Despite the vast increase of available transcriptome datasets, little is known about the evolutionary conservation of those characterized transcriptomes.

Results: The *myTAI* package implements exploratory analysis functions to infer transcriptome conservation patterns in any transcriptome dataset. Comprehensive documentation of *myTAI* functions and tutorial vignettes provide step-by-step instructions on how to use the package in an exploratory and computationally reproducible manner.

Availability and implementation: The open source *myTAI* package is available at https://github.com/HajkD/myTAI and https://cran.r-project.org/web/packages/myTAI/index.html.

Contact: hgd23@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

To investigate phenotypic changes, diseases, environmental stresses, or developmental processes, transcriptome studies are often the approach of choice. Although transcriptomics is based on a solid methodology, little is known about the evolutionary conservation and dynamics of transcriptomes across species (Drost *et al.*, 2017). Understanding the evolutionary processes that change transcriptomes over time, however, might lead to new insights on how diseases emerge or how phenotypic changes are caused by changes in transcriptomes. For this purpose, evolutionary transcriptomics studies aim to capture and quantify the evolutionary conservation of transcriptomes during specific stages of the biological process of interest (Domazet-Lošo and Tautz, 2010). Here, we present the exploratory analysis package *myTAI*, which can combine evolutionary information of genes with their transcript levels to infer transcriptome conservation patterns. Evolutionary information is given as input to the package and can range from classical phylogenetic or orthology relationships between genes to more recent approaches such as phylogenetic comparative methods (PCMs) (Dunn *et al.*, 2013), phylogenetic reconciliation methods (Doyon *et al.*, 2011), or phylostratigraphy (Domazet-Lošo *et al.*, 2007). In summary, starting with a pre-computed table of *gene age* information and a transcriptome dataset, the R package *myTAI* can be used to screen for stages of high or low transcriptome conservation within a biological process of interest. If highly conserved or variable transcriptomes were found in particular stages or treatments, more specialized experimental studies could subsequently be designed to investigate the functions and mechanistic implications of these conserved or variable stages.

2 Implementation

The R package *myTAI* is released under the GNU General Public License within the CRAN project (R Core Team). The package can be downloaded from https://cran.r-project.org/web/packages/myTAI/index.html. The source code is publicly available at https://github.com/HajkD/myTAI. Internal *myTAI* functions are implemented in C++ and integrated via the *Rcpp* (Eddelbuettel, 2013) Application Programming Interface (API) and unit tested using testthat. The *myTAI* package furthermore depends on the R packages *nortest*, *fitdistrplus* (Delignette-Muller and Dutang, 2015), *dplyr*, *RColorBrewer*, *taxize* (Chamberlain and Szöcs, 2013), *reshape2*, *ggplot2* (Wickham, 2009), *biomartr* (Drost and Paszkowski, 2017), *readr*, *tibble*, *scales* and *gridExtra*.

3 Functions and Examples

More than fifty functions are provided by the *myTAI* package. We recently used *myTAI* to investigate the developmental hourglass model of embryo development (Raff, 1996) on the transcriptomic level (Quint *et al.*, 2012; Drost *et al.*, 2017). Others used *myTAI* to investigate transcriptome conservation in plant organ development (Lei *et al.*, 2017). To illustrate an example workflow with *myTAI*, we here use the developmental transcriptome of *Arabidopsis thaliana* embryo development (Quint *et al.*, 2012) (see Supplementary Material for more details about data formats):

```
# Import the myTAI package and load example dataset
library(myTAI); data(PhyloExpressionSetExample)
```

One metric to quantify transcriptome conservation on a global scale is the Transcriptome Age Index (TAI) (Domazet-Lošo and Tautz, 2010), which denotes the average transcriptome age throughout the biological process of interest.

```
# Plot the Transcriptome Age Index of A. thaliana embryo development
PlotSignature (PhyloExpressionSetExample)
```

To quantify the transcript level of each gene age category to the overall transcriptome for each developmental stage, the gene expression level distributions for each gene age category can be visualized by:

```
# Plot gene expression level distributions
PlotCategoryExpr (PhyloExpressionSetExample,
                  legendName = "PS",
                  log.expr = TRUE)
```

A linear transformation of the mean expression levels into the interval [0, 1] enables the comparison of mean expression level patterns between gene age categories independent of their actual mean expression magnitude. A relative expression level of 0 denotes the minimum mean expression level compared to all other stages, and a relative expression level of 1 denotes the maximum mean expression level compared to all other stages:

```
# Plot relative expression levels
PlotRE (PhyloExpressionSetExample,
        Groups = list (c(1: 3), c(4: 12)),
        legendName = "PS",
        adjust.range = TRUE)
```

Finally, we compare relative expression levels between groups of age categories and quantify their difference:

```
# Compare relative expression levels between groups of age categories
PlotBarRE (PhyloExpressionSetExample,
           Groups = list(group_1 = 1: 3, group_2 = 4: 12),
           xlab = "Ontogeny",
           ylab = "Mean Relative Expression",
           cex = 1.5)
```

In addition to these exploratory functions, *myTAI* provides functionality for taxonomic information retrieval, *gene age* enrichment analyses, differential gene expression analyses of age categories, and additional metrics for quantifying trancriptome conservation. A detailed description and interpretation of *myTAI* functions is available at https://github.com/HajkD/myTAI#tutorials and also in the Supplementary Material.

4 Conclusions

Evolutionary transcriptomics studies can serve as a first approach to screen *in silico* for the potential existence of evolutionary constraints within a biological process of interest. This is achieved by quantifying transcriptome conservation patterns and their underlying gene sets in biological processes. The exploratory analysis functions implemented in *myTAI* provide users with a standardized, automated and optimized framework to investigate evolutionary signatures in any transcriptome dataset of interest.

Funding

We thank ERC grant EVOBREED [grant number 322621], the SKWP Research Foundation, and the German Science Foundation (grants Qu 141/5-1, Qu 141/6-1, Qu 141/7-1, GR 3526/6-1, GR 3526/7-1, GR 3526/8-1, and FZT 118) for financial support.

Conflict of Interest: none declared.

References

Chamberlain,S.A. and Szöcs,E. (2013) taxize: taxonomic search and retrieval in R. *F1000Research*, **2**, 191.

Delignette-Muller,M.L. and Dutang,C. (2015) fitdistrplus: an R package for fitting distributions. *J. Stat. Softw.*, **64**, 1–34.

Domazet-Lošo,T. *et al.* (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.*, **23**, 533–539.

Domazet-Lošo,T. and Tautz,D. (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, **468**, 815–818.

Doyon,J.P. *et al.* (2011) Models, algorithms and programs for phylogeny reconciliation. *Brief. Bioinform.*, **12**, 392–400.

Drost,H.-G. *et al.* (2017) Cross-kingdom comparison of the developmental hourglass. *Curr. Opin. Genet. Dev.*, **45**, 69–75.

Drost,H.-G. and Paszkowski,J. (2017) Biomartr: genomic data retrieval with R. *Bioinformatics*, **33**, 1216–1217.

Dunn,C.W. *et al.* (2013) Phylogenetic analysis of gene expression. *Integr. Comp. Biol.*, **53**, 847–856.

Eddelbuettel,D. (2013) *Seamless R and C++ Integration with Rcpp*. Springer, New York, NY.

Lei,L. *et al.* (2017) Plant organ evolution revealed by phylotranscriptomics in *Arabidopsis thaliana*. *Sci. Rep.*, **7**, 7567.

Quint,M. *et al.* (2012) A transcriptomic hourglass in plant embryogenesis. *Nature*, **490**, 98–101.

Raff,R.A. (1996) *The Shape of Life: Genes, Development and the Evolution of Animal Form*. University Chicago Press, Chicago.

Wickham,H. (2009) *ggplot2*. Springer, New York, NY.