GOexpress: A R/Bioconductor package for the identification and visualisation of robust gene ontology signatures through supervised learning of gene expression data

Kévin Rue-Albrecht[1], Paul A. McGettigan[1], Belinda Hernández[2,4], Nicolas, C. Nalpas[1], David A. Magee[1], Andrew C. Parnell[2], Stephen V. Gordon[3,4] and David E. MacHugh[1,4]

[1] Animal Genomics Laboratory, UCD School of Agriculture and Food Science, University College Dublin, Dublin 4, Ireland. Email: kevin.rue@ucdconnect.ie

[2] UCD School Of Mathematical Sciences, University College Dublin, Dublin 4, Ireland.

[3] UCD School of Veterinary Medicine, University College Dublin, Dublin 4, Ireland.

[4] UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland.

Project Website: http://www.bioconductor.org/packages/release/bioc/html/GOexpress.html
Source Code: https://github.com/kevinrue/GOexpress
License: GPL (>= 3)

Main Text of Abstract

Background

The standardisation and decreasing cost of transcriptomics platforms has allowed for more complex experimental setups, including multiple experimental factors and levels. Identification of gene expression profiles that differentiate experimental groups is critical for discovery and analysis of key molecular pathways and also selection of robust diagnostic or prognostic biomarkers. While integration of differential expression statistics has been proposed to inform gene set enrichment analyses, such approaches are typically limited to single gene lists resulting from two-group comparisons or time-series analyses.

Results

We introduce GOexpress, a software package for scoring and summarising the ability of ontology-related genes to simultaneously classify samples from multiple experimental groups. GOexpress integrates normalised gene expression data (*e.g.* from microarray and RNA-seq experiments) and phenotypic information of individual samples with gene ontology annotations to derive a ranking of genes and gene ontologies using a supervised learning approach. The default random forest algorithm allows interactions between all experimental factors, and competitive scoring of expressed gene features to evaluate their relative importance in clustering the predefined groups of samples.

Conclusions

GOexpress enables rapid identification and visualisation of robust ontology-related gene panels that robustly classify groups of samples, and supports both categorical (e.g. infection, treatment) and continuous (e.g. time-series, drug concentrations) experimental factors. The use of standard Bioconductor extension packages and publicly available gene ontology annotations facilitates straightforward integration of GOexpress within existing analytical pipelines.