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
[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.

Supported by wellcome trust

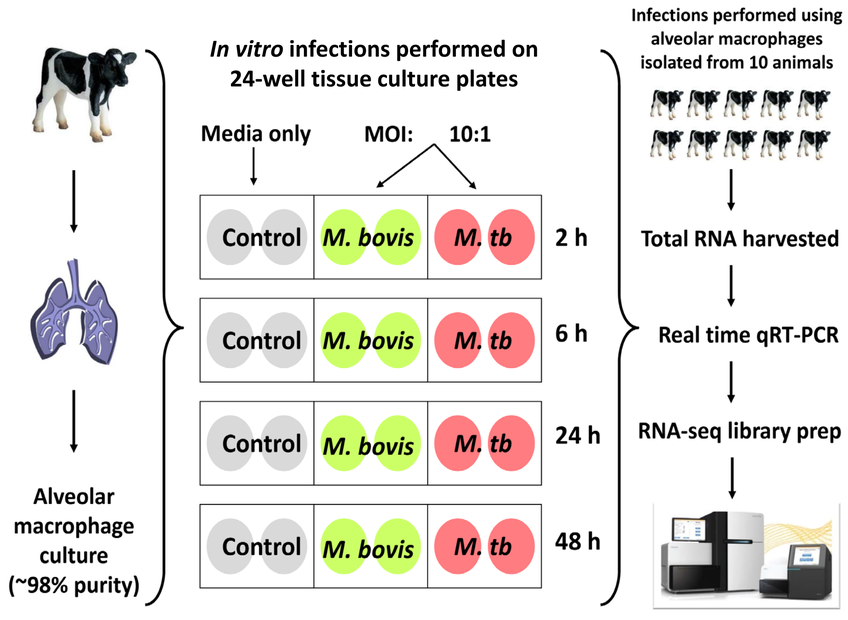
Project Background and Objectives

Identification of gene expression profiles that differentiate experimental groups is critical for discovery and analysis of key molecular pathways and also for selection of robust diagnostic or prognostic biomarkers. Typically, Gene-Set Enrichment Analyses (GSEA) are limited to single gene lists resulting from simple two-group comparisons or time-series analyses. We introduce GOexpress, a software package for scoring and summarising the capacity of gene ontology features to simultaneously classify samples from multiple experimental groups. GOexpress integrates quantitative data (*e.g.* from microarray and RNA-seq experiments) and phenotypic information with gene ontology annotations to derive a ranking of genes and gene ontologies using a supervised learning approach. The default random forest algorithm allows competitive scoring of expressed genes to evaluate their relative importance in classifying predefined groups of samples. GOexpress enables rapid identification and visualisation of ontology-related gene panels that robustly classify groups of samples and supports both categorical (*e.g.* infection status, 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 computational biology pipelines.

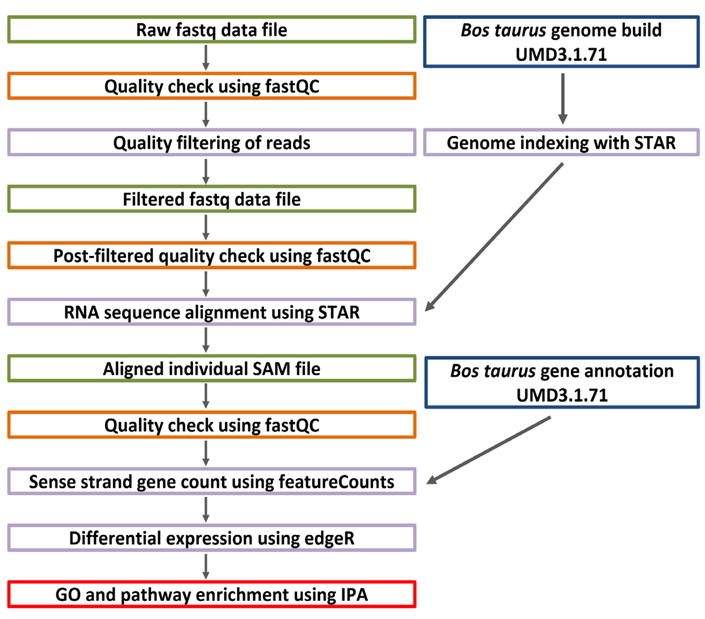
Here we illustrate the features of GOexpress using an RNA-seq data set generated from bovine alveolar macrophages stimulated across an infection time course with two different but closely related mycobacterial pathogens: (a) *Mycobacterium bovis*, the causative agent of tuberculosis in cattle and (b) *M. tuberculosis*, the causative agent of human tuberculosis, which does not cause disease in cattle.

Materials and Methods

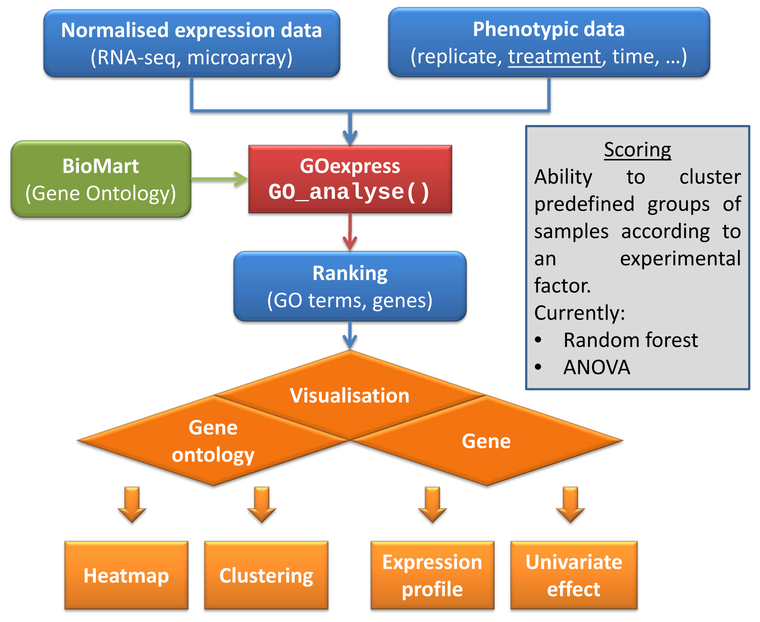
Experimental design



Bioinformatics analysis of RNA-seq data



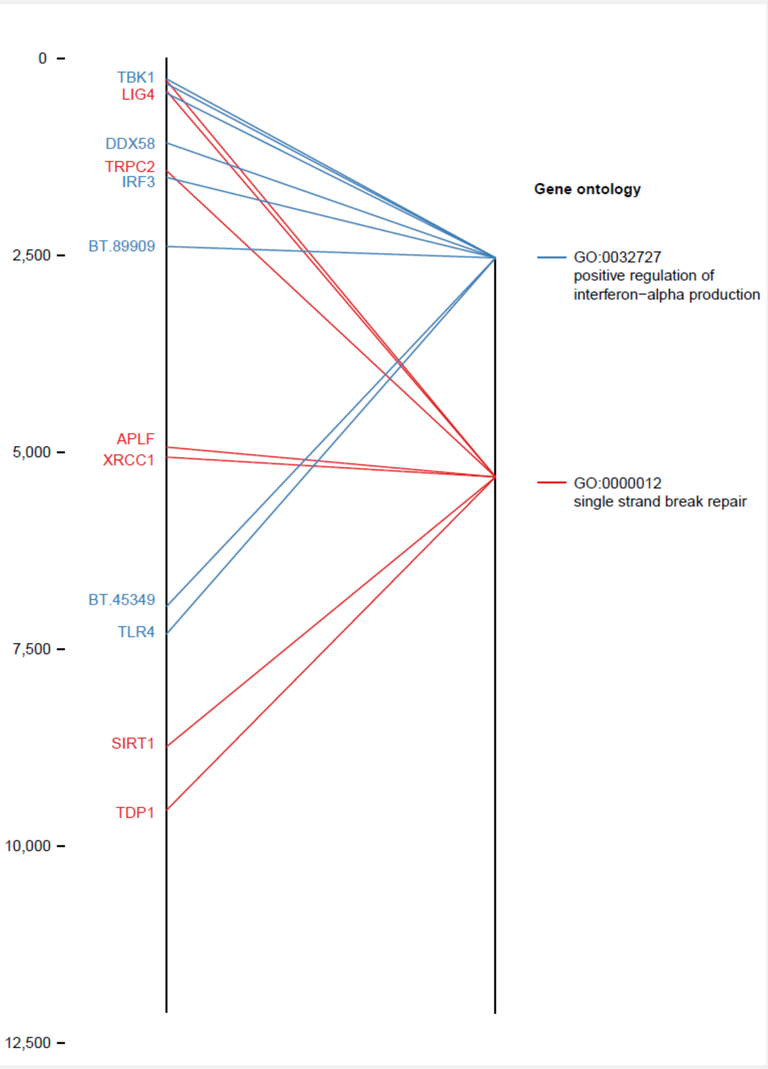
GOexpress analytical pipeline (https://github.com/kevinrue/GOexpress)



Results

1) Ranking of filtered GO terms by summarisation of gene ranks.

Gene ontologies are scored using the average rank of their associated genes.



2) Gene ontology analysis using GOexpress

A) Biological processes that best cluster control, *M. tuberculosis* and *M. bovis* treatments (*filter: 7+ associated genes*)

GO ID	Avg rank	Avg score	Total count	Data count	GO name
GO:0032727	2533.875	0.014	8	8	positive regulation of interferon-alpha production
GO:0006401	2823.125	0.016	8	8	RNA catabolic process
GO:0035518	3177.857	0.011	7	7	histone H2A monoubiquitination
GO:0001841	3364.583	0.011	12	12	neural tube formation
GO:0050718	3640.333	0.014	9	9	positive regulation of interleukin-1 beta secretion

B) Molecular functions that best cluster control, *M. tuberculosis* and *M. bovis* treatments (*filter: 7+ associated genes*)

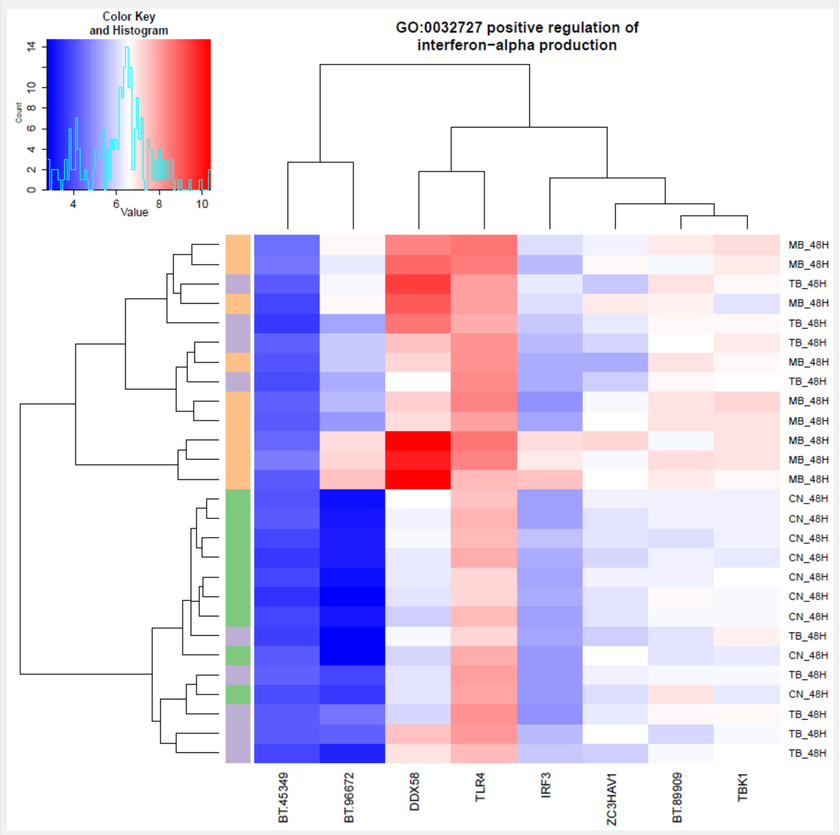
GO ID	Avg rank	Avg score	Total count	Data count	GO name
GO:0070530	4024.5	0.010	8	8	K63-linked polyubiquitin binding
GO:0070628	4375.714286	0.005	7	7	proteasome binding
GO:0070411	4415.090909	0.007	11	10	I-SMAD binding
GO:0042809	4419.714286	0.007	7	7	vitamin D receptor binding
GO:0008440	4644.857143	0.008	7	7	inositol-1,4,5-trisphosphate 3-kinase activity

C) Cellular components that best cluster control, *M. tuberculosis* and *M. bovis* treatments (*filter: 10+ associated genes*)

GO ID	Avg rank	Avg score	Total count	Data count	GO name
GO:0031982	3302.526	0.013	19	19	vesicle
GO:0022624	4767.2	0.008	15	15	proteasome accessory complex
GO:0031519	4959.04	0.007	25	24	PcG protein complex
GO:0016581	5133.813	0.005	16	15	NuRD complex
GO:0000123	5226.056	0.005	18	18	histone acetyltransferase complex

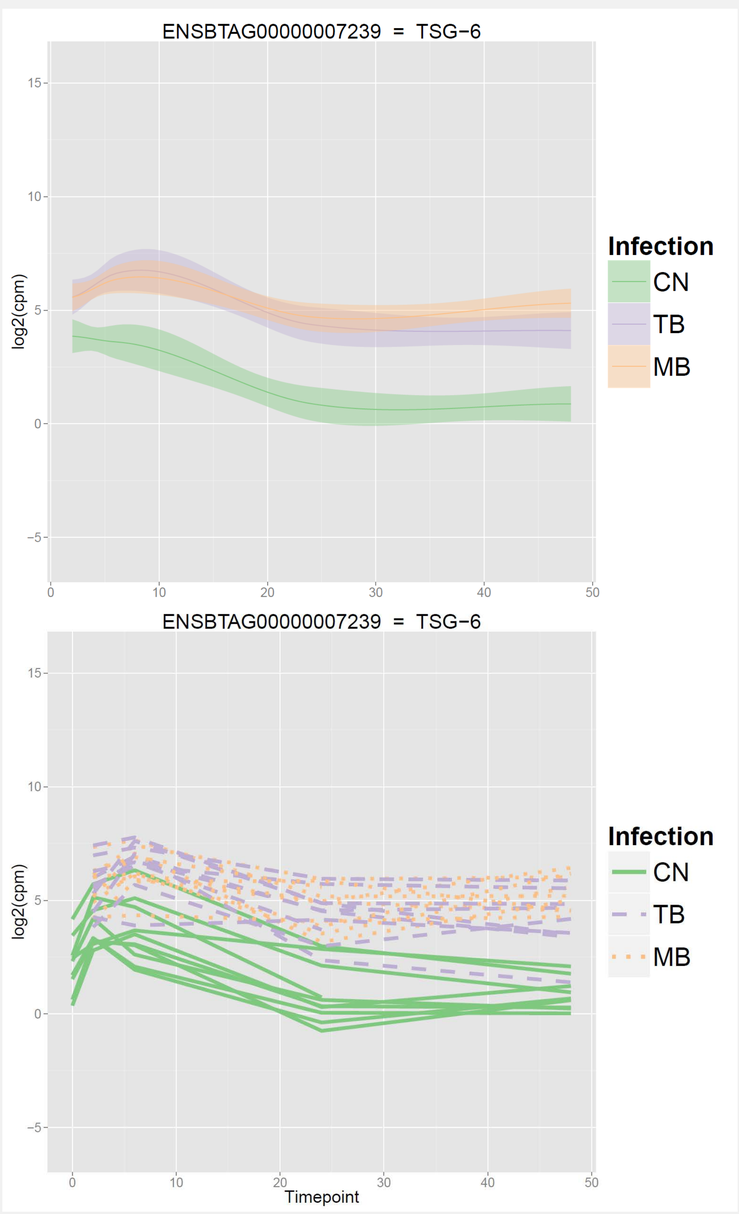
3) Gene ontology overview using GOexpress

http://www.bioconductor.org/packages/release/bioc/html/GOexpress.html



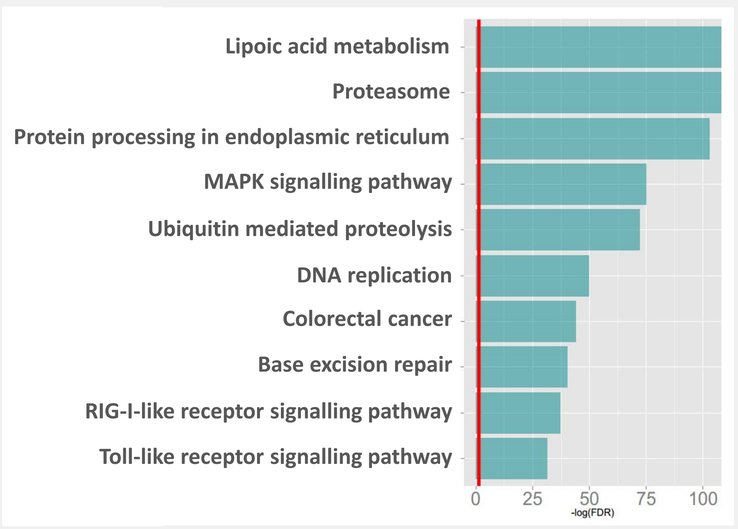
4) Gene expression profiling using GOexpress

Top ranked



5) Comparative canonical pathways analyses

Genes differentially expressed between *M. bovis*- and *M. tuberculosis*-infected macrophages at 48 hours post-infection



Pathways analysis using SIGORA (Foroushani *et al.,* 2013)

Conclusions and Perspectives

- GOexpress allows rapid identification of gene features that best discriminate biological samples according to a given grouping factor.
- The capacity of individual gene features to discriminate groups of samples can be summarised at the gene ontology level.
- GOexpress provides functions to visualise individual gene expression profiles and groups of ontology-related genes
- The algorithms provided (random forest and ANOVA) can be easily extended to include different methods.