Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty721
Advance Access Publication Date: 20 August 2018
Applications Note

Gene expression

ImaGEO: integrative gene expression meta-analysis from GEO database

Daniel Toro-Domínguez[1,2,†], Jordi Martorell-Marugán[1,†], Raúl López-Domínguez [1], Adrián García-Moreno[1], Víctor González-Rumayor[3], Marta E. Alarcón-Riquelme[2,4,*] and Pedro Carmona-Sáez[1,*]

[1]Bioinformatics Unit, [2]Area of Medical Genomics, GENYO Centre for Genomics and Oncological Research: Pfizer/University of Granada/Andalusian Regional Government, PTS Granada, 18016 Granada, Spain, [3]Atrys Health, Barcelona 08025, Spain and [4]Unit of Chronic Inflammatory Diseases, Institute of Environmental Medicine, Karolinska Institutet, 17177 Stockholm, Sweden

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Janet Kelso

Received on February 15, 2018; revised on July 17, 2018; editorial decision on August 16, 2018; accepted on August 17, 2018

Abstract

Summary: The Gene Expression Omnibus (GEO) database provides an invaluable resource of publicly available gene expression data that can be integrated and analyzed to derive new hypothesis and knowledge. In this context, gene expression meta-analysis (geMAs) is increasingly used in several fields to improve study reproducibility and discovering robust biomarkers. Nevertheless, integrating data is not straightforward without bioinformatics expertise. Here, we present ImaGEO, a web tool for geMAs that implements a complete and comprehensive meta-analysis workflow starting from GEO dataset identifiers. The application integrates GEO datasets, applies different meta-analysis techniques and provides functional analysis results in an easy-to-use environment. ImaGEO is a powerful and useful resource that allows researchers to integrate and perform meta-analysis of GEO datasets to lead robust findings for biomarker discovery studies.
Availability and implementation: ImaGEO is accessible at http://bioinfo.genyo.es/imageo/.
Contact: marta.alarcon@genyo,es or pedro.carmona@genyo.es
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Due to the increasing use of high-throughput techniques, the amount of information available in biomedical databases is growing exponentially. In particular, Gene Expression Omnibus (GEO) (Barrett *et al.*, 2013) is a public gene expression repository that contains more than 94 000 datasets and over 2 million of samples. This is an invaluable resource that, with the appropriate methods and tools, can be exploited to integrate gene expression data for applications such as biomarker discovery (Toro-Domínguez *et al.*, 2014), disease classification or phenotype comparisons (Carmona-Sáez *et al.*, 2017), among others. Several software tools have been developed to take advantage of this information. GEO2R was originally available in GEO portal to allow researchers without computational skills to perform differential expression analysis in individual datasets. In the last few years, tools such as ShinyGEO (Dumas *et al.*, 2016) or ScanGEO (Koeppen *et al.*, 2017) have extended some of the GEO2R functionalities to explore, retrieve and analyze gene expression data in an easy-to-use environment.

However, this amount of data offers new possibilities beyond the analysis of individual datasets. In this context, there is an increasing number of studies that integrate different datasets to perform gene expression meta-analysis (geMAs). This technique is

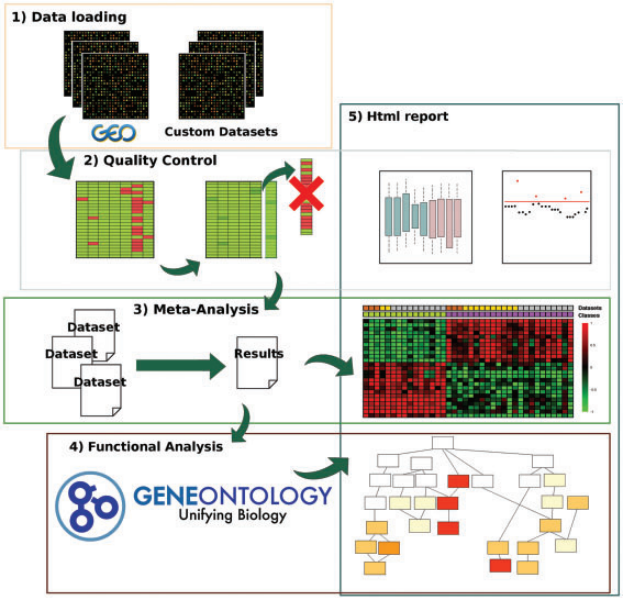

Fig. 1. Workflow of ImaGEO. The image summarizes and orders the five modules of ImaGEO. 1) Data input from GEO or custom data. 2) Quality Control is performed for each dataset followed by sample/gene filtering. 3) Gene expression Meta-analysis. 4) Functional analysis. 5) Results in html report

usually applied to increase the sample size but it can be also used to integrate datasets from different phenotypes in order to discover common biomarkers (Toro-Domínguez *et al.*, 2014). In this context, there are different tools for geMAs such as INMEX (Xia *et al.*, 2013) or ExAtlas (Sharov *et al.*, 2015), but they lack from a complete workflow starting from GEO identifiers that, at the same time, requires a minimal user interaction in terms of data processing. A detailed comparative analysis of available tools is provided in the additional material.

In this work, we present ImaGEO, a web-based application to perform a complete geMAs starting from GEO identifiers. The application provides a step-by-step workflow that guides the user through the entire analysis accelerating the re-use of publicly available gene expression data for biomarker discovery purposes. The application currently supports a curated set of platforms from Illumina, Affymetrix and Agilent for human and model organisms including *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Danio rerio* and *Mus musculus* and others such as *Rattus norvegicus* and *Pseudomonas aeruginosa* (see the online help with a complete list of supported platforms).

2 Materials and methods

ImaGEO has been developed in Shiny, a web application framework for R. Internally, it is divided in five modules (Fig. 1): (i) *Data loading and processing module* where users can enter the GEO IDs or upload custom datasets and establish the parameters of the analysis. GEO data is retrieved and processed using GEOquery package (Davis and Meltzer, 2007). Expression values are transformed to logarithmic scale unless they already are and probe identifiers are annotated to unique gene identifiers. (ii) *Quality control module* that shows data metrics and quality control checking. (iii) *Meta-analysis module* contains a total of nine different meta-analysis methods adapted from functions contained in MetaDE R package, which are effect size (ES), Fisher's, Stouffer's, adaptively-weighted, sum or product of ranks and the selection of minimum or maximum *P*-value across results. (iv) *Functional analysis module*. Enrichment analysis of Gene Ontology terms is performed in the list of over- and under-expressed genes obtained in the meta-analysis. (v) *Report Module:* where an html report is generated to explore all results using Nozzle.R1 R package. The report is divided in four sections that summarize the results of each step. First, a summary section contains an overview of the data and the analysis parameters used. Second, the quality control study shows the distribution of the expression values in boxplots and the missing values for each dataset along with a comparison pre- and post-quality control. Third, the results section displays an interactive table with significant genes annotated with Gene Symbol identifiers, gene names, *P*-values, corrected *P*-values and fold-change values. In addition, heatmaps of the top 100 and all significant genes are available. Finally, if the user chooses the enrichment analysis its results are provided in table format. A detailed documentation of methods and results can be found in the application web site.

3 Case study

As a working example we provide a use case that identify genes deregulated in opposite directions among lung cancer (LC) and Alzheimer (AD), two diseases that display inverse co-morbidity according to epidemiological data (Sánchez-Valle *et al.*, 2017). This is another type of application of geMAs that can be easily conducted in ImaGEO and can be applied to analyze inverse gene expression patterns among phenotypes, for example for drug repurposing analysis. As input we used AD (GEO IDs: GSE5281, GSE48350 and GSE4757) and LC datasets (GEO IDs: GSE33532, GSE19188, GSE19804, GSE7670 and GSE10072). To detect genes that were deregulated in opposite directions we simply switched group labels (cases/controls) in both diseases. Therefore, selecting ESs and random effects model as meta-analysis options we obtained 997 genes that were over-expressed in AD and under-expressed in LC (AD+/LC-) and 1220 genes with opposite patterns (AD-/LC+). Similarly to the results reported by Sanchez-Valle *et al.* functional analysis of AD+/LC-genes yielded biological pathways related to inflammatory responses and processes associated to AD-/LC+ genes were related to synaptic transmission and mitochondrial activities, that the authors stated could be implicated in the inverse co-morbidity between these diseases. This analysis was executed in a few minutes and is a good example of the potential of ImaGEO to perform a comprehensive geMAs. We are confident that it will be a useful application for the research community to exploit and re-use GEO data for deriving new biological knowledge and hypothesis generation.

Acknowledgements

This work is part of the thesis of JMM and DTD in the doctorate program of Biomedicine of the University of Granada.

Funding

This work was partially supported by Junta de Andalucía (PI-0173-2017). JMM was partially funded by Ministerio de Economía, Industria y Competitividad.

Conflict of Interest: none declared.

References

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

Carmona-Sáez,P. *et al*. (2017) Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. *Bioinformatics*, **33**, 3691–3695.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.

Dumas,J. *et al*. (2016) shinyGEO: a web-based application for analyzing gene expression omnibus datasets. *Bioinformatics*, **32**, 3679–3681.

Koeppen,K. *et al*. (2017) ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics*, **33**, 3500–3501.

Sánchez-Valle,J. *et al*. (2017) A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's Disease, Glioblastoma and Lung cancer. *Sci. Rep*., **7**, 4474.

Sharov,A.A. *et al*. (2015) ExAtlas: an interactive online tool for meta-analysis of gene expression data. *J. Bioinform. Comput. Biol*., **13**, 1550019.

Toro-Domínguez,D. *et al*. (2014) Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. *Arthritis Res. Ther*., **16**, 489.

Xia,J. *et al*. (2013) INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res*., **41**, W63–W70.