Data and text mining

GOplot: an R package for visually combining expression data with functional analysis

Wencke Walter[1], Fátima Sánchez-Cabo[2,†,*] and Mercedes Ricote[1,†,*]

[1] Department of Cardiovascular Development and Repair, [2] Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain.

ABSTRACT

Summary: Despite the plethora of methods available for the functional analysis of omics data, obtaining comprehensive-yet detailed understanding of the results remains challenging. This is mainly due to the lack of publicly available tools for the visualization of this type of information. Here we present an R package called GOplot, based on ggplot2, for enhanced graphical representation. Our package takes the output of any general enrichment analysis and generates plots at different levels of detail: from a general overview to identify the most enriched categories (bar plot, bubble plot) to a more detailed view displaying different types of information for molecules in a given set of categories (circle plot, chord plot, cluster plot). The package provides a deeper insight into omics data and allows scientists to generate insightful plots with only a few lines of code to easily communicate the findings.

Availability: The R package GOplot is available via CRAN-The Comprehensive R Archive Network: http://cran.r-project.org/web/packages/GOplot. The shiny web application of the Venn diagram can be found at: https://wwalter.shinyapps.io/Venn/

Contact: fscabo@cnic.es; mricote@cnic.es

Supplementary information: A detailed manual of the package with sample figures can be found at https://wencke.github.io/

1 INTRODUCTION

Omics technologies have become standard tools in biological research for identifying and unraveling transcriptional networks, building predictive models, and discovering candidate biomarkers. Gene/protein/metabolomic expression data is especially challenging for investigators due to its high-dimensional nature. Exploratory data analysis techniques are used to get a first impression of the important characteristics of the dataset and to reveal its underlying structure. Statistical analyses are then performed to detect subsets of elements owing the ability to provide valuable insight into the underlying patterns of the investigated biological process. One common approach is to map the molecules to their associated biological annotations, e.g. gene ontology (GO) terms, and to further perform an enrichment analysis. Although many visualization methods, tools (Supek *et al.*, 2011) and packages (Yin *et al.*, 2012; Zhang *et al.*, 2013) have been developed, none of them enables the user to combine expression data with the results of functional analysis in a way that guarantees the preservation of the power of both analyses.

R is commonly used by the *omics* community to analyze high- dimensional expression data. We therefore developed the R package `GOplot` based on the implementation of the grammar of graphics (Wickham 2009). `GOPlot` follows the path of deductive reasoning to allow the user to go from the most general to the most specific details of the functional analysis results. The package implements novel, original, high-quality plots which also allow the integration of other quantitative information about molecules, ie the expression levels. `GOplot` improves understanding of *omics* data and aids the communication of biologically relevant findings in publications.

[*] To whom correspondence should be addressed.

† Joint senior authorship.

2 PACKAGE DESCRIPTION

`GOplot` imports various R packages and was built on ggplot2, one of the three graphic systems in R. The package makes use of the complexity of `ggplot2` to provide the user with a collection of pre-specified and multilayered charts. Each layer adds valuable information to the displayed context to convey the intended message. Two kinds of datasets are required as input: a list of selected molecules with their expression levels and the results of a functional analysis.
We implemented two kinds of function: preprocessing and plotting.

Preprocessing functions: Although using the preprocessing functions is not mandatory, it is highly recommended to ensure easy workflow. `circle_dat()` allows the user to easily combine expression and functional enrichment data and generates the appropriate input for most of the plotting functions. `chord_dat()` generates a binary matrix that assigns the molecules to each predefined functional term. This kind of binary matrix is used as input for `GOChord()`. The input format for the preprocessing functions can be checked with the help function in R.

Plotting functions: The package provides the user with a guide to explore the data and to select lists of elements and terms of interest. The exploratory part of the data analysis starts at a very general level with the `GOBubble()` and `GOBar()` plotting functions for comparative charts. Both charts display information about the significance of the enrichment (-$\log_{10}$ of the adjusted p-value) and the z-score of the term. `GOBar()` allows processes to be sorted by z-score or p-value in order to provide a better overview. The circles of the bubble plot are area-proportional to the number of molecules in the given category. Based on these charts a list of relevant terms can be selected. With `GOCircle()`, `GOChord()` and `GOCluster()`, the user can add quantitative molecular information to the terms of interest (Figure1).

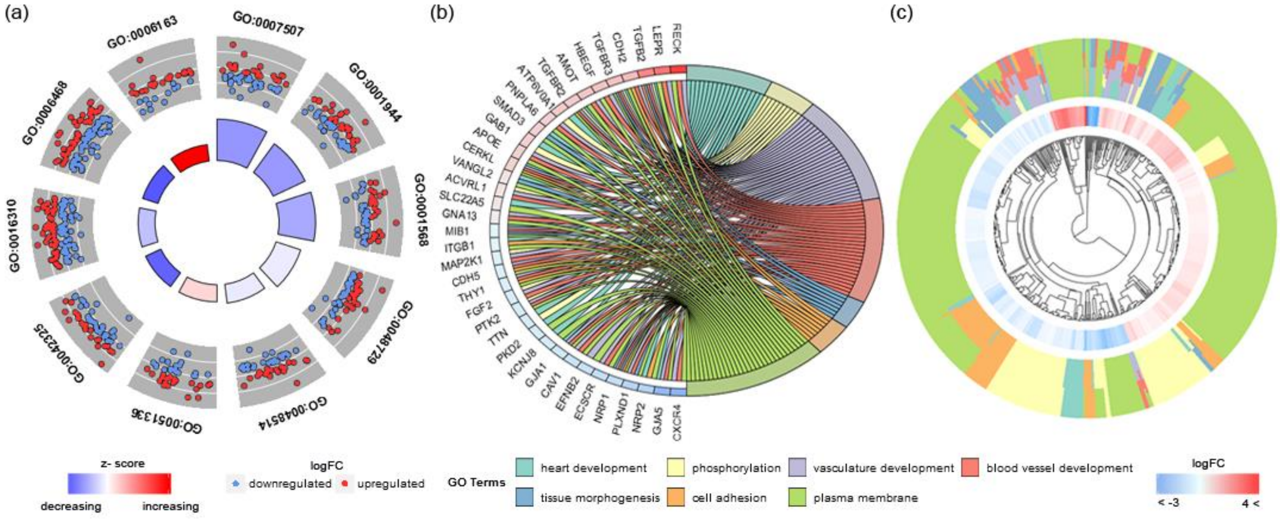

Fig. 1. **(a)** GOCircle plot; the inner ring is a bar plot where the height of the bar indicates the significance of the term (-$\log_{10}$ adjusted p-value), and color corresponds to the z-score. The outer ring displays scatterplots of the expression levels (logFC) for the genes in each term. (**b)** GOChord plot; the genes are linked via ribbons to their assigned terms. Blue-to-red coding next to the selected genes indicates logFC. **(c)** GOCluster plot displaying a circular dendrogram of the clustering of the expression profiles. The inner ring shows the color-coded logFC (up to three conditions), the outer ring the assigned functional terms.

With `GOVenn()` we have implemented a Venn diagram, that displays the number of overlapping elements as well as their expression patterns (commonly upregulated, commonly downregulated or contraregulated). In addition we have used shiny (Chang *et al.*, 2015), the web application framework for R, to create an interactive web application of `GOVenn()`.

3 EXAMPLE

This section briefly exemplifies some of the functionalities of the `GOplot` package. Further details of the available functions and their usage can be found in the `GOplot` vignette.
`GOplot` comes with a manually compiled sample dataset (EC). Selected samples were downloaded from the Gene Expression Omnibus (accession number: GSE47067). The data were normalized and a statistical analysis was performed to determine differentially expressed genes. The DAVID functional annotation tool (Huang *et al.*, 2009) was used to perform a gene-annotation enrichment analysis of the set of differentially expressed genes

(adjusted p-value < 0.05). Figure 1 shows three sample plots from the package created by the example code below.

```
> library(GOplot)
# Load the dataset
> data(EC)
# Generate the plotting object
> circ <- circle_dat(EC$david, EC$genelist)
# Generate the binary matrix
> chord <- chord_dat(circ, EC$genes, EC$process)
# Create the plots
> GOCircle(circ)
> GOChord(chord, ribbon.col = brewer.pal (7, ‘Set3’))
> GOCluster(circ, EC$process)
```

ACKNOWLEDGEMENTS

We thank Manuel Gómez (Bioinformatics Unit, CNIC) for his valuable contributions in the initial steps of this work and Carlos Torroja (Bioinformatics Unit, CNIC) for his comments that helped to improve and refine the implementation. We thank S. Bartlett (CNIC) for editorial assistance.

Funding: This work was supported by grants from the Spanish Ministry of Economy and Competitiveness (SAF2012-31483), Fundación Marató TV3 and the European Commision FP7 (CardioNext-ITN-608027) to MR. The CNIC is supported by the Spanish Ministry of Economy and Competitiveness and the Pro-CNIC Foundation.

REFERENCES

Huang DW, Sherman BT and Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*., **4(1)**, 44-57.

Nolan DJ et al. (2013) Molecular signatures of tissue- specific microvascular endothelial cell heterogeneity in organ maintenance and regeneration. *Dev Cell.*, **26(2)**, 204-19.

Supek F et al. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, **6.7**, e21800.

Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer- Verlag, New York.

Winston Chang et al. (2015) shiny: Web Application Framework for R. R package version 0.11.1 http://CRAN.R-project.org/package=shiny

Yin T, Cook T and Lawrence M (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol*, **13.8**, R77.

Zhang H, Meltzer PS and Davis SR (2013) RCircos: an R package for Circos 2D track plots. *BMC bioinformatics*, **14**, 244.