

GOCmpare: an R package to compare functional enrichment analysis results among gene lists and two species.

Chrystian C. Sosa^{1,2}, Diana Carolina Clavijo-Buriticá¹, Nicolas López-Rozo¹, Camila Riccio-Rengifo¹, Maria Victoria Diaz³, Victor Hugo García Merchán², Mauricio Alberto Quimbaya¹

¹ Departamento de Ciencias Naturales y Matemáticas, Pontificia Universidad Javeriana, Cali, Cali, Colombia

² Grupo de Investigación en Evolución, Ecología y Conservación EECO, Programa de Biología, Facultad de Ciencias Básicas y Tecnologías, Universidad del Quindío, Armenia, Colombia

³International Center for Tropical Agriculture (CIAT), Cali, Colombia

Email: ccsosaa@javerianacali.edu.co

Functional enrichment analysis is a cornerstone in bioinformatics since it is widely used to identify functional information associated with gene lists. Such analysis is based on the existence and agreement of a gene ontology (GO) community that has generated a standard and controlled vocabulary containing a hierarchical information on (i) molecular functions (FM), (ii) biological processes (BP), and (iii) cellular compartments (CC) available for annotated genes in different organisms. Functional enrichment approaches based on GO terms worked under a hypothesis test and multiple comparisons framework to obtain a filtered list of biological features as output. However, to date, most of the implemented tools to process gene lists using GO information use false discovery rate or gene ratio values to generate the functional enrichments. Nevertheless, none of the available tools compare multiple gene lists and GO enriched terms from an analytical approach. To prioritize GO terms and provide a straightforward way to analyze functional enrichment outputs, we introduce the R package GOCmpare (available at CRAN and <https://github.com/ccsosa/GOCmpare>). This R package comprises a descriptive computational procedure to obtain insights from several gene lists (categories), allowing to analyze from a comparative genomics perspective, either GO terms or categories by applying five different and complementary functions. The comparative strategy uses different tools: (i) undirected-weighted graphs, (ii) hypothesis tests, (iii) Jaccard distances, and (iv) visualizations in a principal coordinates analysis. Thus, the software provides a straightforward way to study biological problems where a comparative genomics approach is required. As a proof of concept we want to use the GOCmpare R package to understand from a functional perspective the main biological processes associated with aluminum tolerance in rice, introducing the model plant *Arabidopsis thaliana* as a genomic platform for comparison.