Deciphering functional annotation of multiple gene sets with MOTVIS

Aarón Ayllón-Benítez, Fleur Mougin, Manuel Quesada Martinez, Jesualdo Tomás Fernández Breis, Romain Bourqui, Patricia Thebault

▶ **To cite this version:**

Aarón Ayllón-Benítez, Fleur Mougin, Manuel Quesada Martinez, Jesualdo Tomás Fernández Breis, Romain Bourqui, et al.. Deciphering functional annotation of multiple gene sets with MOTVIS. Jornées ouvertes de biologie informatique et matematiques, Jul 2018, Marseille, France. hal-02401858

HAL Id: hal-02401858
https://hal.science/hal-02401858v1

Submitted on 10 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deciphering functional annotation of multiple gene sets with MOTVIS

A. Ayllón-Benítez[1,2], F. Mougin[1,2], M. Quesada-Martínez[3],

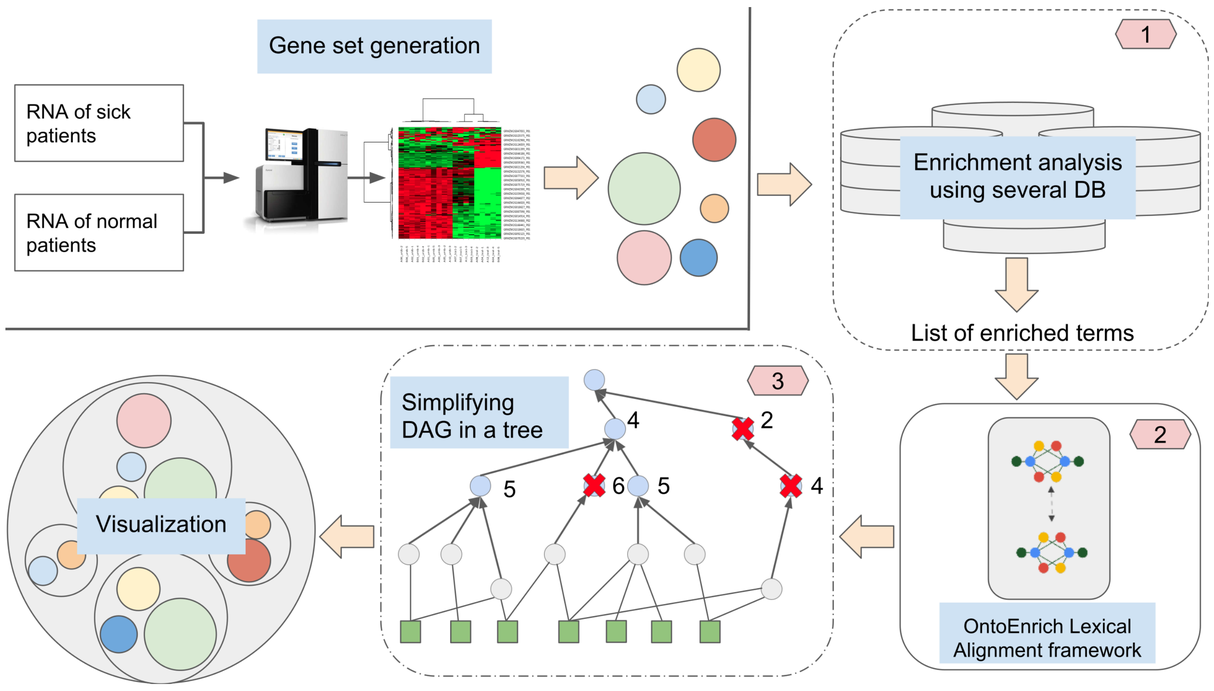
J.T. Fernández-Breis[4], R. Bourqui[1], P. Thébault[1,2]

[1]Univ. Bordeaux, LaBRI CNRS UMR 5800, France

[2]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, France

[3]Center of Operation Research (CIO), Universidad Miguel Hernández (UMH), Spain

[4]Universidad de Murcia, TECNOMOD group, Spain



1 - Context

- The advances in **omics technologies**, such as single-cell, RNA-SEQ or microarray facilitate **understanding the link between expression profiling and phenotype**.
- Statistical approaches or clustering aim at grouping genes according to their expression levels [1] to detect **gene signatures** and **understand the biological processes**.

For that, **annotating gene sets** will be crucial to:

- Elucidate the biological role of these specific cells.
- Highlight their specificity.

Problem:

- Managing a **large number of annotation terms** associated with a gene set is **very difficult**.
- **Enrichment methods** have been proposed [1] but they show an important pitfall related to **redundancy** in the results [2] due to the **lack or under-exploitation of semantic relations** between terms.

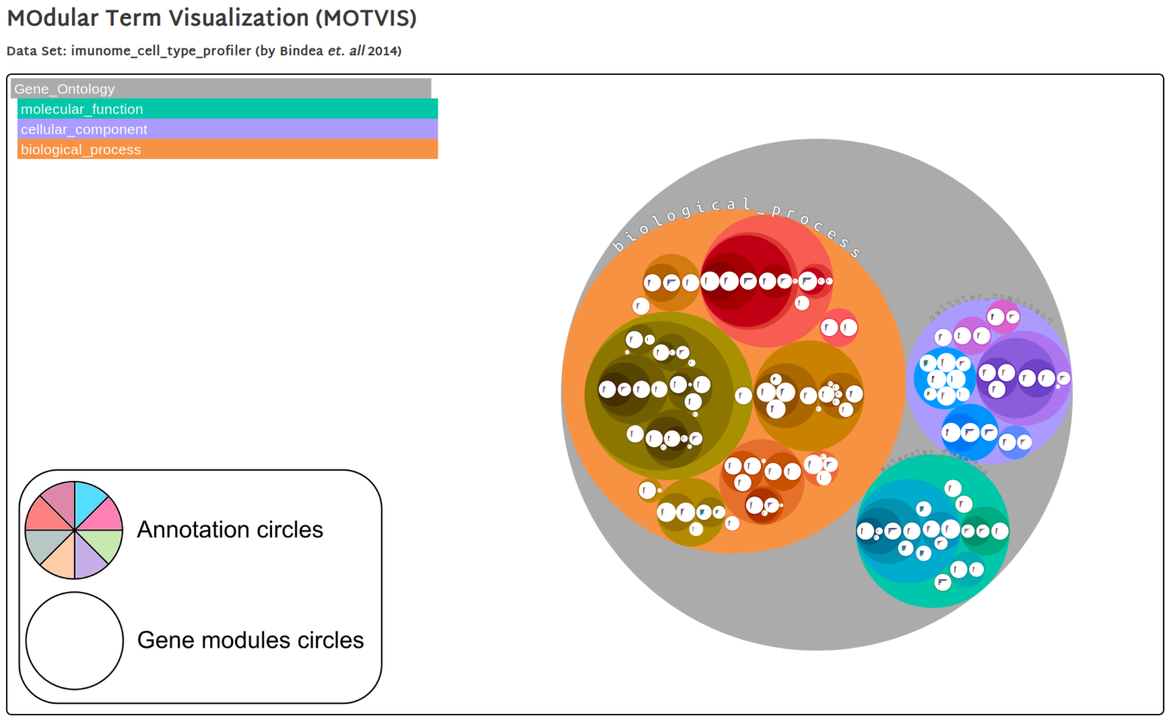
Objective

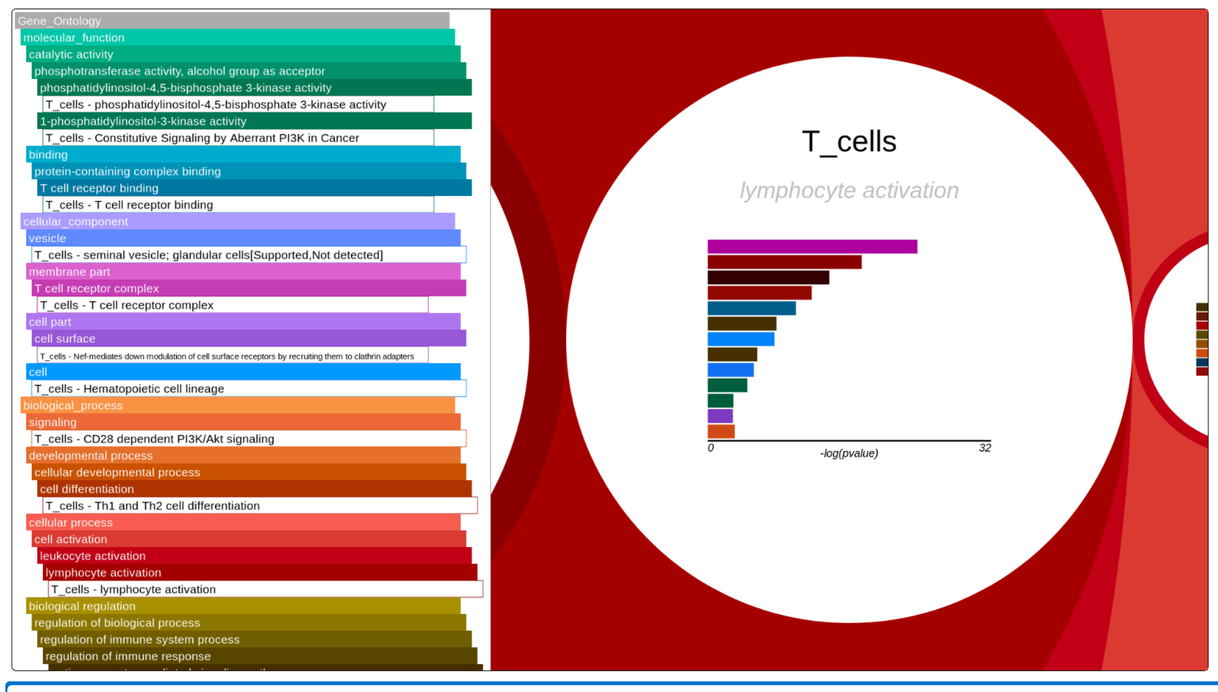
To illustrate the usefulness of the MOTVIS workflow to decipher the main roles in immune cell signatures

2 - Workflow

We proposed **three main steps** to compute gene sets annotations before presenting the visualization system [3]:

1. Gene sets are annotated using an enrichment approach. **g:Profiler** has been chosen because it uses **several annotation databases.** This enables to combine complementary knowledge for enriching functional information about gene sets.
2. **Lexical analysis** is done to infer relations between terms coming from different sources to **eliminate redundant terms**. The **OntoEnrich framework** [4] has been integrated for associating annotation terms with GO terms according to the following strategy:
   - Decomposing annotations into words
   - Searching groups of consecutive words that correspond to a GO term or its synonyms
   - Removing words included in other ones.
3. Because of the **large size of GO**, we selected only the most **relevant terms** that **synthesize the functional information** of the input gene sets. The most informative parent terms of each GO term found at the previous step are recursively processed until the root term is reached.





3 - Case study

To **demonstrate the efficiency and reproducibility** of the pipeline, the signature profiling of different types of cells has been analyzed using the data from **The immunome compendium of immune cell** subpopulations [5]. They isolated **28 subpopulations of innate and adaptive immune cells**, including normal mucosa and colon cancer cell lines. Each cell type presents different transcriptional profiles that can be considered as gene sets.

- **323 annotations for 24 gene sets** using a hierarchical filter proposed in the tool.
- **98 annotations** would have been obtained for only **16 gene sets**.
- After using the lexical mapping, **264 out of the 323 annotations were kept**.
- **5 out of the 24 gene sets were ignored** by our pipeline.
- Hierarchy simplification **decreased the number of annotations from 264 to 119.**

References

[1] HUANG, DW. et al. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 2008.

[2] THEBAULT, P. et al. Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. Briefings in bioinformatics, 2015.

[3] AYLLÓN-BENITEZ, A. et al. Deciphering gene sets annotations with ontology based visualization. Int. Conf. on Information Visualisation, 2017.

[4] QUESADA-MARTÍNEZ, M. et al. Ontoenrich: A platform for the lexical analysis of ontologies. Int. Conf. on Knowledge Engineering and Knowledge Management, 2014.

[5] BINDEA, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity, 2013.