Deciphering functional annotation of multiple gene sets with MOTVIS.

Aarón Ayllón-Benítez *† [1,2], Fleur Mougin‡ [2], Manuel Quesada-Martínez [3], Jesualdo Tomas Fernández-Breis [4], Romain Bourqui [1], Patricia Thebault§ [1]

[1] Université de Bordeaux – LaBRI – France
[2] Université de Bordeaux – Bordeaux Population Health Center – France
[3] University Miguel Hernandez (UMH) – Spain
[4] University of Murcia – Spain

Introduction

Currently, the advances in omics technologies have opened new opportunities in a large range of biological applications. Such advances may include single-cell, RNA-SEQ or microarray approaches that facilitate expression profiling according to a phenotype or a cell type of interest. As an illustration, these gene profiles are crucial to address the complexity of immune signatures [1]. As these approaches generate a large amount of information, they require bioinformatics pipelines to be understandable by biologists.

In practice, the detection of gene signatures is carried out by applying statistical approaches or clustering. Such methods aim at grouping genes according to their expression levels [2]. Then, deciphering the biological roles of these gene sets becomes a major research challenge to better understand and investigate the biological processes that are involved.

A relevant example is given by the human immunome where each cell has to play a specific role in the immune response. Then, an extensive cell type analysis can be carried out by gene sets that are specifically expressed in each cell type, making use of their gene profiles. For example, a group of genes associated with natural killer cells may be related to the innate immune response, antigen processing, presentation, and cytotoxicity. Thus, annotating gene sets is crucial to: (i) elucidate the biological role of these specific cells and (ii) highlight their specificity. Moreover, making use of these results as a whole can lead to pertinent applications for inferring the role of new type of cells. Furthermore the gene signature of each cell type has to be contextualized with the other types.

The annotation stage consists in associating a gene to a term described in a controlled vocabulary (inferred from experimental or automatic methods) describing functions, pathways, diseases, interactions, etc. This information is stored in various knowledge sources that are continuously evolving.

*Speaker
†Corresponding author: aaron.ayllon-benitez@u-bordeaux.fr
‡Corresponding author: fleur.mougin@u-bordeaux.fr
§Corresponding author: patricia.thebault@labri.fr

Managing the large number of annotation terms associated with a gene set level is usually very difficult. To address this issue, statistical methods, called enrichment methods, have been proposed [2,3]. These tools show an important pitfall related to redundancy in the results [4], resulting from the lack or under-exploitation of semantic relations between terms. In order to solve that, structure knowledge like the ontologies are proposed. The most widely used biological ontology is the Gene Ontology (GO) that provides almost 45 000 terms describing gene roles according to three sub-ontologies: biological processes, molecular functions and cellular components.

Few bioinformatics tools use multiple knowledge sources and aim at decreasing the redundancy and/or quantity of annotation terms by making use of semantic relations between terms [3,4]. However, to the best of our knowledge, no tool addresses these two features combined with a visualization system to analyze together related gene sets. In this context, visualization techniques provide real added-value for the expert when dealing with the additional level of complexity resulting from the multiple sets. So far, such aspects have been partially used to present enrichment results. For example, g:Profiler [5] uses a simple heatmap showing the presence or absence of a term for a given gene in the set. ClueGO [6] provides a node-link visualization between terms sharing the same genes. REVIGO [4] displays results according to three types of visualization: treemap, node-link and space diagram. However, the options avaible are very limited for dealing with multiple gene sets, . Moreover, these tools provide interaction options in the visualization to allow a deep exploration of results. In such context, we recently proposed a prototype of visualization tackling these issues [7], called MOTVIS (MOdular Terms Visualizations).

In this summary, we presents improvements of the MOTVIS pipeline and apply it, to the analysis of signatures of different types of cells. This type of analysis is becoming more interesting and requires new solutions to explore the functional signature of compared expression results since the emergence of single cell sequencing.

Methods

The workflow consists of three main steps to compute gene set annotations plus the dedicated visualization system to examine the results .

First, gene sets are annotated using an enrichment approach. g:Profiler has been chosen for this task because its uses several annotation databases. This permits to combine complementary knowledge for enriching functional information about gene sets (as gene annotations may have been done at different cell organization levels).

The second step involves a lexical analysis to infer relations between terms coming from different sources in order to eliminate redundant terms (same information about the functional roles). To do so, the OntoEnrich framework [8] has been integrated in for associating annotation terms with GO terms by following the strategy:(i) decomposing annotations into words, (ii) searching groups of consecutive words that correspond to a GO term or any of its synonyms, and (iii) removing words included in other ones.

Because of the large size of GO, the third step selects only the most relevant terms that synthesize the functional information of the input gene sets. Then, the most informative parent terms of each GO term found at the previous step are recursively processed until the root term is reached. The selection of the most informative parent term is computed using the information content score proposed in [9]. Once the subgraph of GO is created, the structure is explored to identify the GO terms associated with the gene sets. When a term is associated to the same gene sets as its ancestors, the ancestors are removed.

The last step is to explore these multi-set annotation results, for which a visualization tool has been designed (see Figure 1). The chosen visual structure combines an indented tree (to interactively move across the hierarchy of the ontology) and a circular treemap. The colored visualization of circular treemap represents the different hierarchy between terms and take into account the various scales of biological information that go from general to specific information. The proximity of some circles (representing annotation terms) require to use a visualization technique based on colours. We chose and adapted the three-colors algorithm [10] for automatically assigning gradients of colors to nodes according to their neighborhood distance while preserving a comprehensive cognitive understanding of their relative inclusion. The algorithm uses a color space that is recursively divided into intervals of colors associated with a node and its children. Then, increasing/reducing the luminance/sharpness improves the perception of depth in the tree. This visualization allows to explore the annotation results thanks to interactions as zoom and pan in the circular treemap, or click to expand the branch in the indented tree. Actions performed on the circular treemap impact on the indented tree (and vice-versa). In the circular treemap, the leaf node (white color) represents a gene set, in which a barplot summarizes all the annotations of this gene set (represented as colored circles).

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 [1].The authors 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.

By applying the g:Profiler tool, we obtained 323 annotations for 24 gene sets using a hierarchical filter proposed in the tool. 98 annotation would have been obtained for 16 gene sets if only GO enrichment would have been done This demonstrates the great advantage of using several sources to characterize a larger number of gene sets. After using the lexical mapping, 264 out of the 323 annotations were kept (the 59 remaining annotations were discarded because they could not be mapped to GO). Five out of the 24 gene sets were ignored by our pipeline. Then, the hierarchy simplification stage (third stage) making use of the GO structure has decreased the number of annotations from 264 to 119. This 2.2-fold decrease demonstrates that the enrichment produces a significant quantity of redundant information.

Figure 1. Global view of the visualization tool. At this level, the global information that is displayed allows to define the three ontologies of GO (orange circle for biological process, purple for molecular function and blue for cellular component). The inclusive colored circles correspond to annotation terms that are included in the previous ones. At last, gene sets are represented as white circles.

To illustrate an application of MOTVIS (see a global view of MOTVIS in Figure 1), focusing on the cellular activation and migration, the indented tree can be interactively used to localize

these annotation terms (Figure 2). They fall within "cellular process" and appear there as direct children of this general term (due to the simplification stage). Going into details within the "cell activation" circle, more specific annotations can be depicted. Moreover, if the "activation to lymphocyte" is the focus, zoom facilities are provided to identify the specific type of involved cells, in this case, T cells. At the leaf level (white circle related to T cells), the other annotations related to the type of focused cells can be observed. The pertinence of all the annotations is provided in the white circle thanks to the barplot (Figure 3).

Figure 2. Zoom in on "cell activation" and "cell migration" annotation terms. It shows the gene sets concerned by these annotations. The gradient of colors is correlated to the depth of terms within GO.

Figure 3. Zoom in to represent the leafs or white circles that are related to a type of cells. In this example, the information for the T cells is displayed. For this type of cell, all the annotation terms (corresponding to the gene profile) are represented within a barplot.

Conclusion

In this work, we present and apply the pipeline MOTVIS, dedicated to the annotation of multiple gene sets. Taking advantage of enrichment analysis and the use of several source knowledge, MOTVIS provides computation stages to: (i) perform an original lexical mapping that enables to make use of different knowledge sources, (ii) reduce the annotation redundancy and (iii) filter out the most relevant annotation to synthesize the functional information summarizing multiple gene sets. This new original pipeline has been applied to analyze, compare and visualize the results of a reference compendium of immune cells. According to the transcriptomic profiles of each cell type, MOTVIS offers an interactive way for both identifying the main roles where a type of cell may be involved, and deciphering common features between different cell types. According to the hierarchical relations between GO terms, biology experts can also choose the appropriate level of information (details on demand by interacting with the visualization system) to analyze the results.

BINDEA, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity, 2013, p. 782-795.

HUANG, DW. et al. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 2008, p. 1-13.

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

p. 795-805.

SUPEK, F. et al. REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one, 2011, vol. 6, no 7, p.

REIMAND, J. et al. g:Profiler-a web-based toolset for functional profiling of gene lists from

large-scale experiments. Nucleic acids research, 2007

BINDEA, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 2009, p. 1091-1093.

AYLLÓN-BENITEZ, A. et al. Deciphering gene sets annotations with ontology based visualization. International conference in Information Visualization. 2017.

QUESADA-MARTÍNEZ, M. et al. Ontoenrich: A platform for the lexical analysis of ontologies. In: International Conference on Knowledge Engineering and Knowledge Management. Springer. p. 172-176.

ZHOU, Z. et al. A new model of information content for semantic similarity in WordNet. In: Future Generation Communication and Networking Symposia, 2008. p. 85-89.

TENNEKES, M. et al. Tree colors: color schemes for tree-structured data. IEEE transactions on visualization and computer graphics, 2014, p. 2072-2081.

Keywords: Visualization, Gene Ontology, functional annotation