GOlandscape: A novel and user-friendly GO analysis tool

Abstract

The target of this method is to provide a threshold-free and easy-to-use Gene Ontology (GO) analysis tool for Differential Expression (DE) studies. The method ultimately provides a single heatmap connecting the most differentially expressed genes with the most relevant GO categories associated to the DE study.

1. Introduction

A standard DE study involves a group of treated- and control-condition samples. For each gene, the numeric comparison of its quantitative (normalized) expression among the two groups gives information about the effect of the treatment on the expression level of that gene. The statistical significance of the DE of a gene *g* is commonly measured in terms of a p-value *pDE(g)*, calculated through a *t*-test between the samples of the treated- and the control-condition related to the gene *g*. We will refer to the $-\log_{10}[pDE(g)]$ as to the DE significance *sDE(g)*, in such a way that $pDE(g)=0.001$ (or $pDE(g)=10^{-3}$) will be simply written $sDE(g)=3$. We will often omit the string "*(g)*" for the sake of simplicity.

There are many popular tools to perform a DE analysis, among with limma (edgeR). Usually, these popular DE bioinformatics tools include in the output (for each gene) the fold change (FC), often in terms of its logarithm in base 2, and the significance in terms of a p-value (and/or an adjusted p-value). In this work, we will focus mainly on the statistical significance, rather than on the FC or other statistical parameters, using the FC information only to assess the gene up/down regulation.

Standard GO analysis

The standard GO analysis consists in defining a list of differentially expressed genes (DEList) according to an arbitrary threshold *sDET* on their significance, often being $sDET=2$ (or higher): all the genes whose *sDE* is greater than or equal to *sDET* are then considered differentially expressed. DEList is then clearly a function of the selected *sDET*, DEList=DEList(*sDET*). We will drop the string "(*sDET*)" if not necessary.

Once the threshold *sDET* is chosen (and, thus, a DEList is stated), it is possible to associate the DEList to some annotated categories (GO categories or terms). These categories consist on lists of genes that are already known to be involved in some biological pathway or process. The statistical strength of the association between the DEList and a GO term *G* is again measured by a GO p-value *pGO(G,sDET)* or by the related significance $sGO(G,sDET)=-\log_{10}[pGO(G,sDET)]$, or simply *sGO*. This quantity is usually recovered by means of the hypergeometric test (*k*-test), which involves the size of the gene universe, the size of the annotated category, the size of the DEList and the size of the intersection between the genes present in the DEList and the ones present in the GO term *G*.

An additional, arbitrary threshold *sGOT* on the significance *sGO* of these associations, will select a GOList: all the GO terms whose *sGO* is greater than or equal to *sGOT* are considered statistically relevant in relation with the genes

belonging to the chosen DEList. Eventually, the GOList will give hints about the pathways/processes involved in the effect of the treatment in study, at least in relation with the already assessed knowledge therein.

Limits of the standard GO approach

There are several tools based on this method: Gorilla, goseq, DESeq and David are among the most popular. However, the arbitrary definition of the DEList and the GOList represents a weak point of this standard method. The choice of *sDET*, in fact, has a strong impact on the establishment of the GOList (i.e. on the genes that are passed to the *k*-test to evaluate the *sGO* for each term), leading to an intrinsic instability (especially in case the distribution of the DE significance is far to be normal): A variation of 1 unit of significance may include/exclude key-genes for a specific category which could considered significant/not-significant only on the basis of such choice. This issue cannot be circumvented using the adjusted p-values (at the level of DE), since the false discovery rate correction would anyway only *scale* the relevance of the DE genes, leaving the question of the arbitrary choice on the thresholds untouched (especially at the level of the GOList definition).

There are indeed other methods (e.g. GSEA) which addresses this problem by means of non-parametric tests and a ranking approach. GOlandscape is very similar to these approaches. The main differences are two:

1. The way it is implemented. In fact, GOlandscape still makes use of the *sDE* and *sGO* collected with the standard method described above (*t*-test and *k*-test, respectively).

2. The way it shows the results. The tool is designed to quickly and visually associate relevant genes categories in one look. The key point of this analysis is a heatmap connecting DEList and GOList at different threshold levels *sDET* and *sGOT*.

In the Results section, we will compare the results of the GOlandscape analysis with the ones of other popular GO tools on the XXXXXXXXX case-study, together with the comparison of the GOlandscape results with Goseq and GSEA.

2. Methods

The GOlandscape method is composed of two parts, both integrated in a simple heatmap: The *GO Landscape* and the *Gene Landscape*. The former provides the statistical basis of a threshold-free approach to the GO analysis (a stepwise threshold sampling on the *sDE* span), while the latter (more user-oriented) helps the biologist to evaluate in one look the connection between the most DE genes with the most relevant GO terms obtained with the GO Landscape.

Stepwise threshold sampling

One of the aims of this method is to get rid of the arbitrary definition of the significance thresholds, both for the definition of the DEList and the associated GOList. An easy way to attain this task is to apply a stepwise

threshold sampling (STS) on the whole available range of *sDE* and, at each *sDE* step, to evaluate the *sGO* associated to that *sDE* step for all the available GO terms (see Fig. 1).

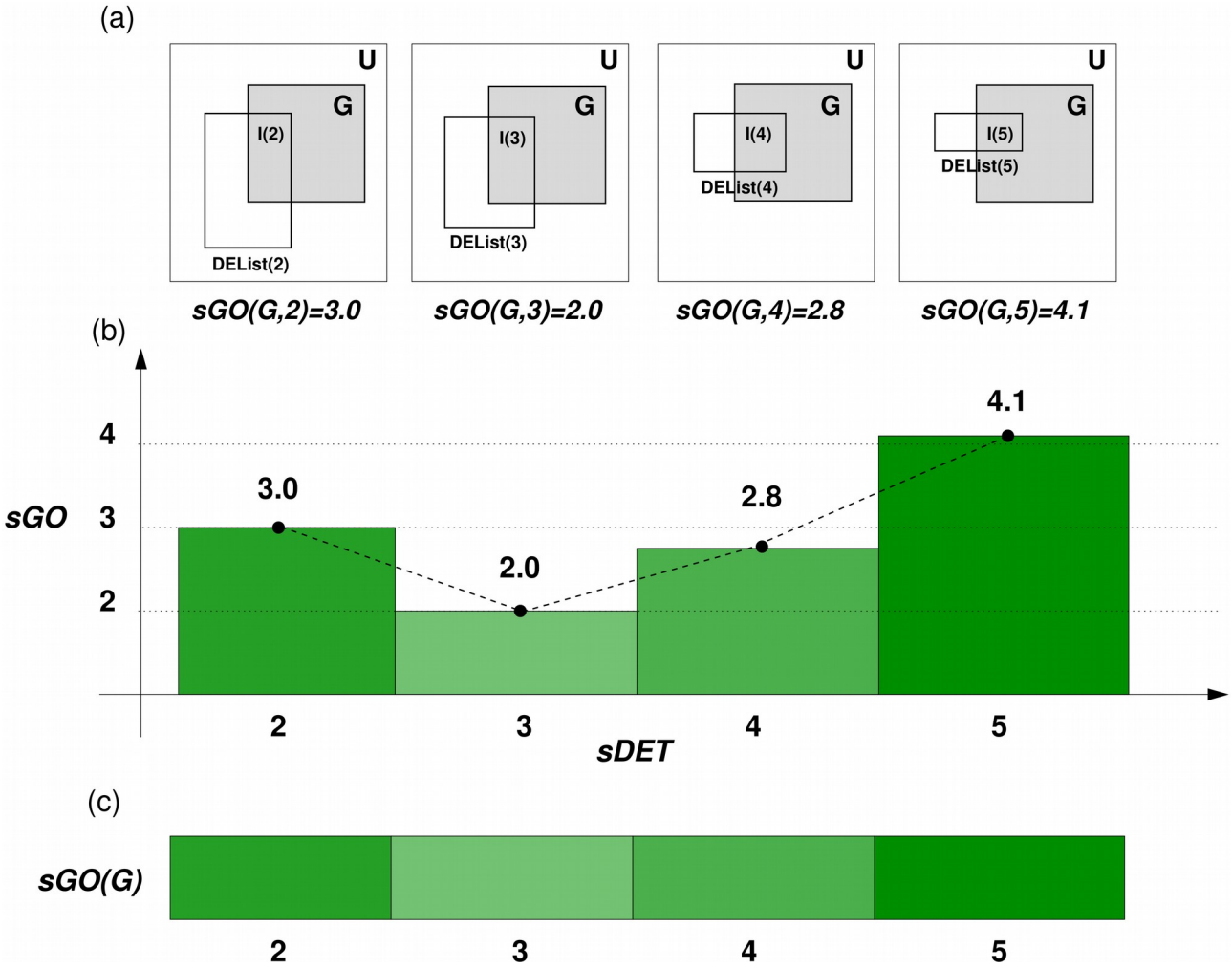

Figure 1. Illustrative example of the STS of the relationship between *sGO* and *sDE* (or, better the *sDE* threshold, *sDET*). We focus the attention on a specific GO term *G.* In panel (a) we show the *k*-test at work: The gray square represents the list of genes annotated for the GO term *G,* while the outer square marked with U represents the gene universe (or background), which are constant at each step of increasing stringency in the DE definition (left to right). The more the threshold on the DE significance *sDET* becomes stringent (here ranging from 2 to 5 with a *sDET* step of 1), the less the related DEList (rectangles) intersects the GO term *G* (the inner rectangles marked with *I*), because it contains less and less genes. Taking into account all the variables involved in the test (the gene universe U, the *G* size, the DEList size and the intersection size) the *k*-test returns a *sGO* value at each step, representing the probability that the DEList(*sDET*) is associated to the *G* gene-set at *sDET.* The panel (b) shows the resulting discrete function *sGO(sDET,G)* (green histogram/black-dashed line). Note that, due to the interplay between these variables and the nature of the *k*-test, the *sGO(sDET,G)* is not necessarily a monotone function of *sDET*. The STS related to the GO term *G* can be eventually synthesized in a compact heatmap-wise fashion, reported in panel (c), where the color-code gives a simplified visual description of the strength of the association between the DEList and the GO term *G* at each step.

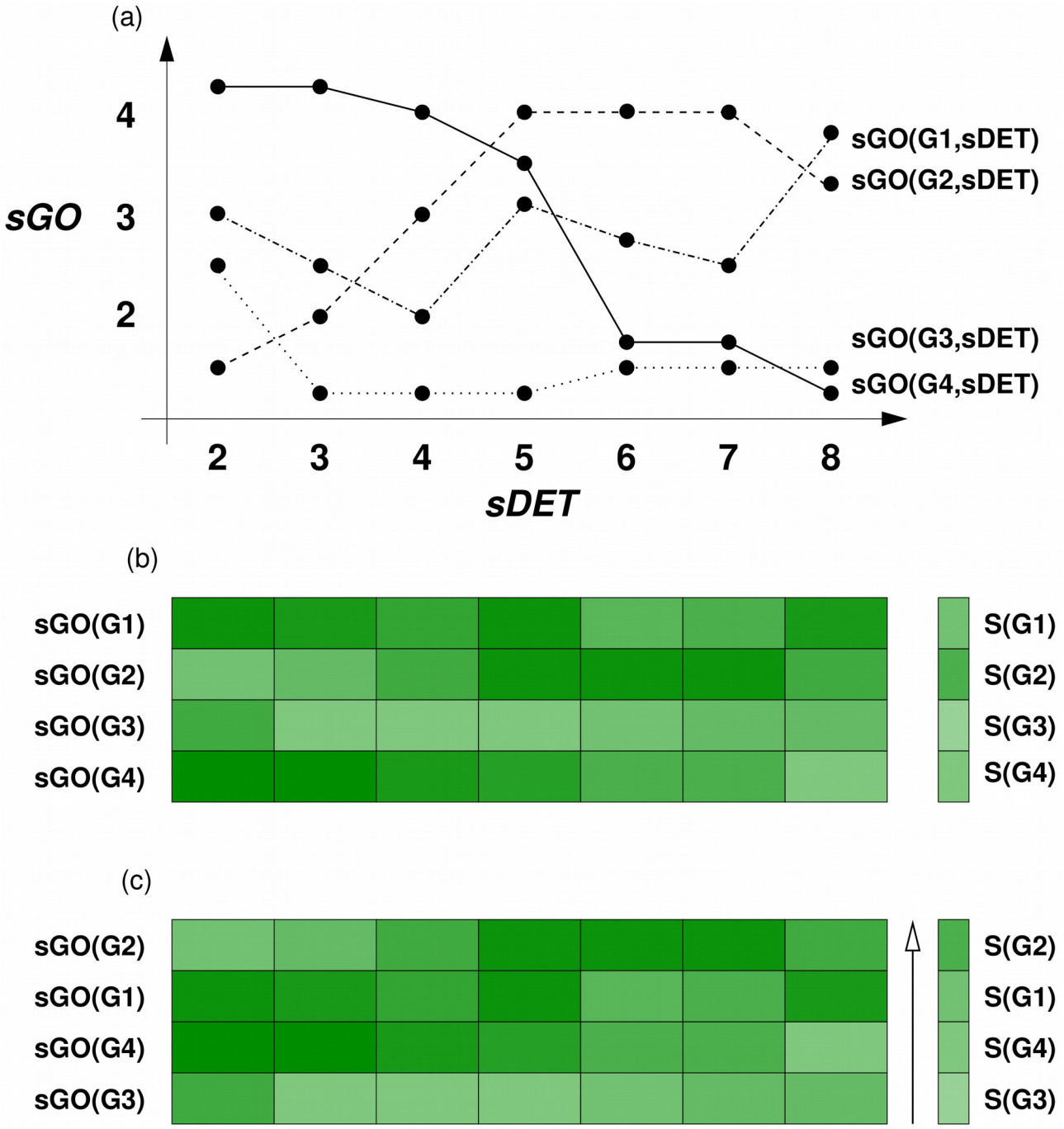

Figure 2. Illustrative survey (up to 4 GO terms) of the *sGO* versus *sDET* stepwise-sampled trend. Panel (a) shows 2 terms, *G4* and *G2,* that are more significantly associated (*sGO,* y-axis) with the DEList at low and high stringency of *sDE*, respectively, while the *G1* and *G3* terms show intermediate and low values, respectively, all along the *sDE* range (x-axis). Panel (b): All these functions can be summarized in a compact form with a heatmap, as already shown in Fig. 1, with an additional column (right) representing the global significance *S* for each GO term. In panel (c) we eventually reorder the rows of the heatmap according to the global significance *S*. Note that although the *sGO(G4,sDET)* reaches the highest *sGO* values, it ranks only at the thirds position in this qualitative sketch: this is due to the fact that the high *sGO* values are associated to low values at the corresponding *sDET*. The heatmap in panel (c) will be referred as to the *GO Landscape* related to the DE study.

Overall relevance of GO terms in a STS scenario: the GO Landscape

Since we are interested in targeting the *most relevant* GO terms associated to our study, we have to evaluate the *sGO(sDET,G)* for *all the GO terms G* in order to compare them to one another. The result is a collection of discrete functions, one for each GO term, whose x-axis represent the increasing degree of stringency in the choice of the DEList definition and whose y-axis represent the trend of the GO significance as a function of *sDET*.

Fig. 2 illustrates a naive sketch of the STS scenario for some GO terms. The analysis of such scenario suggests an overall criterion to assess which GO term, among the analyzed ones, should be considered as the most relevant. As it appears from Fig. 2(a)-2(b), the STS trend can be very diverse for different GO terms. The idea we propose here is to weight, all over the *sDE* range, the *sGO(sDET,G)* with the related *sDET*, in such a way to avoid overestimation of high *sGO* values associated to low *sDE* and vice versa. The global significance *S(G)* of a GO term *G* will be thus simply defined as follows:

$$S(G)=\sum sGO(G,sDET)\times sDET,$$

where the sum is intended over the *sDET* thresholds (or steps) all over the range of variation of *sDE*. The discrete *sGO(G,sDET)* functions can then be ranked according to their *S(G)* value, as qualitatively depicted in Fig. 2(c). The *S(G)* function can be easily transformed in a proper (average) significance dividing it by the sum of all the *sDE* steps times the number of steps, with no influence on the ranking results. This heatmap will be referred as to the *GO Landscape*. In the Results section a strictly quantitative example of a GO Landscape will be introduced.

Clearly, the GO Landscape can be built on the DEList containing only up regulated genes, only down regulated genes or pooling all the DE genes together, independently on their fold change. Figure 3 contains an example of a final results with GOlandscape.

Figure 3 Example (constant stepwise sampling)

Sampling refinement

The question how to properly sample the *sDE* spectrum is here addressed. In the previous paragraph we mentioned a constant-step sampling (STS), whose step is arbitrary decided (in Fig. 1 we used a step of 1 *sDE* unit). This choice is not connected to the characteristics of the p-value distribution, which can vary considerably for each data-set.

However, the sampling can be improved on the basis of a simple observation: the number of reliable *significant digits* of a p-value cannot, in general, exceed 2 [citation]. This allows us to *round* the significance *sDE* to the second decimal, thanks to a property of the logarithm. In this condition, it is possible to create a summary histogram of the *sDE* distribution, attributing at each (rounded) *sDET* available in the data-set the number of genes *NDEg(sDET)* carrying such significance (while, considering all the full digits of the p-value, it would very likely reflect in a single gene for each p-value). The product *sDET*NDEg(sDET)* represents the global amount of differential expression at *sDET* across the data-set. The list of the different *sDET* in the

summary histogram represent the most precise (and, at the same time, statistically robust) sampling we can apply to the *sDE* spectrum.

Once this histogram is produced, the sampling of the distribution can be easily obtained with the request of adding (approximately) a constant amount of differential expression across the genes, to each step (rather than a rough constant step on the *sDE* alone). See Figure 4 for details.

Figure 4 with the stepwise comparisons

The number of different *sDET* (or, in other words, the number of steps) in the summary histogram is usually greater than the one obtained with a constant thresholding. This number (that is actually the only free parameter left for the sampling) can be arbitrary decreased, accordingly to the available power of computation: a lower number of step will reflect in a higher amount of significance added from one step to the following. In any case, the resulting sampling (list of *sDET* steps) will be strictly connected to the shape of the distribution. In short, instead of arbitrary fixing the step-size (e.g. 1 unit of *sDE* as in Fig. 1) irrespectively on the p-value distribution, we can arbitrary fix the number of steps and modify the step-size, according to the p-value distribution, in such a way that, at each step, the same amount of DE significance in the data-set is added. We have to keep in mind that each DE step, no matter how we do establish it, requires the calculation of the GO p-values for each GO category. Thus, a balance between the refinement and the computational time consumed should be found.

A more refined way to sample the distribution, based on the same observation on the significant digits of a p-value, would be to select the number of steps in such a way that the *sGO* of each category varies, among two contiguous steps, above the second decimal: but this sampling would be obviously different from GO term to GO term and it would be frankly computationally unfeasible.

Association between DE genes and relevant GO terms: the Gene Landscape

The GO landscape gives at the same time an overview of the most relevant categories and the statistics needed to support the ranking among them. However, this information is not of great use for the biologist, who is usually more interested in which genes are most relevant in her/his DE study, possibly according to their up/down regulation with respect to the control condition. To this end, but on the statistical basis of the GO Landscape described before, we provide also another heatmap, called the *Gene Landscape*, which will cross the genes contained in the relevant categories revealed by the GO Landscape with the genes present in the DEList. A quantitative example of this Gene Landscape will be introduced in the Result section.

3. Results

Based on

Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease (Nature. 2015 Feb 19; 518(7539): 365–369).

4. Discussion

With this approach, the arbitrary choice of the *sDE* threshold used in the standard method to assess a GOList is translated in the number of GO terms to be listed in the GO Landscape. The clear advantage of this transformation is that the problem of instability mentioned in the Introduction has disappeared, since if a GO term is relevant at a certain threshold, it will stick out from the heatmap, allowing the scientist to investigate about its role in the context of the study. Actually, the standard method is *contained* in one column of the GO Landscape. But this approach allows also to evaluate the impact of the choice of a certain threshold in the framework of many other possible choices, taking into account that the definition of *S(G)* (the ranking function) tends to give more relevance to the GO term with highest (*sDE*, *sGO*) values, i.e. with higher statistical relevance.

The disadvantage of this method is that the user has to deal with an entire *panorama* of GO terms which is on the one hand strictly based on statistics but, on the other hand, strongly dependent on the GO database used (e.g. consider the different level of annotation for BP and KEGG terms). About the first problem, the *sGO* values evaluated with the *k*-test tend to favor in particular the GO term with small intersections and, thus, the DEList with fewer genes. However, Fig. X shows that the ranking of the *G* terms (i.e. the *S(G)* value) is not dependent on the size (number of genes) contained in the *G* terms, which is a critical issue in the GSEA approach. The second issue is not circumventable by mathematical tools and it is a shared problem with the standard method. Thus, our opinion is that an overview of all the available associations could be of great help to the biologist whose task is to find the most relevant pathways related to her/his study, avoiding to omit only on the basis of an arbitrary choice (that could be too stringent or to loose) some relevant process.

The other critical point of this approach, is the definition of the STS itself. In case the distribution of the *sDE* is considerably different than the normal distribution, a constant stepwise sampling is not suitable for a correct evaluation of all the possible DEList. On the other hand, a further refined technique to sample the DEList (e.g. adding one single DE gene at each step) may not necessarily be beneficial. In fact, as mentioned in the Methods, the p-values resulting from the *k*-test would differ only to the extent of few decimals, far lower than the two significant digits which, as a general rule, is the limit for a meaningful and reliable p-value. The refined sampling described takes into account both these aspects. However, during the test of this work, various other sampling techniques were explored to sample the DEList range but with few changes to the global significance (and, thus, to the ranking) of a category (data not shown).

Remarks

In case the maximum p-value of a DE study is low (say, sDE=2) the method cannot expected to be reliable.

In case the labeling of the genes is not unique, the method cannot work, so one should be careful on the input gene id.

Different GO types (namely BP and KEGG) differ for internal number of annotation which reflects in intrinsic absolute difference in the *sGO*. This means that these different GO types cannot be directly compared. The choice is again left to the user: to select the first *N* GO term for each GO term types.

Correction to the k-test (Daniel)

Introduction of the onion problem (Mathieu)

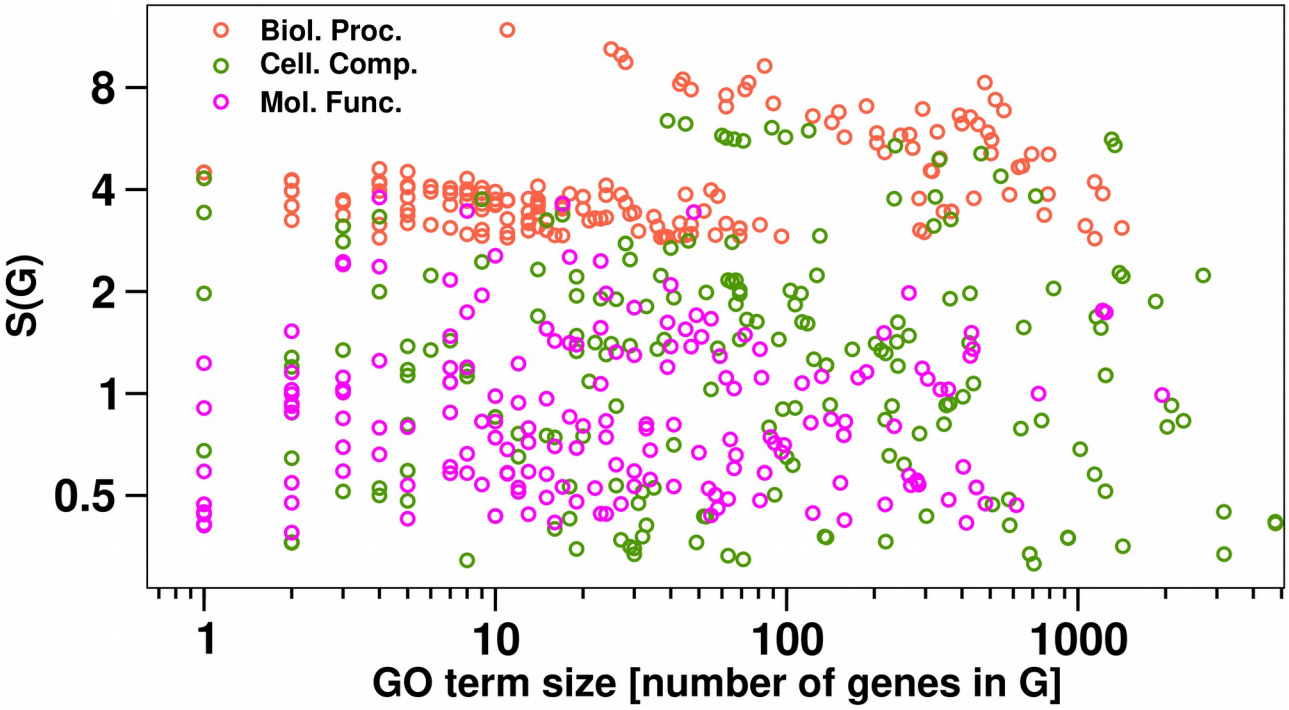

Figure X. Numerical relationship between GO term size *G* and the relevance of the association *S(G)* for three subset of gene annotation (Biological Process, Cellular Component and Molecular Function terms). As it appears from these 600 terms (200 terms for each subcategory, coming from four different DE studies), there is no correlation between the number of genes contained in a GO category and the associated relevance (or rank) in the GO Landscape heatmap (Fig. 2, panel (c), right-most column).