Traditionally, gene-set enrichment analysis methods, including Fisher’s exact test, Kolmogorov-Smirnov test coupled with permutation, which is known as GSEA [reference], and the CAMERA method [reference], all operate independently on each gene set, implicitly assuming independency between gene sets. While the assumption may hold true when the gene sets are derived from single data source and the redundancy of genes between gene sets are low. However, when applied to gene-sets derived from multiple resources, such analyses often identify multiple significantly enriched gene-sets with very similar compositions (see Supplementary Figure X for an example). Though post-processing algorithms are available to cluster significantly enriched gene-sets by their composition, for instance the DAVID clustering tool [references], it is desired to explicitly model the dependencies between gene sets within a rigorous statistical test procedure. This is especially important in the context of annotating modules identified from network analysis, because one wishes to describe the modules by annotating them with few (sparse) yet informative gene sets.

To this end, we propose to formulate the gene-set enrichment problem within a regression framework: given *M*, a module of genes, and *B*, the background, or the union of genes from various database such as GO and Reactome, as well as ***S***, a list of gene sets, we take each gene set *s* as a feature of the any gene *g* in B, which in the simplest form can take the value of either 0 (*g* does not belong to *s*)or 1 (*g* belongs to *s)*. Thus, the problem of gene-set enrichment is transformed to the problem of predicting/explaining the membership of *M* given the background *B* using the features ***S****,* which can be solved for instance by logistic regression.

To account for the fact that gene sets can have overlapping genes, we propose to apply regularised regression methods, such as *lasso* (*l1* regularization), *ridge (l2 regularization),* or *elastic net* (hybrid of *l1* and *l2* regularization controlled by the hyperparameter *alpha)*, so to adjust the treatment of similar or redundant gene-sets. If two gene-sets are highly redundant, *lasso* will assign a higher coefficient to one of them randomly, *ridge* will assign equal coefficients to both of them, whereas *elastic net* will behave between *lasso* and *ridge*. In our analysis, we applied the elastic-net variant with alpha=0.5 [reference].

The last detail of the algorithm is that in order to make the models interpretable, we set the constraints that the coefficients must be non-negative(?????), namely we only consider over-representation of gene-sets and ignore cases of under-representation.

We constructed gene sets from the latest version of Gene Ontology (format-version: 1.2

data-version: releases/2018-07-15) and the Reactome database (download time: 2018-07-16). Genes belonging to a GO term or a Reactome pathway are considered as one gene set, independent of positions of either the term or the pathway in the respective hierarchies. Next, we used the gene sets to construct a gene-by-gene-set matrix binary *G,* whose rows are genes (B) and columns are gene sets (S). Gij equals 1 if and only if gene *i* belongs to gene set *j*; otherwise Gij = 0.

Given a module *M* derived from the network analysis described previously, as well as the background gene list *B*, which contains the union of genes identified in GO and Reatome, we construct a vector *y* representing all genes in *B*. We assign yi = 1 if and only if when gene *gi* belongs to the module *M*. Next, we train the regression model: y=GTβ+ε using elastic net. Gene sets with coefficients significantly larger than zero are selected and reported (alpha=0.5). By adjusting the alpha parameter of the elastic-net method, users can control the number of selected pathways.

From my side, I think in this method both positive and negative coefficients can represent the importance of certain gene sets. Actually, I also tried to use absolute coefficient value before. But interestingly, all selected coefficients having absolute value large than 0 are also positive. My explanation for this is that we label y here as 1 and 0 instead of 1 and -1, so negative coefficients will not be needed in regression.   
  
As we only consider ownership relationship between genes and gene sets and ignore the expression level of genes. So I think coefficients whose values are large than 0 only mean the corresponding gene sets are important and have a larger amount of genes belong to the annotated modules. And importance here actually doesnt mean "enriched" or "depleted ". It only means its gene set has "certain amount of overlap genes " with annotated module. But compared with simply using overlapped genes as annotation criteria, regression based method can get more sparse results