# Gene-set enrichment with regularized regression

Tao Fang[1,2], Iakov Davydov[1], Daniel Marbach[1], Jitao David Zhang[1]
[1] Roche Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland
[2] Current Address: European Bioinformatics Institute, Wellcome Genome Campus Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

## Abstract

**Motivation**: Canonical methods for gene-set enrichment analysis assume independence between gene-sets. While the assumption may be reasonable when the redundancy is low, its validity breaks down when gene-sets are overlapping or even redundant with each other. In practice, heterogeneous gene-sets from different sources are often used, leading to hit gene-sets that are partially or fully overlapping, which compromises statistical modelling and complicates results interpretation.

**Results**: We rephrase gene-set enrichment as a regression problem by treating genes-of-interest membership as a binary target variable, and gene-set membership as binary dependent variables. The goal is to identify a minimum set of gene-sets that best predict whether or not a gene belongs to a set of genes of interest. To accommodate redundancy between gene-sets, we propose to solve the problem with regularized regression techniques such as the *elastic net*. We found that regression-based results are consistent with established methods, but much more sparse and therefore interpretable.

**Availability**: We implement the model in an R package, *gerr* (gene-set enrichment with regularized regression), which is freely available at https://github.com/TaoDFang/gerr and has been submitted to *Bioconductor*. The scripts and the data used in this paper are available at https://github.com/TaoDFang/GeneModuleAnnotationPaper.

**Contact**: Jitao David Zhang (jitao_david.zhang@roche.com), Roche Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd. Grenzacherstrasse 124, 4070 Basel, Switzerland.

## Introduction

A plethora of gene-set analysis methods have been proposed. Popular choices include Fisher's exact test, GSEA (A. Subramanian *et al.*, 2005), CAMERA (Wu and Smyth, 2012), while many other tools using various statistical models and procedures are available (Rahmatallah *et al.*, 2014; de Leeuw *et al.*, 2016; Rahmatallah *et al.*, 2016). Methodologically, they can be classified into self-containing and competitive methods

(Goeman and Bühlmann, 2007). Practically, they are applied in contexts where the user wishes to gain biological insight from a set of genes of interest (GOI hereafter).

We observe that most proposed methods are based on two implicit assumptions. First, there is only one set of GOI to be tested for enrichment of gene-sets at a time. If there is more than one set, each set is tested independently and the results of multiple sets are simply merged. Second, most methods operate independently on each gene-set and thereby implicitly assuming independence between gene-sets. There are good reasons for both assumptions. First, it is not rare that users are only interested in one set of GOI and in finding out whether a few gene-sets are over- or under-presented in the set. Second, treating both sets of GOI and gene-sets independently simplifies the software implementation and allows enrichment analysis even of one single gene-set for one set of GOI. Computational techniques such as parallelization (Sergushichev, 2016) and approximation (Zhang *et al.*, 2017) can be easily implemented to speed up execution. Finally, when the gene-sets are derived from a single data source, the redundancy between the gene-sets, i.e. proportion of shared genes between two gene-sets, can be small, and therefore the gene-sets may indeed be treated empirically as independent. Under the circumstances when gene sets are organized in a hierarchical structure, such as Gene Ontology (GO), it is possible to reduce results redundancy of the gene sets using graph decorrelation (Alexa *et al.*, 2006; Grossmann *et al.*, 2007). More often than not, however, gene-sets are not hierarchically organized and the independence is implicitly assumed.

In practice, however, the two assumptions, particularly the independence between gene-sets, are often violated. Gene-sets from different sources are commonly aggregated for enrichment analysis, therefore redundant gene-sets can be called as hits. For instance, DAVID (Huang *et al.*, 2009), a popular web tool for gene-set enrichment analysis, aggregates by default gene-sets that reflect disease association (e.g., OMIM, Hamosh *et al.*, 2005), functional category (e.g., UniProt keywords, The UniProt Consortium, 2019), Gene Ontology (Ashburner *et al.*, 2000, The Gene Ontology Consortium, 2018), pathways (e.g., KEGG pathway, Kanehisa *et al.*, 2016) and protein domains (e.g. Interpro, Mitchell *et al.*, 2019). Additionally, users can select other gene-sets that will be appended. Gene-set enrichment analysis is then performed on one set of GOI uploaded by the user, using a modified version of Fisher's exact test running on each gene-set independently. Such an aggregation-and-test-independently strategy is not only used by the DAVID tool, but is followed by many published studies using different sets of GOI, gene-sets, and methodologies.

While some gene-sets are hardly overlapping with each other, others do share a significant proportion of genes as common members. For instance, genes associated with the keyword *chemotaxis* in UniProt are highly redundant with genes associated with the biological-process term *cell chemotaxis* in GO. If a set of GOI is indeed enriched of chemotaxis-relevant genes, both gene-sets will be reported as hits. It is apparent that when many enriched gene-sets are partially or fully overlapping, not only the independence assumption underlying statistical modelling is compromised, the interpretation will become complicated because a common set of genes may underlie many hit gene-sets with different names.

One way to assist human interpretation is to cluster gene-sets with similar compositions *post hoc*. For instance, the DAVID Gene Functional Classification Tool is based on the kappa statistics, a similarity measure of gene-sets based on gene composition, and a fuzzy heuristic multiple-linkage partition algorithm to cluster gene-sets that are similar with each other into so-called *annotation clusters* (Huang *et al.*, 2007). The results are often very useful because they organize gene-sets reflecting identical or relevant biological aspects together, and the resolution of the clustering can be modulated by user-defined parameters. Nevertheless, users still have to examine gene-sets within each annotation cluster to derive a high-level understanding of pathways and gene-sets that are enriched. Apparently, this does not scale when many annotation clusters are identified or when many sets of GOI are to be tested simultaneously, which is sometimes the case in network biology and especially in the area of community detection, where network modules are identified and the functions of these modules need to be elucidated. For instance, Choobdar *et al.* (2019) recently reported a community challenge that assesses network module identification methods across complex diseases, where more than three hundred consensus modules are reported. It will be slow, error-prone, and most likely not reproducible to manually curate gene-set enrichment analysis results to remove redundancy in such cases.

Another issue often met in practice is that one may have more than one set of GOI, for instance, multiple gene modules, and wishes to identify gene-sets that are either enriched in each set of GOI irrespective of their enrichment in other sets, which we name *simple enrichment* to contrast against *characteristic enrichment*, where uniquely enriched gene-set(s) when compared with other sets of GOI are of interest. A simple approach would be to run gene-set enrichment analysis using the same set of gene-sets individually and independently on each set of GOI. Consequently, in the case of characteristic enrichment, significantly enriched gene-sets can be filtered *post hoc* to identify gene-sets that distinguish each set of GOI from others. Note that such an approach will require two levels of *post hoc* operations: firstly treating gene-sets as independent from each other, and secondly treating sets of GOI as independent from each other.

Instead of applying *post hoc* methods to enrichment problems with multiple sets of GOI and many gene-sets with potential redundancy, we asked whether it is possible to explicitly model the overlapping nature of gene-sets and the two questions of interest - simple enrichment and characteristic enrichment - using a single unified statistical framework. This led us to transpose the task of gene-set enrichment as a regression problem. The key insight is that we can treat gene-sets as dichotomous feature vectors of genes with two possible values: one if a gene is in the gene-set, and zero otherwise.

From this perspective, the problem of simple enrichment, namely enrichment of gene-sets in a set of GOI among all possible genes of consideration (commonly known as *background* or *universe*), can be seen as a regression problem with many partially correlated features (gene-sets in this case) as predictors, and the membership of GOI as a dichotomous response variable, which takes the value of one if a gene belongs to the GOI, and zero otherwise. In this setting, the problem of characteristic enrichment can be addressed by multinomial regression, a natural extension of linear and logistic regression.

Given that we are interested in a minimum set of gene-sets that describe the GOI, many regularized regression algorithms can be used, including the Lasso method (Tibshirani, 1996), Ridge regression (Hoerl and Kennard, 1970), or linear support vector machines (Hastie *et al.*, 2004). In the current study, we demonstrate the principle and feasibility using the elastic net (Zou and Hastie, 2005), an established statistical procedure that combines regularization and feature selection. We show that regularized regression is able to derive biologically meaningful and succinct lists of gene-sets that reflect biological functions enriched in GOI, without *post hoc* human interference.

## Algorithm and Implementation

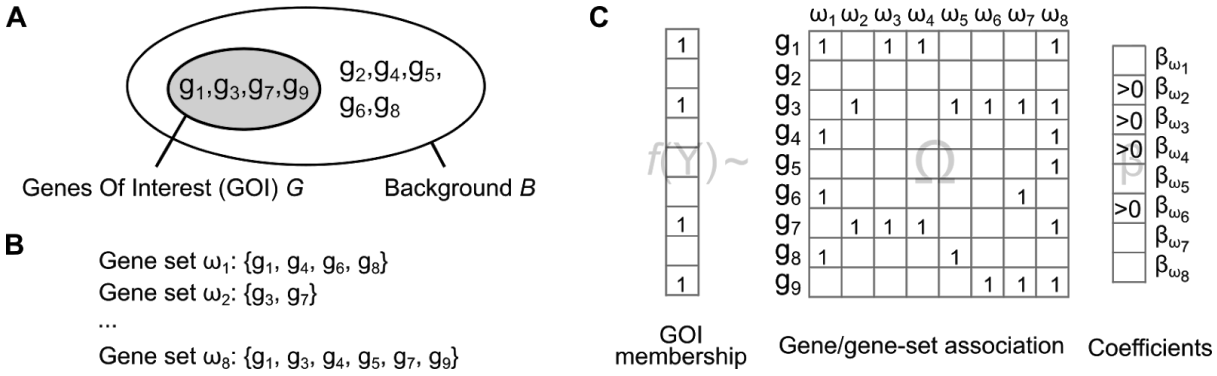


Figure 1: Schematic representation of the algorithm, with a toy example with four genes of interest in a background of nine genes and eight gene-sets. Generally, the user specifies the set of GOI, the set of background genes (panel A), and a list of gene-sets that are potentially redundant with each other (panel B). Gene-enrichment with regularized regression formulates the task of gene-set enrichment analysis as a regression problem with GOI membership as a dichotomous target variable and gene/gene-set association as the independent variables, using a link function *f* that can be specified by the user (panel C). Feature selection is achieved by identifying gene-sets with coefficients that are non-zero, more specifically that are positive for over-representation. In this example, gene-sets $\omega_2$, $\omega_3$, $\omega_4$, $\omega_6$ are selected. Empty cells in column vectors and the matrix indicate zero.

A schematic representation of the algorithm can be found in Figure 1. We use $G$ to denote a set of GOI and $B$ to denote the set of all genes of consideration, i.e, *background* or *universe*. $B$ can be for instance all genes encoded in a genome or a subset of it that is measured. We assume that $G \subseteq B$, namely any gene in $G$ must also in $B$. We use $Y$ to denote a binary vector indexed by genes in $B$; we assign $y_i = 1$ if and only if when $g_i \in G$ (gene $i$ belongs to the set of GOI).

We use $\Omega$ to denote the set of gene-sets that we use to annotate $G$, with $n$ gene-sets $\omega_1, \omega_2, ..., \omega_n$ as elements. Each gene-set $\omega_i\ (i = 1, 2, ..., n)$ can be expressed as a binary vector indexed by genes in $B$, with a value of $1$ if the gene is a member of $\omega_i$ and $0$

otherwise. Therefore, equivalently, we can use $\Omega$ to denote a binary matrix, with genes in rows and gene-sets in columns, that associates gene-sets with genes in $B$ (therefore also genes in $G$).

Following these notations, the statistical model of generalized linear regression (Agresti, 2015) has the form of

$$f(Y) = \Omega\beta + \varepsilon,$$

where function $f$ denotes the inverse of the link function, which in the case of linear regression is the identity function, and in the case of logistic regression is the logit function that maps from the domain of $[0, 1]$ to the real-number domain $R$. $Y$ is the column vector of gene membership in $G$, $\beta$ denotes the column vector of coefficients of gene-sets, and $\varepsilon$ denotes the error term that is assumed to be independently and identically distributed following the normal distribution. A similar definition of the problem was used by Mi *et al.* (2012). The important distinction, however, is that we use regularization to mitigate non-independence between gene set, while in the former approach the logistic regression was used to rank gene sets, not to select a relevant and non-redundant set.

With regard to the choice of appropriate link function, previous studies (Hellevik, 2009) and our observations (supplementary document 1) showed that the results of linear models are both stable and meaningful, therefore the linear model is constructed by default in the *gerr* package, though logistic models are equally supported.

To accommodate redundancy between gene-sets, we propose to use the regularization technique. Typical choices include (1) Lasso, a type of *L1* regularization, which, loosely speaking, shrinks coefficients of less important features to zero, (2) Ridge, a type of *L2* regularization, which, loosely speaking, halfs the coefficient if two variables are identical, and (3) the elastic net, a hybrid of *L1* and *L2* regularization, controlled by the hyperparameter $\alpha$. If two gene-sets are highly redundant, *Lasso* will assign a higher coefficient randomly to one of them, while keeping coefficients with low importance equal to zero; *Ridge* will assign equal coefficients to both of them, while non-zero estimates even for features with low impact on the model. *Elastic net* combines advantages of both approaches: it will estimate non-zero values for coefficients of correlated features while setting coefficients for features of low importance exactly to zero. In our analysis, we applied the elastic net variant with $\alpha$ =0.5, using the implementation in the R *glmnet* package (Friedman *et al.*, 2010). By adjusting the parameter, users can control the sparsity of the results, i.e. numbers of selected pathways.

We are interested in the coefficient vector $\beta$, particularly the gene-sets with large positive values. They correspond to gene-sets that associated stronger with $G$ compared with other gene-sets. We set the constraint that the coefficients must be non-negative, namely, we only consider over-representation of gene-sets, where genes in a gene-set are more frequently present in GOI than random as specified by a null model, and ignore cases of under-representation.

Practically, the gene-sets can be derived from any data sources. For the purpose of demonstration, we use a union set derived from GO and the Reactome pathway database (Fabregat *et al.*, 2018).

The regression model, the example gene-set collection for demonstration, as well as a number of helper functions are implemented in the R package *gerr* (gene-set enrichment with regularized regression) that is published under the Artistic-2.0 open-source license. *gerr* allows users to extract enriched gene-sets from the elastic net regression analysis, and to control the sparsity of results by adjusting the parameters. For gene-sets identified by the regression analysis, *gerr* also returns the enrichment analysis result using Fisher's exact test for comparison. In case that selected gene-sets are originated from GO and/or the Reactome database, which implement tree data structures, *gerr* returns the distance from the selected node to the root nodes as well as the subtree structure to help users understand the biological context of the selected gene-sets.

# Results

## Model verification and performance evaluation

We verified the model of *gerr* and evaluated its performance with simulation studies, which are described in full detail with reproducible codes and data in the vignette of the package (supplementary document 2). Here we highlight the key concepts and results.

We used 500 randomly selected curated gene-sets from MSigDB (A. Subramanian *et al.*, 2005) for simulations. The gene-sets are of varying sizes, containing tens up to more than a thousand genes, and a subset of them share common genes.

To verify the model, we select one gene-set, artificially assign its member genes as GOI, and use the *gerr* package to identify enriched gene-sets. The procedure is run once for each gene-set. An ideal model will return one and only one gene-set, namely the input gene-set, as the positive hit. For 88% of all the cases, *gerr* managed to do so. Otherwise, *gerr* returned true-positive hits without exception and no more than three false-positive hits among 500 tests (false positive rate <= 3/500, or 0.006).

As a comparison, we performed analysis with the same set of gene-sets using one-sided Fisher's exact test, testing for over-presentation, and FDR-correction with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). According to previous reviews and analysis (Khatri and Drăghici, 2005; Hackenberg and Matthiesen, 2008), FDR is probably the best choice if gene-sets are likely to be related. By setting cut-off at FDR<0.05, this procedure (FET+FDR hereafter) returned many more false-positive hits in most cases (Figure 2A, median false positive rate 0.014, with a median absolute deviation of 0.021).

Furthermore, we note that the false-positive hits of *gerr* are mainly caused by gene-sets with high level overlapping with the true positive hit. The average overlap coefficient, defined by

$|A \cap B|/min(|A|, |B|)$ for any two sets $A$ and $B$, between false positive hits and the true positive hit is much higher than the case in FET+FDR or than the expected values if gene-sets are drawn randomly (Figure 2B). Therefore, the verification step, despite of its simplicity, suggests that the elastic net model underlying *gerr* may have good sensitivity and specificity for the task of gene-set enrichment.

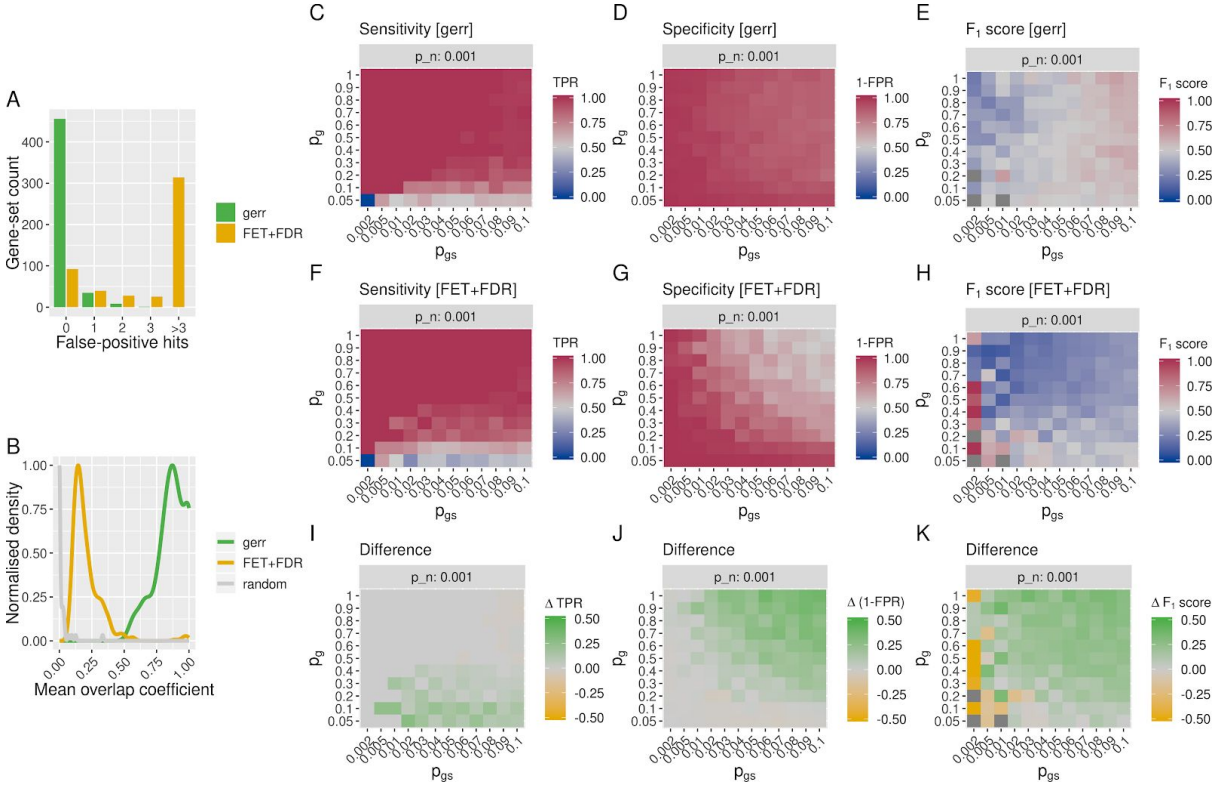


Figure 2: Model verification and performance evaluation with simulation studies. (A) In model verification, both *gerr* and FET+FDR had 100% sensitivity. However, FET+FDR often returned false-positive hits (median: 7) whereas *gerr* returned no more than 3 false positive hits. (B) Mean overlap coefficient between false positive hits returned by *gerr* and the true positive hit is much higher in case of *gerr* than FET+FDR or otherwise randomly expected. (C-E) Sensitivity (true positive rate, or TPR), specificity (1-false positive rate, or 1-FPR), and $F_1$ score (harmonic mean of precision and sensitivity) of *gerr*, by varying parameters. Six values of $p_n$ were tested. Only one is shown here, though the patterns apply to other values (see the full visualizations in supplementary document 2). (F-H) Like (C-E), but results of FET+FDR. (I-K) the difference of performance between *gerr* and FET+FDR.

Next, we evaluated the specificity, sensitivity, and precision of the *gerr* model in a probabilistic framework. We assumed a generative model of GOI, where the probability of a gene $g$ belonging to a set of GOI $G$ is modelled by $p_{g \in G} = \sum_{\Omega} p_{g|\omega} p_{\omega} + p_{g_n}$. The model specifies an additive model of the probability that is modelled by the probability of the gene-set $\omega$ contributes to GOI, expressed as $p_{\omega}$, multiplied by the probability that $g$ is selected to contribute to GOI given that $\omega$ contributes to GOI, expressed as $p_{g|\omega}$, summed over all genet-sets, and then adding the gene-specific term $p_{g_n}$ that models the probability

that the gene $g$ contributes to GOI independent of its associations with gene-sets. The two parts on the right side of the equation can be observed as a gene-set dependent and a noise term, respectively. The total probability is defined in the range $[0, 1]$.

For simplicity, we make the assumption that both $p_\omega$ and $p_{g|\omega}$ follow binomial distributions, and a single noise level $p_n$ applies to all genes. Under these assumptions, we varied the values of $p_\omega$, $p_{g|\omega}$, and $p_n$ in a set of hopefully plausible ranges and assessed sensitivity (false positive rate), specificity (1-false negative rate), and $F_1$ score of the *gerr* model. As a reference, performances of the FET+FDR procedure and the difference between the two are reported using the same parameter settings (Figure 2C-K). The results suggest that within most of the tested parameter space, *gerr* is robust against the noise item of the generative model of GOI, and shows higher sensitivity, specificity, and precision than the FET+FDR procedure.

In short, model verification and performance evaluation with simulation studies suggest that *gerr* may work reasonably well if a set of GOI is constructed either by genes of a single gene-set or by the proposed additive model with noise term. In reality, GOI can be constructed in many different ways and the additive model may not always hold. Nevertheless, the simulation studies suggest that the performance of *gerr* can rival or even exceed the performance of well-established FET+FDR procedure.

## Case-study with consensus modules identified by a community effort in the DREAM challenge

Given the good performance of our approach in simulations, we applied the gene-set enrichment with regularized regression model to a real-world data set in order to evaluate its performance. To this end, we leveraged the consensus modules identified by a community effort in the *Disease Module Identification DREAM Challenge* (Choobdar *et al.*, 2019). In this project, researchers from all over the world used diverse methods to detect disease-related gene modules from diverse human molecular networks such as STRING (Szklarczyk *et al.*, 2015) and InWeb (Li *et al.*, 2017). As part of the outcome, the crowd-sourcing approach generated 377 consensus modules from the STRING molecular interaction network.

We first applied gene-set enrichment analysis with the FET+FDR procedure to each consensus module as a set of GOI, using the union of Gene Ontology and Reactome gene-sets. Next, we applied gene-set enrichment analysis with regularized regression using *gerr* and compared the results of both procedures.

Across modules, we found that regularized regression returned far more sparse results (median=9.0, interquartile range/IQR=8.0) compared with FET+FDR (median=54.0, IQR=53.0), i.e. many fewer gene-sets are called for each consensus module (Figure 3A), though the numbers of selected gene-sets are positively correlated (Figure 3B, Spearman correlation's coefficient $\rho = 0.70$). Importantly, the biological information represented by the gene-sets selected by the two procedures is highly consistent. This was assessed by

manually curating the titles and associated descriptions of the enriched gene-sets for many modules, and we attempt to capture the consistency with visualizations in Figure 3C and D, using enriched GO terms. The distribution of gene-set level overlap coefficient in Figure 3C shows that for most modules, *gerr* essentially captures a subset gene-sets of the results of FTE+FDR. Alternatively, we can consider the *leading-edge* genes following the convention of GSEA (Aravind Subramanian *et al.*, 2005), which are defined genes within GOI that are associated with significantly enriched gene-sets. The distribution of overlap coefficients of leading-edge genes identified by either method is shown in Figure 3D. For most modules, *gerr* effectively identifies a subset of leading-edge genes that are also identified by FTE+FDR.

The enrichment of biological information is demonstrated by the distribution of normalized ranks (0 for gene-sets with the lowest FDR values, 1 for gene-sets with the highest FDR values) of gene-sets in Figure 3E, and more compellingly, of genes in Figure 3F. To generate both figures, gene-sets are ranked by the FDR values reported by the FET+FDR procedure. Gene-sets that are identified by both *gerr* and FET+FDR are visualized in contrast to the gene-sets that are only identified by FET+FDR. Figure 3E already shows that hits of *gerr* rank higher than FET+FDR hits. When considering genes underlying hit gene-sets, genes associated with gene-sets identified by *gerr* are almost three times more likely to rank at the top than genes of gene-sets identified by FET+FDR alone (Figure 3F).

Furthermore, we found that gene-sets identified by *gerr* contain far more genes that belong to GOI than gene-sets identified by FET+FDR (supplementary document 3, figure 1); and that for many modules, *gerr* even identified gene-sets that are not selected by FET+FDR (supplementary document 3, figure 2). These findings are consistent with our observations in the simulation studies that *gerr* can have higher sensitivities, especially when the noise term in the generative model is strong, and when relatively few genes in gene-sets contribute to GOI.

Taken together, gene-set enrichment with regularized regression is able to identify a succinct list of gene-sets that are representative of the biological functions associated with many sets of GOI as defined by the consensus modules identified in a community challenge.

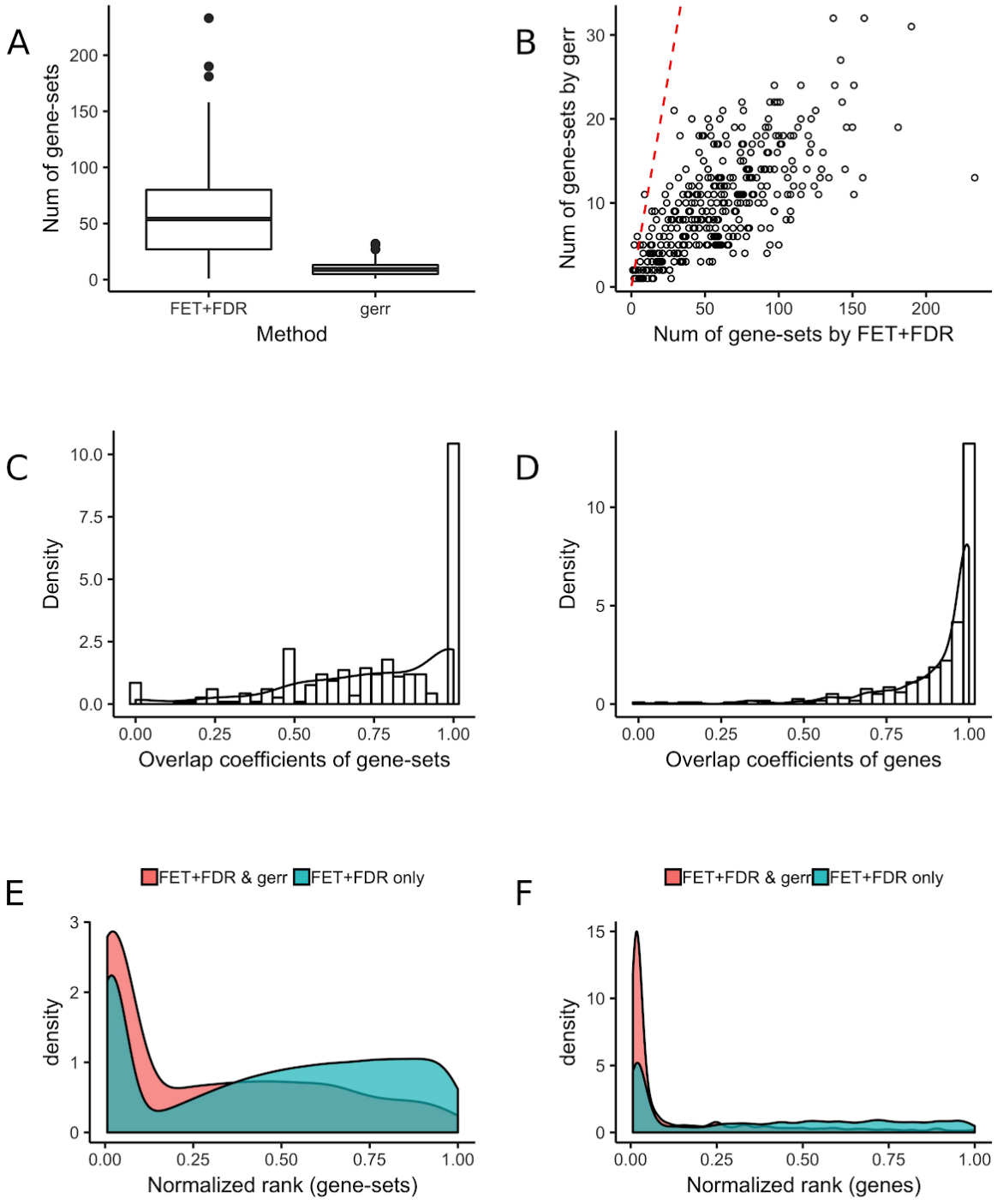


Figure 3. Comparison of hit gene-sets returned by *gerr* ($\alpha = 0.5$) and by the FET+FDR procedure (5% false discovery rate) from 377 consensus disease-related gene modules detected from the STRING network (Choobdar *et al.*, 2019). (A) Whisker-box plot of counts of hit gene-sets of each module. The horizontal bar indicates the median value, the boxes span between the first and the third quartile, outliers that are beyond 1.5 times the

interquartile range (IQR) are plotted individually as dots. (B) Scatter plot of the number of hit gene-sets of each module selected by both methods. The red dashed line represents $y = x$. (C) Distribution of overlap coefficients of hit GO terms returned by both methods of each module. The overlap coefficient in most modules is between 0.5 and 1, suggesting that the gene-sets returned by *gerr* are often a subset of gene-sets returned by FET+FDR. (D) Distribution of overlap coefficients of genes that are both present in a module and associated with the hit GO terms. (E) Normalised rank between 0 (top-ranking by ascending FDR values) and 1 (bottom-ranking) of gene-sets that are selected by *gerr* and FET+FDR (red), and gene-sets that are selected by FET+FDR alone (green). Gene-sets selected by *gerr* tend to enrich towards the top of the ranking. The enrichment is much strong if we consider the normalised rank of *genes*, as shown in (F). Genes that are associated with one or more gene-sets in a module are ranked by the lowest FDR value of gene-sets it belongs to. Apparently, gene-sets identified by *gerr* contain many genes that belong to top-ranking gene-sets. Therefore *gerr* enriches biological information despite sparse solutions.

## Conclusions and discussions

In this paper, we present a new perspective to the problem of gene-set enrichment. By considering the membership of genes in GOI as a binary response and gene-sets as features, we transform the gene-set enrichment problem to a regression problem explicitly allowing a dependency between individual gene sets. From this viewpoint, we propose the method of gene-set enrichment with regularized regression, and demonstrate its value by an implementation based on the elastic net regularization technique. Software codes and data are freely available through the open-source R package *gerr* and the GitHub repository.

The motivation and outcome of gene-set enrichment with regularized regression is quite different from canonical gene-set enrichment analysis techniques that have been comprehensively reviewed in multiple studies including (Huang *et al.*, 2009) and more recently (Rahmatallah *et al.*, 2016). While most other methods seek to identify all gene-sets that are over- or under-presented in a set of GOI, or more broadly, all gene-sets that are associated with some classification scheme of genes and finally some phenotype, *gerr* aims at identifying a minimum set of gene-sets that are representative. Apparently, there can be a potential loss of information in the latter approach since some gene-sets may be ignored simply due to redundancy with other gene-sets, or due to the parameter setting that controls the sparsity of the results. The gain is nevertheless also obvious: the regression approach is able to identify a few gene-sets that are strongly associated with the set of GOI, and the application of regularization techniques such as the elastic net ensures that the regression problem is well-posed even there are overlapping genes between gene-sets. The gain can be particularly valuable for humans to interpret the results when many gene-sets are aggregated from different sources and therefore redundancy and dependency can hardly be ignored.

When many redundant gene-sets are used, a sparse solution to the gene-set enrichment problem has been so far only possible if the gene-sets are organized in a tree structure that

can be exploited by the gene-set enrichment analysis method, *e.g.* topGO (Alexa *et al.*, 2006). Otherwise, most methods assume independence between gene-sets, an assumption that is unfortunately often invalid thanks to the ever-increasing volume of biological knowledge that is embodied in heterogeneous, partially redundant gene-sets. Gene-enrichment with regularized regression, in contrast, applies both to structured gene-sets, e.g., GO and Reactome, and loosely structured or unstructured gene-sets, such as those in the MSigDB database (Aravind Subramanian *et al.*, 2005) and the CREEDS database (Wang *et al.*, 2016) that are constructed by manual or semi-automatic curation of multiple datasets. Given that technologies such as single-cell and spatial omics (Sturm *et al.*, 2018; Rodriques *et al.*, 2019) and functional genomic screening (Haney *et al.*, 2018; Pluvinage *et al.*, 2019) are enriching our knowledge in cell-type and cell-state-specific gene expression and function at an unprecedented high rate, we envision that number of gene-sets that are available for enrichment analysis will grow exponentially, and most of them will be unstructured (at least for the near future), necessitating redundancy-agonistic methods like *gerr.*

Another major advantage of the regression approach is that it can be naturally extended to accommodate other data types of the dependent variable. In this work, we used linear regression to demonstrate the feasibility, though logistic regression is equally legitime. In this vein, different types of dependent variables can be modelled by choosing an appropriate link function (McCullagh and Nelder, 1989). For instance, in contrast to the simple enrichment model that we described here, the characteristic enrichment, in which one set of GOI is compared against other sets of GOI and therefore can be seen as a classification problem, can be tested using multinomial regression. On the other side of the model, the association between gene-sets and genes can be extended to more complex relationships beyond binomial (for instance effect size as continuous variables), and in case of necessity, covariates can be modelled in the framework of linear regression. For instance, it is almost trivial to control for gene sets biases, such as gene length, simply by incorporating such a bias it into the model, similar to Mi *et al.* (2012). The great flexibility of generalized linear models (Dobson and Barnett, 2008) allows developers and users extend the scope of *gerr* to be used in many areas of bioinformatics and genomics analysis where gene-set level interpretation is important, including disease understanding and drug discovery (Moisan *et al.*, 2015; Drawnel *et al.*, 2017).

## Acknowledgements

The authors thank Manfred Kansy, Martin Ebeling, and Fabian Birzele for the support of the work. The work has benefited from discussions within the Bioinformatics and Exploratory Data Analysis (BEDA) team. The internship of Mr. Tao Fang was sponsored by the Pharmaceutical Sciences department of Roche Pharma Research and Early Development, Roche Innovation Center Basel.

## Funding information

The work was supported by F. Hoffmann-La Roche Ltd.

## Conflict of Interest

None declared.

## References

Agresti,A. (2015) Foundations of linear and generalized linear models John Wiley & Sons.

Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.

Choobdar,S. *et al.* (2019) Open Community Challenge Reveals Molecular Network Modules with Key Roles in Diseases. *bioRxiv*, 265553.

Dobson,A.J. and Barnett,A. (2008) An introduction to generalized linear models Chapman and Hall/CRC.

Drawnel,F.M. *et al.* (2017) Molecular Phenotyping Combines Molecular Information, Biological Relevance, and Patient Data to Improve Productivity of Early Drug Discovery. *Cell Chem. Biol.*, **18**, 624–634.

Fabregat,A. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res*, **46**, D649–D655.

Friedman,J. *et al.* (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.*, **33**, 1–22.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Grossmann,S. *et al.* (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. *Bioinformatics*, **23**, 3024–3031.

Hackenberg,M. and Matthiesen,R. (2008) Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics*, **24**, 1386–1393.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Haney,M.S. *et al.* (2018) Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. *Nat. Genet.*, **50**, 1716--1727.

Hastie,T. *et al.* (2004) The Entire Regularization Path for the Support Vector Machine. *J. Mach. Learn. Res.*, **5**, 1391–1415.

Hellevik,O. (2009) Linear versus logistic regression when the dependent variable is a dichotomy. *Qual. Quant.*, **43**, 59–74.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55–67.

Huang,D. *et al.* (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, **8**, R183.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.

Khatri,P. and Drăghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

de Leeuw,C.A. *et al.* (2016) The statistical properties of gene-set analysis. *Nat. Rev. Genet.*, **17**, 353–364.

Li,T. *et al.* (2017) A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods*, **14**, 61–64.

McCullagh,P. and Nelder,J.A. (1989) Generalized linear models Chapman and Hall/CRC.

Mi,G. *et al.* (2012) Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression. *PLOS ONE*, **7**, e46128.

Mitchell,A.L. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.

Moisan,A. *et al.* (2015) White-to-brown metabolic conversion of human adipocytes by JAK inhibition. *Nat. Cell Biol.*, **17**, 57–67.

Pluvinage,J.V. *et al.* (2019) CD22 blockade restores homeostatic microglial phagocytosis in ageing brains. *Nature*, **568**, 187--192.

Rahmatallah,Y. *et al.* (2014) Comparative evaluation of gene set analysis approaches for RNA-Seq data. *BMC Bioinformatics*, **15**, 397.

Rahmatallah,Y. *et al.* (2016) Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief. Bioinform.*, **17**, 393–407.

Rodriques,S.G. *et al.* (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, **363**, 1463–1467.

Sergushichev,A. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, 060012.

Sturm,G. *et al.* (2018) Comprehensive evaluation of cell-type quantification methods for immuno-oncology. *bioRxiv*, 463828.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U A*, **102**, 15545–50.

Subramanian,Aravind *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102**, 15545–15550.

Szklarczyk,D. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447-452.

The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

The Gene Ontology Consortium (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.

Tibshirani,R. (1996) Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.*, **58**, 267–288.

Wang,Z. *et al.* (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.*, **7**, 12846.

Wu,D. and Smyth,G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res*, **40**, e133.

Zhang,J.D. *et al.* (2017) Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics*, **18**, 277.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 301–320.