W84–W88 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 10 June 2013*
doi:10.1093/nar/gkt516

FIDEA: a server for the functional interpretation of differential expression analysis

Daniel D'Andrea[1], Luigi Grassi[1], Mariagiovanna Mazzapioda[1] and Anna Tramontano[1,2,*]

[1]Department of Physics, Sapienza University, Rome, 00185, Italy and [2]Istituto Pasteur—Fondazione Cenci Bolognetti, Sapienza University, Rome, 00185, Italy

Received February 3, 2013; Revised May 10, 2013; Accepted May 18, 2013

ABSTRACT

The results of differential expression analyses provide scientists with hundreds to thousands of differentially expressed genes that need to be interpreted in light of the biology of the specific system under study. This requires mapping the genes to functional classifications that can be, for example, the KEGG pathways or InterPro families they belong to, their GO Molecular Function, Biological Process or Cellular Component. A statistically significant overrepresentation of one or more category terms in the set of differentially expressed genes is an essential step for the interpretation of the biological significance of the results. Ideally, the analysis should be performed by scientists who are well acquainted with the biological problem, as they have a wealth of knowledge about the system and can, more easily than a bioinformatician, discover less obvious and, therefore, more interesting relationships. To allow experimentalists to explore their data in an easy and at the same time exhaustive fashion within a single tool and to test their hypothesis quickly and effortlessly, we developed FIDEA. The FIDEA server is located at http://www.biocomputing.it/fidea; it is free and open to all users, and there is no login requirement.

INTRODUCTION

Differential expression analysis typically results in a long list of differentially expressed genes derived from the comparison of one or more samples. Although the results provide an essentially complete view of the analyzed transcriptomes, their functional interpretation is not always straightforward.

Once the differentially expressed genes have been identified and their statistical significance correctly assessed, it is essential to interpret the data to formulate hypotheses about the specific mechanisms involved and to select the most biologically significant transcripts for further validation.

The first step of the functional analysis is almost invariably an enrichment analysis (1) aimed at verifying whether a significant number of the identified genes belong to one or more specific pathways or functional categories. This is usually performed by statistically assessing whether a pathway or process is enriched in differentially expressed genes (2).

The enrichment analysis should be performed using different classifications of the genes, for example, KEGG pathways (3), Interpro (4), Gene Ontology Molecular Function, Biological Process and Cellular Component categories (5). Furthermore, one should explore the effect of selecting different thresholds for the *P*-value threshold as well as for the ratio of the gene expression values between the experimental system under study and the respective control to define the subset of differentially expressed genes. The matter becomes even more complex if more than two comparisons are required to interpret the experiment.

The correct interpretation of the results and, especially, the identification of particularly interestingly genes or functions among the differentially expressed ones are much more effective if associated to a deep knowledge of the biological system at hand and should, therefore, be done by the experimentalists. However, often they are not sufficiently expert to be able to effectively exploit the power of different tools and databases or to perform the comparison of the results of more than one experiment (which usually requires some scripting). This remains true even though there are a number of publicly available servers that allow functional enrichment analysis given a list of genes, the most used of which are DAVID (6,7), g:Profiler (8,9), Gorilla (10), High-Throughput GoMiner (11), Babelomics (12) and GeneCodis 3 (13). All these tools require the user to provide lists of genes, which implies that the identification of genes the transcripts of which are up- and downregulated needs to be performed separately and in advance. Furthermore, should the user wish to see how the results change when a different *P*-value or fold-change threshold is applied to identify

*To whom correspondence should be addressed. Tel: +39 064 991 4550; Fax: +39 064 957 697; Email: anna.tramontano@uniroma1.it differentially expressed transcripts, the list has to be rebuilt, resubmitted to the server and the results compared. Two of the aforementioned servers, Babelomics and GeneCodis 3, give the possibility of performing the analysis in parallel on two different lists of genes, for example, upregulated and downregulated ones or deregulated in different experiments, but the burden of comparing the results is still left to the user.

The aforementioned considerations prompted us to provide experimentalists with a more user-friendly tool for analyzing their data from a functional point of view. The FIDEA tool was developed through a number of iterations with our collaborating experimental groups, and we believe that the resulting system is sufficiently easy to use and at the same time complete and flexible to be useful in the functional analysis of differential expression experiments.

DESCRIPTION

The FIDEA server allows the user to directly input the results of a differential expression analysis, for example, by uploading the output of cufflinks (14,15) (one of the most used tools for RNA-Seq analysis) or, alternatively, a formatted result file that can be easily obtained through other tools such as EdgeR (16) or DESeq (17). The input format in the latter case is simple and can be defined by the user (these tools do not have a single output format; therefore, the columns of the file where the different fields are stored needs to be specified).

The most commonly used gene IDs (Gene symbol, Entrez gene ID, Ensembl gene ID, UCSC gene ID and Refseq ID) are accepted as input. They can refer to one of the following species: *Homo sapiens, Mus musculus, Drosophila melanogaster, Danio rerio* and *Saccharomyces cerevisiae*.

On loading the input data, the system immediately shows the distribution of genes with *P*-value below a selected threshold (values >0.1 are not allowed) thereby permitting to quickly appreciate the fraction of up- and downregulated genes in the specific experiment. The fold change and *P*-value thresholds can be interactively modified to further filter the data, and this leads to the direct display of the updated distributions (Figure 1A).

If more than one entry for the same gene is present in the input, the user is warned, and information about their annotations is displayed. The entry with the lowest *P*-value is used in all subsequent analyses.

Once a *P*-value and a fold change thresholds are selected, the enrichment analysis is performed considering the upregulated and downregulated genes both together and separately. This is relevant, as the detection of a statistically significant enrichment depends on the number of deregulated genes in a functional category compared with what is expected by chance and thereby, in specific cases, the results might differ if the upregulated and downregulated genes are considered together or separately. If a pathway, for example, has a significant number of upregulated genes and a few downregulated genes, the total number of differentially expressed genes in the pathway might turn out not to be statistically significant, whereas computing the enrichment of the upregulated genes separately might highlight an implication of the pathway in the system under study.

The categories that are considered for the analysis are KEGG, Interpro (Families, domains, sites and repeats), Gene Ontology Molecular Function (all evidence codes), Gene Ontology Biological Process (all evidence codes), Gene Ontology Cellular Component (all evidence codes) and GoSlim.

The statistical significance of the enrichment is computed using the hypergeometric test, the resulting *P*-values are corrected using the Benjamini and Yekutieli FDR method (18). The background distribution is, by default, the distribution of all the genes for the selected organism, but the user can also provide his/her own list of genes.

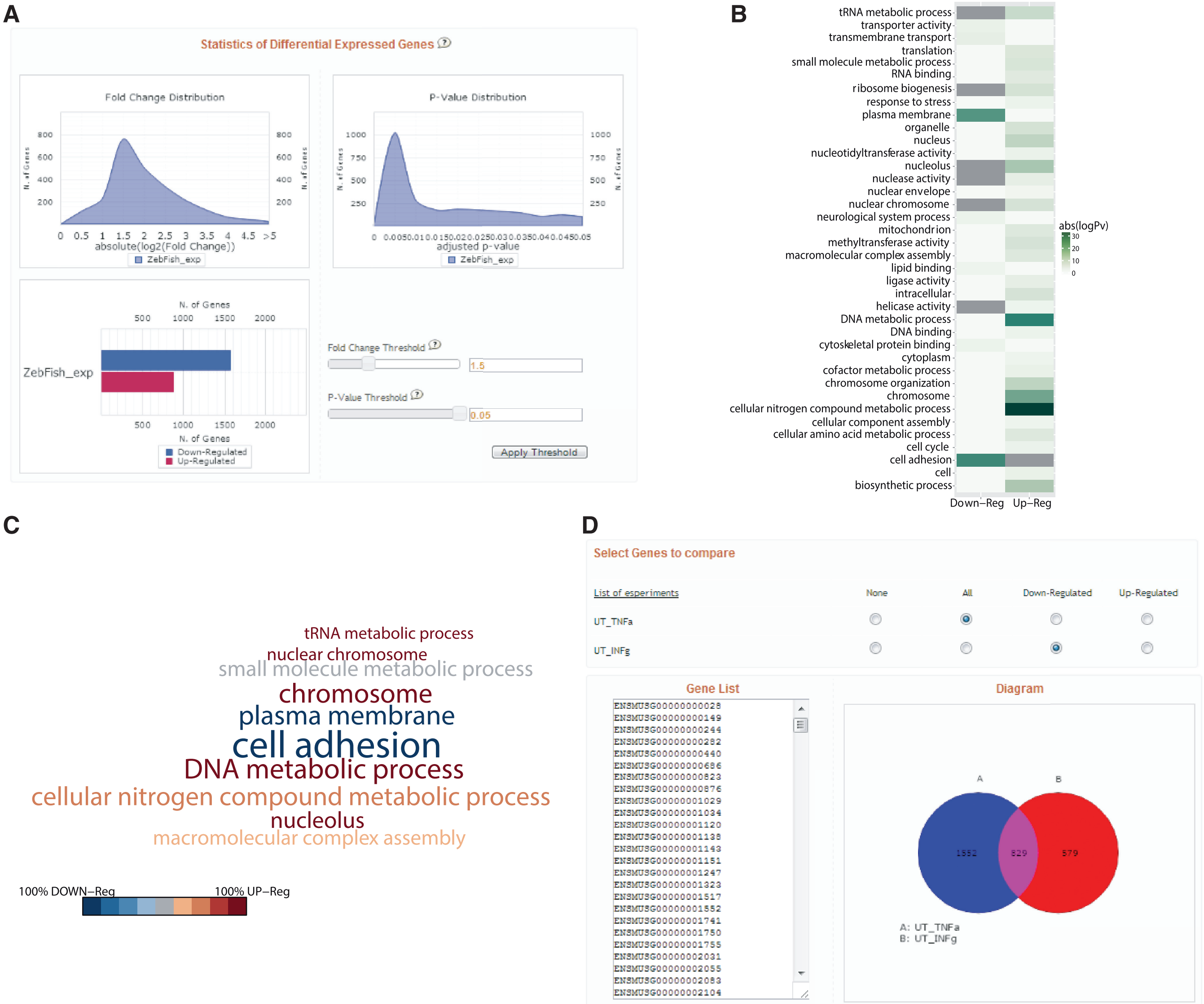
The significantly enriched functional categories (according to the corrected *P*-value and fold change thresholds selected by the user) can be displayed in different ways.

For the analysis performed on the upregulated and downregulated genes taken separately, the user obtains:

(i) an interactive dynamic heat map showing the absolute log10 of the corrected *P*-value. The rows of the heat map can be interactively ordered according to the *P*-value, the number of differentially expressed genes belonging to the category or alphabetically by category name. The list of the genes contributing to the category can be obtained by clicking on the corresponding cell;
(ii) an interactive table reporting the category name, the *P*-value and the corrected *P*-value, the fold enrichment and the number of differentially expressed genes. On clicking the latter, the system shows the list of the genes;
(iii) a static publication-ready heat map (Figure 1B) reporting the 60 categories with the lowest corrected *P*-values;
(iv) a text table (csv format and downloadable).

For the analysis that considers up- and downregulated genes together, the system provides:

(i) a dynamic interactive barplot listing the various enriched categories in ascending order of corrected *P*-value and the percentage of down-regulated and up-regulated genes in each of them in different colors. The list of the genes contributing to the category can be obtained by clicking on the corresponding bar;
(ii) a ‘word cloud’ where the functional categories are shown with a character size related to their enrichment (according to the corrected *P*-value) and in different colors according to the extent by which the pathways or categories are enriched by up- or downregulated genes (red to blue, respectively) (Figure 1C);
(iii) an interactive table reporting the category name, the *P*-value and the corrected *P*-value, the fold enrichment and the number of differentially expressed genes. On clicking the latter, the system shows the list of the genes;
(iv) a text table (csv format and downloadable).



List of esperiments	None	All	Down-Regulated	Up-Regulated
UT_TNFa	○	●	○	○
UT_INFg	○	○	●	○

```
ENSMUSG00000000028
ENSMUSG00000000149
ENSMUSG00000000244
ENSMUSG00000000282
ENSMUSG00000000440
ENSMUSG00000000686
ENSMUSG00000000823
ENSMUSG00000000876
ENSMUSG00000001029
ENSMUSG00000001034
ENSMUSG00000001120
ENSMUSG00000001138
ENSMUSG00000001143
ENSMUSG00000001151
ENSMUSG00000001247
ENSMUSG00000001323
ENSMUSG00000001517
ENSMUSG00000001552
ENSMUSG00000001741
ENSMUSG00000001750
ENSMUSG00000001755
ENSMUSG00000002031
ENSMUSG00000002055
ENSMUSG00000002083
ENSMUSG00000002104
```

Figure 1. The figure shows an example of the results of FIDEA. Panel (**A**) is the first page where, on uploading the data, the user has an overview of the distributions of fold changes and *P*-values for up- and downregulated genes and can interactively modify the *P*-value and fold-change thresholds. The results of the GOSlim analysis are shown as both a heat map (**B**) and a word cloud (**C**). Panel (**D**) shows an example of how data from different experiments can be combined and subsequently analyzed.

When more than one experiment is uploaded, the user can obtain a list of genes that are in common among experiments. These can include genes that are upregulated in all the experiments, upregulated in one and downregulated in another and so forth (Figure 1D). The results are shown as a Venn diagram and as a list of genes that can be directly submitted to the functional enrichment analysis or downloaded.

The server is regularly updated in parallel with new releases of Ensembl and of the functional classification annotations.

As an example, Figure 1 and Supplementary Figures 2–6 show some of the results obtained using the server for the data described previously (19), the authors of which performed an RNA-seq experiment of Id2a-deficient retinae obtained from zebrafish embryos in which Id2a expression was blocked by morpholino-mediated knockdown. As described by the authors, the data show an enrichment of downregulated genes belonging to the ‘‘cell adhesion’’ GO biological process and an enrichment of upregulated genes in the ‘RNA processing’ and ‘nitrogen compound biosynthesis’ processes.

The different results that can be obtained from the system according to the different functional categories are shown in the Supplementary Figures. These were obtained using the data from RNA sequencing experiments aimed at identifying transcripts expressed in human islets of Langerhans under control conditions or following exposure to pro-inflammatory cytokines (20).

IMPLEMENTATION

The FIDEA core consists in a Perl code and a set of Perl modules. The Perl modules are used to process the input, convert the gene IDs, perform the functional annotation and create the textual output files. The Benjamini and Yekutieli FDR is performed by the Statistics::Multtest Perl package. The R language and the ggplot2 package are used to create publication-quality PDF output images. The server front-end is implemented in standard HTML markup language using the Javascript programming language and AJAX technologies using the jQuery library. CanvasXpress is used for displaying the interactive images. The server runs under the Linux (Debian) operating system on a machine with 4# Intel Xeon E7-4820 2.00 GHz processors and 80 GB random access memory.

The associations between functional annotations and gene ID are stored in a local MySQL database (see Supplementary Figure S1 for a detailed scheme of the DB). Because of the large number of independent databases, the identifiers are, in most cases, redundant. FIDEA maps all functional annotations to the ENSEMBL gene IDs (21) and converts all supported gene IDs (Entrez, UCSC, gene symbol and Refseq) to ENSEMBL gene IDs. Organism-specific IDs can also be used, namely, Flybase ID, Gene and Annotation Symbol for *D. melanogaster*, Gene Name, Zfin Gene ID for *D. rerio* and SGD Systematic Name, Primary SGD and Gene Name for *S. cerevisiae*.

Regardless of this internal conversion, the final results are given using the original gene ID provided by the user.

For all three ontologies included in Gene Ontology (Biological Process, Molecular Function and Cellular Component) FIDEA applies the 'true path rule' (22): any gene associated with a given GO term is always associated with the ancestors of that term leading back to one step before the ontology root. This ensures an exhaustive annotation, even though it may produce some redundancy. Because of this, FIDEA also includes an enrichment analysis using the GO Slim annotations (23,24) that contain a smaller subset of GO terms.

Functional and structural annotations for protein families, domains and functional sites are retrieved from INTERPRO. The functional annotations for metabolic pathways are derived from the KEGG pathway database.

DISCUSSION

We describe here a publicly available and regularly updated server devoted to the functional analysis of differentially expressed genes.

The main features of the tool that make it different from what is already available are the possibility of directly processing the results of the differential expression analysis, of interactively modifying the *P*-value and fold change thresholds used for selecting the genes, of analyzing up- and downregulated genes separately or together and to directly analyze and compare lists of genes obtained from more than one comparisons. This, together with an easy to use interface and with the possibility of displaying the data in different ways (tables, heat maps and word clouds), makes the tool especially appropriate to be used in the functional interpretation of data derived from microarrays or RNA-seq experiments by the investigators themselves.

In the future, we plan to include the possibility for the user to upload newly sequenced and annotated genomes and to link to information from public data such as, for example, expression levels of the genes of interest in different tissues or disease states.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–12.

ACKNOWLEDGEMENTS

The authors are grateful to the groups of Prof. Gulino, Prof. Bozzoni and Prof. Levrero who helped debugging and improving the server and to Paolo Marcatili and all the other members of the Biocomputing Unit for useful discussions and for testing the server. They are indebted to the developer of CanvasXpress, Isaac Neuhaus, for his responsiveness and kindness.

FUNDING

Epigenomics Flagship Project—EPIGEN; KAUST [KUK-I1-012-43]; PRIN Project [20108XYHJS]. Funding for open access charge: Epigenomics Flagship Project—EPIGEN.

Conflict of interest statement. None declared.

REFERENCES

1. Glazko,G.V. and Emmert-Streib,F. (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, **25**, 2348–2354.
2. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
3. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
4. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
5. GO Consortium. (2013) Gene ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
6. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
7. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
8. Reimand,J., Arak,T. and Vilo,J. (2011) g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.*, **39**, W307–W315.
9. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
10. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

11. Zeeberg,B.R., Qin,H., Narasimhan,S., Sunshine,M., Cao,H., Kane,D.W., Reimers,M., Stephens,R.M., Bryant,D., Burt,S.K. *et al.* (2005) High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, **6**, 168.
12. Medina,I., Carbonell,J., Pulido,L., Madeira,S.C., Goetz,S., Conesa,A., Tarraga,J., Pascual-Montano,A., Nogales-Cadenas,R., Santoyo,J. *et al.* (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.*, **38**, W210–W213.
13. Tabas-Madrid,D., Nogales-Cadenas,R. and Pascual-Montano,A. (2012) GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res.*, **40**, W478–W483.
14. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
15. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
16. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
17. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
18. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
19. Uribe,R.A., Kwon,T., Marcotte,E.M. and Gross,J.M. (2012) Id2a functions to limit Notch pathway activity and thereby influence the transition from proliferation to differentiation of retinoblasts during zebrafish retinogenesis. *Dev. Biol.*, **371**, 280–292.
20. Eizirik,D.L., Sammeth,M., Bouckenooghe,T., Bottu,G., Sisino,G., Igoillo-Esteve,M., Ortis,F., Santin,I., Colli,M.L., Barthson,J. *et al.* (2012) The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. *PLoS Genet.*, **8**, e1002552.
21. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
22. GO Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
23. GO Consortium. (2012) The gene ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
24. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.