Hassan and Shanak *BMC Bioinformatics* (2019) 20:20
https://doi.org/10.1186/s12859-018-2581-8

BMC Bioinformatics

SOFTWARE Open Access

GOTrapper: a tool to navigate through branches of gene ontology hierarchy

Hezha Hassan[1,2*] and Siba Shanak[3]

Abstract

Background: Gene Ontology (GO) is a useful resource of controlled vocabulary that provides information about annotated genes. Based on such resource, finding the biological function is useful for biologists to come up with different hypotheses and help further investigations of an experiment. The biological function for desired genes and gene associations is picked up from a randomly chosen list or through the analysis of differential gene expression. Many tools have been developed to utilize GO knowledge and cluster genes according to relevant biological functions. The retrieved GO terms include both specific and non-specific terms, which is not user-friendly in terms of data analysis. Thus one approach is still missing, which allows navigating through different levels of GO hierarchy manually.

Result: We developed a tool, GOTrapper, which allows moving up or down to the very bottom of the GO hierarchy. This is performed manually by the user, based on an assigned threshold. This tool grabs the shared terms by the desired set of input genes of *Homo sapiens*. Here, two inputs are possible. "Within" is to find associated terms within one gene list, and "Between" is to find associated terms between two lists. The tool also provides the option to return the terms with the pre-selected evidence codes.

Conclusion: GOTrapper is a user-friendly Java tool that helps the user move up and down the ontology tree, which leads to new hypotheses and devising new association of the input genes. It also allows returning terms of associated genes based on selected evidence codes. This tool can be accessed and is freely available at https://github.com/BioGeneTools/GOTrapper.

Keywords: Gene ontology, GO term refinement, Gene association

Background

The Gene Ontology (GO) is a controlled vocabulary of gene annotations, which was founded in 1998 to provide interpretation of biological functions that are associated with individual genes [1, 2]. The GO terms were placed in a hierarchy and are structured as an acyclic directed graph. They are classified into three vocabularies: Biological Processes, Molecular Functions, and Cellular Components. Each term may have more than one parent and more than one child. Going down the graph, the terms get more specific.

Gene Ontology is a powerful tool and the largest resource for cataloguing gene function continuously used in data analysis and functional prediction. The usage of this tool by inexperienced users might draw false conclusions [3, 4]. In microarray and RNA-seq experiments, GO is used broadly as a tool to group genes as well as to determine term enrichment of different biological processes, molecular functions, and cellular components. This helps explain the biology of the sample conditions.

Many methods and tools have been developed to find terms and perform enrichment analysis from expression data. Nonetheless, there is still some hidden information needed to be revealed from GO, many redundant terms, and a lack of simplicity of the tools; especially for biologists.

One strategy to speculate the gene ontology list for an experiment is to find the enriched GO terms. Several statistical methods can be used for this analysis, such as the hypergeometric distribution, Fisher Exact test, and binomial test [5]. These methods serve in

* Correspondence: hezha.hassan@hezhalab.com
[1]Public Health Laboratory, Sulaimaniyah, Kurdistan Region 46001, Iraq
[2]Genome Informatics, Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany
Full list of author information is available at the end of the article

mining the statistically significant enriched terms and suffer from redundancies, due to the inclusion of less specific terms. There exist tools and algorithms that manipulate different techniques to reduce those redundancies, through removing parent terms from the list of enriched terms [6–8]. Still, the remaining 'last' children terms, which are extracted by the different statistical methods mentioned above, have important information that could be lost at the expanded level of the maintained children terms.

With increasing biological information and expanding ontological annotations, it is highly beneficial for biologists to have on hand tools to find the associations between the different desired sets of genes with less redundancy. Some tools have made this option available [6–14]. Some of these tools require the user to provide extra information such as *p*-values or expression data, which may be obtained from differential expression analysis. Other tools allow provisioning of the gene lists alone but they handle enrichment analysis. This causes the loss of specific associations between genes at the end of the branch of the GO tree.

There are also a number of tools, e.g.; web-based and plugins that provide a variety of functions but require internet connection or third-party software. This could be complicated or less helpful; especially for inexperienced users [8, 12–22].

It is important, especially for wet lab experimentalists, to utilize gene ontology resources in finding different gene associations and in making new hypotheses via the manual crawling through stages of hierarchy for the ontology. To our knowledge, there is no tool to provide such options.

In this paper, we developed a user-friendly, open source, and cross-platform tool to help experienced and inexperienced users in finding gene set associations. This tool offers manual navigation through ontology hierarchy by using the gene names only, and without the need for expression data, *p*-value, fold change calculations, or other inputs.

Implementation

Figure 1b depicts the workflow of the tool. The tool is open source and built in Java. GOTrapper does not rely directly on the GO database. It derives all the mapping and annotations from two databases, GO.db [23] and org.Hs.eg.db [24], from Bioconductor [25].

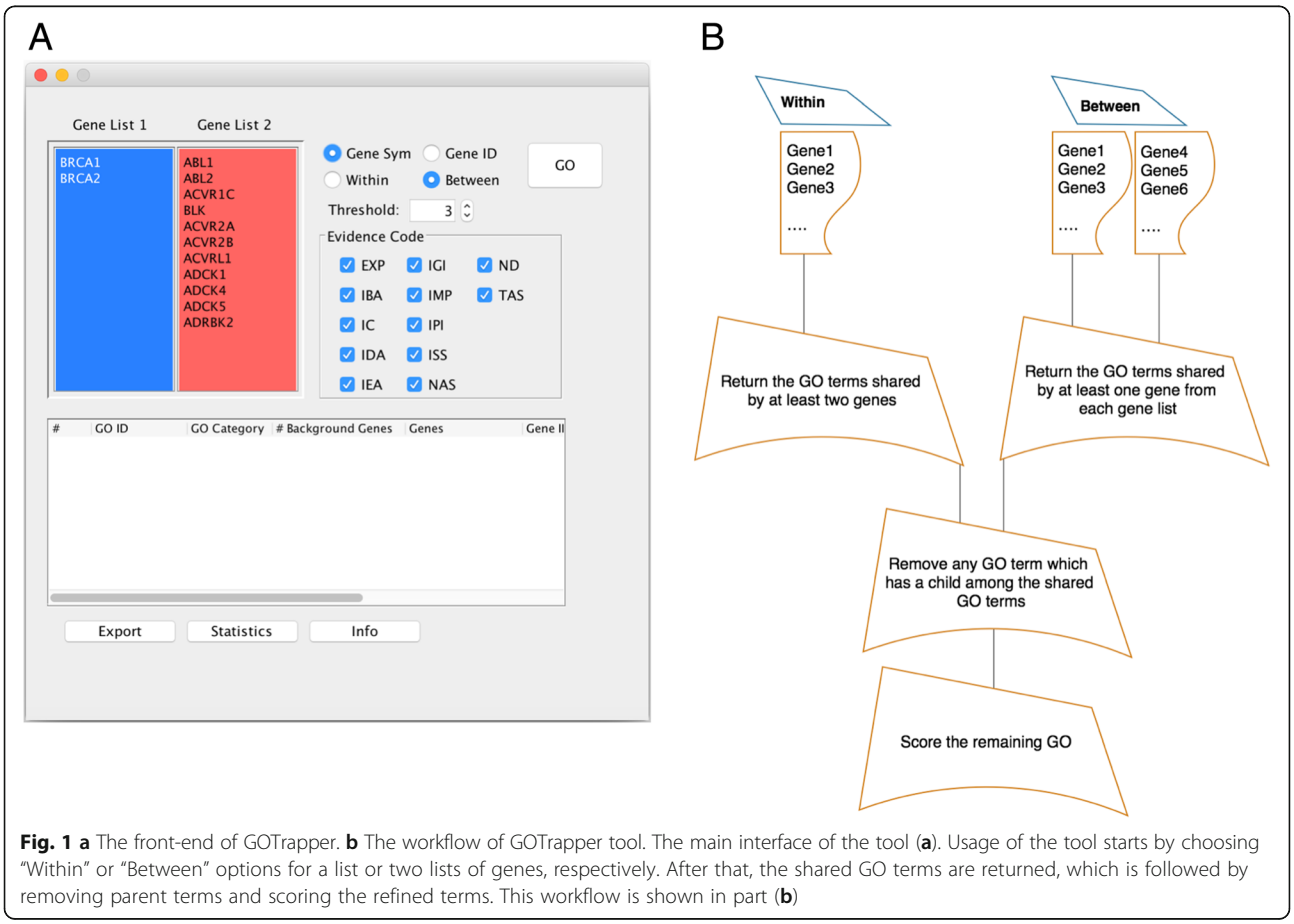

Fig. 1 **a** The front-end of GOTrapper. **b** The workflow of GOTrapper tool. The main interface of the tool (**a**). Usage of the tool starts by choosing "Within" or "Between" options for a list or two lists of genes, respectively. After that, the shared GO terms are returned, which is followed by removing parent terms and scoring the refined terms. This workflow is shown in part (**b**)

Finding most specific GO terms

In this first part of the algorithm, the GO terms which are shared by the desired number of genes would be defined (Fig. 2). After that, any shared terms with one or more children are removed. This helps get the most specific GO terms. Namely, the last shared terms remain at the end. The tool makes use of an option called "Threshold" to allow the users to control and pick up different levels of the tree.

Scoring of the resulting GO terms

After retaining the most specific shared terms, we applied a scoring system to provide more meaningful information to the user for ranking the GO terms. The terms are scored based on the negative log likelihood:

$$Score(t) = -\log(p(t))$$

where *p(t)* is calculated by:

$$p(t) = \frac{2}{g(t)}$$

where the constant number of '2' was assigned to it in the tool as the number of the minimum background genes in a shared GO term is two, and *g(t)* is the number of background genes, which is the total number of genes, annotated to the *t* term. The lower the number of background genes annotated to a term, the lower the *score(t)* would be. We assume that the lesser the *score(t)*, the more specific the term *t* is, as the number of annotated genes decreases in the terms going down the hierarchy.

Threshold

The flexibility of GOTrapper increases by introducing the "Threshold" option. The minimum threshold is "2", i.e.; the retained GO terms must be shared by at least two input genes. This option also provides the user with the ability to control the returned level (going up or down the tree) of the shared GO terms in the GO hierarchy by increasing or decreasing the threshold.

Examples

We use different sets of genes [26, 27] to implement both functionalities ('Within' and 'Between') of GOTrapper.

Grouping a list of genes using the "within" option

In using high-throughput microarray and next generation sequencing technologies, researchers compare the expression data for a large number of genes in two (or more) different states. Exemplary research was conducted on human prostate cancer using RNA-seq data [26], where malignant samples were compared with non-malignant. The study ended up with a large number of genes being expressed differently between the two conditions. The comparison held by the researchers resulted in a large number of GO terms with an exceedingly large number of background genes. The most common groups of GO terms achieved by the researchers were related to metabolic and cellular processes; which are known to be fundamental needs for the establishment of cancer. Other groups were related to regulation, development, nucleic acid binding, localization, biological adhesion, catalytic activity, structural molecule activity, immune response, and multicellular organism activity. We aim to compare a list of 815 differentially expressed genes (> = 2fold change) (Additional file 1), from the prostate cancer research mentioned above. We want to find possible associations among the genes and understand biological processes as well as molecular functions of the genes in highly specific terms based on the GO annotations. In this example, we used a threshold value of 10 (each GO term to be shared by at least 10 genes). Out of the total 1077 GO terms, which are shared among the respective genes, 319 most highly specific shared GO terms were trapped (Additional file 2). We classified the group of genes the same way discussed above. We could find a large number of genes related to regulation and developmental processes in the most specific GO terms. A large number of GO terms also allocate to metabolic and cellular processes. Nonetheless, genes associated with cell adhesion were rather so scarce. Other groups of GO terms met nicely with the classifications held by the researchers relating to prostate cancer. Indeed, after assigning our score scheme to the study, the terms got more specific and less redundant. This in turn aids in the easier and more efficient interpretation of biological data than when handling a large number of nonspecific redundant GO terms.

Comparing two lists of genes using the "between" option

Here we want to compare two lists of genes. A list of six genes, known to be related to urea cycle disorders (CPS1, OTC, ASS1, ASL, ARG1, and NAGS) [27], is compared to a list of 114 Chromatin Remodeling genes (Additional file 3), which modify the chromatin architecture and make it accessible for transcription. Current research investigates how aberrant chromatin remodeling, among other epigenetic factors, is correlated with a wide spectrum of diseases [28]. Many diseases associated with chromatin remodeling are related to metabolism [29]. One such example of metabolic diseases is the urea cycle disorder. Current research has investigated the epigenetic modifications expected to be correlated with urea cycle disease [30]. We assume that a researcher has intention to

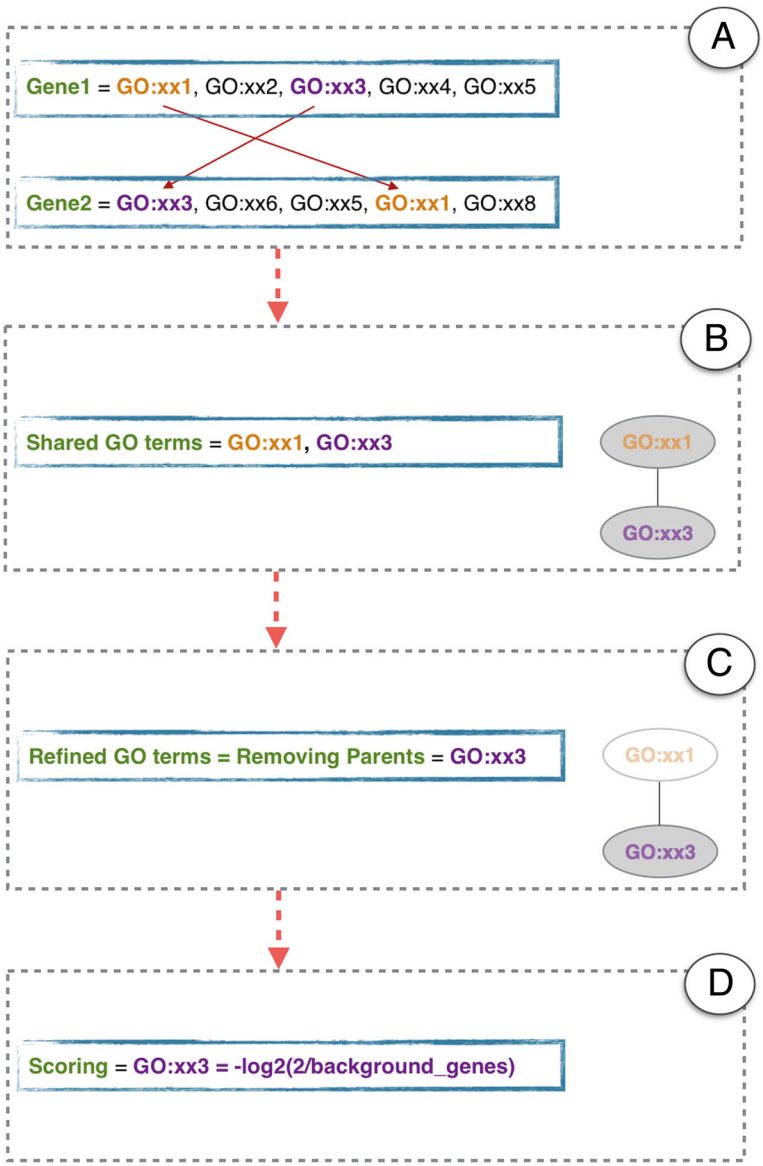

Fig. 2 GOTrapper Algorithm. The algorithm is shown in Parts A-D. The GO terms of two genes are returned in Part A. The terms that are shared by the two genes remain in Part B. For the two returned shared terms, the one which is a parent term is removed in Part C. The refined shared term without a parent is scored in Part D

investigate the correlation between the list of chromatin remodeler genes and the genes related to urea cycle disorder. Using this option to compare these two lists of genes, we find GO terms that are shared by at least one gene from each list by setting the threshold to two. This comparison resulted in 295 total shared GO terms and was refined to 72 highly specific shared terms (Table 1, Additional file 4). Interestingly, many shared GO terms between the two groups were associated with metabolism, including biosynthetic and catabolic processes. Many cellular processes are linked to the response to internal metabolites, including the ammonium ion, among others. Regulation involved metabolic processes associated with nitrogen compounds. Some abundant transport processes were also related to nitrogen compound transport. Additionally, response to amine stimulus was also involved in the set of GO terms. Many other processes were associated with development. Since the threshold was set to the minimum value, results are highly specific and the derived number of GO terms is much lesser. This could nicely help in supporting the hypothesis that urea cycle disease has a strong correlation with epigenetic modifications that can predispose as a result of, e.g., environmental factors.

Table 1 Top 10 terms shared by urea cycle disorder and chromatin remodeling genes

GO id	GO category	Background Genes	Genes	Score	GO term
GO:0071242	BP	27	CPS1, HDAC4	3.7549	cellular response to ammonium ion
GO:0045909	BP	29	CPS1, HDAC4	3.858	positive regulation of vasodilation
GO:0032964	BP	38	ARG1, ASL, ASS1, CPS1, NPM1, OTC, METTL3, NAGS	4.2479	collagen biosynthetic process
GO:0071398	BP	42	ARG1, ASS1, BNIP3, CPS1, HDAC2, HDAC5	4.3923	cellular response to fatty acid
GO:0014075	BP	47	BNIP3, CPS1, RB1, SIRT1	4.5546	response to amine stimulus
GO:0060416	BP	50	ARG1, ASS1, CPS1, HDAC2, KMT2A, OTC, HDAC4, HDAC5, CBX7	4.6439	response to growth hormone stimulus
GO:0055081	BP	53	ARG1, ASS1, CPS1, KAT2A, OTC, RB1, CHD8	4.7279	anion homeostasis
GO:0060135	BP	57	CPS1, OTC, CHMP3	4.8329	maternal process involved in female pregnancy
GO:1901655	BP	63	ASS1, CPS1, PHC1, HDAC2, SMARCD1	4.9773	cellular response to ketone
GO:0070301	BP	66	CPS1, OTC, CHMP3	5.0444	cellular response to hydrogen peroxide

Results

We present GOTrapper; a methodology and user-friendly tool to devise new hypotheses and gene associations by going through the branches of the gene ontology tree by providing only the gene symbols or IDs.

The goal of GOTrapper is to assist researchers in finding gene associations and grouping the genes according to GO knowledge. A scoring system is provided to show the specificity of the terms. In addition, the tool allows the pre-selection of the evidence codes to be considered in the downstream analysis.

GOTrapper enables two types of input:

- '*Within*': This option allows the input of a list of gene symbols or IDs to find the GO terms and association within this list (Fig. 1).
- '*Between*': The purpose of this option is to find an association between two lists of genes in which the output GO terms have to be shared by at least two genes, each from a list (Fig. 1).

Conclusions

GOTrapper is a user-friendly and multi-platform tool designed for experienced and non-bioinformaticians to cluster and group input genes of *Homo sapiens*. This allows the prediction of new hypotheses and helps find associations among the genes based on GO terms. Thus, the branches of the GO tree can be analyzed manually. The tool allows selection of the desired evidence code to be included in the process. A scoring system is also provided to determine the specificity of the returned GO terms.

Availability of data and materials

Project name: GOTrapper.

Project home page: https://github.com/BioGeneTools/GOTrapper.

Operating system(s): Platform-independent.

Programming language: Java.

Other requirements: Java (v1.7 or higher).

License: GNU GPL.

Any restrictions to use by non-academics: No.

Additional files

Additional file 1: 815 DEGs from a prostate cancer study. (XLS 20 kb)

Additional file 2: 319 highly specific shared terms among 815 DEGs of the prostate cancer study. (TXT 693 bytes)

Additional file 3: 114 chromatin remodelers. (XLS 94 kb)

Additional file 4: 72 highly specific shared terms between 114chromatin remodelers and 6 urea cycle disorders. (TXT 5 kb)

Abbreviation

GO: Gene ontology

Acknowledgements

We would like to thank Prof. Dr. Volkhard Helms for his suggestions and support during the beginning of the tool development. We also wish to thank Prof. Dr. Jens Stoye for his insightful support and to the reviewers for their constructive comments that strengthened the manuscript.

Funding

This work had no source of funding except from the authors themselves.

Availability of data and materials

GOTrapper and the source code is publicly available on Github at: https://github.com/BioGeneTools/GOTrapper.

Authors' contributions

HH conceived the idea, developed the tool and wrote the manuscript. SS participated in developing the original manuscript and supervised the project. All authors have read, revised, and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Public Health Laboratory, Sulaimaniyah, Kurdistan Region 46001, Iraq. [2]Genome Informatics, Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany. [3]Faculty of Sciences, Arab American University-Palestine, P.O Box 240, Jenin, Palestine.

Received: 24 June 2018 Accepted: 11 December 2018
Published online: 11 January 2019

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Consortium GO. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.10614036.
2. Blake JA, et al. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43(D1):1049–56. https://doi.org/10.1093/nar/gku1179.
3. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008;9(7):509–15. https://doi.org/10.1038/nrg2363.NIHMS150003.
4. du Plessis L, Skunca N, Dessimoz C. The what, where, how and why of gene ontology–a primer for bioinformaticians. Brief Bioinform. 2011;12(6):723–35. https://doi.org/10.1093/bib/bbr002.
5. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13. https://doi.org/10.1093/nar/gkn923.
6. Jantzen SG, Sutherland BJ, Minkley DR, Koop BF. GO trimming: systematically reducing redundancy in large gene ontology datasets. BMC Res Notes. 2011;4:267. https://doi.org/10.1186/1756-0500-4-267.
7. Moutselos K, Maglogiannis I, Chatziioannou A. GOrevenge: a novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies. IEEE Trans Biomed Eng. 2011;58(12 PART 2):3522–7. https://doi.org/10.1109/TBME.2011.2164794.
8. Supek F, Bo˘snjak M, S˘kunca N, S˘muc T. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6(7). https://doi.org/10.1371/journal.pone.0021800.
9. Zambon AC, Gaj S, Ho I, Hanspers K, Vranizan K, Evelo CT, Conklin BR, Pico AR, Salomonis N. GO-elite: a flexible solution for pathway and ontology over-representation. Bioinformatics. 2012;28(16):2209–10. https://doi.org/10.1093/bioinformatics/bts366.
10. Bauer S, Grossmann S, Vingron M, Robinson PN. Ontologizer 2.0 - a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics. 2008;24(14):1650–1. https://doi.org/10.1093/bioinformatics/btn250.
11. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. Bioinformatics. 2007;23(22):3024–31. https://doi.org/10.1093/bioinformatics/btm440.
12. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol. 2007;8(1):3. https://doi.org/10.1186/gb-2007-8-1-r3.
13. Prüfer K, Muetzel B, Do H-H, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics. 2007;8(41):1–10. https://doi.org/10.1186/1471-2105-8-41.
14. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane H, Lempicki RA. DAVID: database for annotation, visualization, and Integrated Discovery. Gen Biol. 2003;4(9):60. https://doi.org/10.1186/gb-2003-4-9-r60.
15. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics. 2004;20(4):578–80. https://doi.org/10.1093/bioinformatics/btg455.
16. Martin D, Martin D, Brun C, Brun C, Remy E, Remy E, Mouren P, Mouren P, Thieffry D, Thieffry D, Jacq B, Jacq B. GOToolBox: functional analysis of gene datasets based on gene ontology. Genome Biol. 2004;5(12):101. https://doi.org/10.1186/gb-2004-5-12-r101.
17. Beißbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies with a group of genes. Bioinformatics. 2004;20(9):1464–5. https://doi.org/10.1093/bioinformatics/bth088.
18. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. BMC Bioinformatics. 2004;5:16. https://doi.org/10.1186/1471-2105-5-16.
19. Lee JSM, Katari G, Sachidanandam R. GObar: a gene ontology based analysis and visualization tool for gene sets. BMC Bioinformatics. 2005;6:189. https://doi.org/10.1186/1471-2105-6-189.
20. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: A GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38(SUPPL. 2):64–70. https://doi.org/10.1093/nar/gkq310.arXiv:1011.1669v3.
21. Zhou X, Su Z. EasyGO: gene ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics. 2007;8(1):246. https://doi.org/10.1186/1471-2164-8-246.
22. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for gene ontology enrichment analysis. Nucleic Acids Res. 2008;36(Web Server issue):358–63. https://doi.org/10.1093/nar/gkn276.
23. Carlson M. GO. db: A set of annotation maps describing the entire. Gen Ontol. 2013;3(0):2016 R package version 3.4.0.
24. Carlson, M.: org.Hs.eg.db: Genome wide annotation for Human. (2016) R package version 3.4.0.
25. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5. https://doi.org/10.1186/gb-2004-5-10-r80.
26. Myers JS, von Lersner AK, Robbins CJ, Sang Q-XA. Differentially expressed genes and signature pathways of human prostate Cancer. PLoS One. 2015;10(12):0145322. https://doi.org/10.1371/journal.pone.0145322.
27. Summar ML, Tuchman M. Urea cycle disorders overview. In: Pagon RA, Bird TD, Dolan CR, et al., editors. Gene Reviews; 2005. p. 1993–2002.
28. De Chiara F, Jalan R, Marrone G, Heeboll S, Montoliu C, Hamilton-Dutoit S, Garcia-Torres M, Andreola F, Rombouts K, Grønbæk H. Others: epigenetic modification of urea cycle enzymes in NAFLD animal models and patients: implications for novel therapeutic approaches. J Hepatol. 2018;68:359–60.
29. Etchegaray J-P, Mostoslavsky R. Interplay between metabolism and epigenetics: a nuclear adaptation to environmental changes. Mol Cell. 2016;62(5):695–711.
30. Feinberg AP. The key role of epigenetics in human disease prevention and mitigation. N Engl J Med. 2018;378(14):1323–34.