GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 3 No. 3 (2017): e45

Research Article

From Protein Variations to Biological Processes and Pathways with NET-GE

Samuele Bovo[1], Pietro Di Lena[2], Pier L. Martelli[1,*], Piero Fariselli[3], Rita Casadio[1,4]

[1]Biocomputing Group, CIG, Interdepartmental Center "Luigi Galvani" for Integrated Studies of Bioinformatics, Bio-physics and Biocomplexity, University of Bologna, Bologna, Italy
[2]DISI, University of Bologna, Bologna, Italy
[3]BCA, University of Padova, Padova, Italy
[4]Interdepartmental Center "Giorgio Prodi" for Cancer Research, University of Bologna, Bologna, Italy

*Correspondence: Email: gigi@biocomp.unibo.it

Received 2016-10-07; Accepted 2017-04-19

ABSTRACT

Gene enrichment analysis is a common technique for highlighting molecular pathways and biological processes of a phenotype. Such technique has recently evolved exploiting the information contained in biological networks. We developed NET-GE, a web server for network-based gene enrichment analyses. NET-GE defines functional associations between a list of genes/proteins and biological processes or pathways by identifying function-specific modules in a molecular interaction network. The peculiarity of NET-GE is the possibility to enrich terms not detectable by standard enrichment procedure. Here, we highlight with two specific applications the performances of NET-GE by computing which functional phenotypes can be associated with two different sets of genes related to Attention Deficit Hyperactivity Disorder and to an Obsessive-compulsive disorder, respectively.

KEYWORDS

Gene enrichment analysis; network-based gene enrichment analysis; functional association

INTRODUCTION

Technologies capable of investigating the organism complexity at different levels of resolution have been revolutionizing healthcare practice [1]. Genomic data are generated more and more to better define, at molecular levels, the origin of the different phenotypes. From a precision/genomic medicine prospective, such phenotypes need annotations in order to reconcile specific variations with common biological processes and pathways, such as GENE ONTOLOGY [2], KEGG [3] and REACTOME [4] pathways. For this purpose, functional association is routinely performed by applying gene enrichment analysis, a technique that assesses the statistically over-represented biological processes and pathways of a given gene/protein set [5].

Presently, enrichment analysis methods mainly group into two classes, standard and network-based. While standard methods rely only on the annotations characterizing the genes/proteins included in the input set, network-based methods consider them in the context of their interaction network. Thus, such methods exploit information derived from functional biological networks, modelling the complexity of the processes occurring in the cell, and implement algorithms that exploit graph properties (such as shortest paths and node degrees).

In the last year, several approaches exploiting the interaction networks for functional association analysis have emerged (see [6–8] for a comprehensive list of available tools). They may be classified into two main categories: A) methods that exploit the topology of the network to infer how similar are sets of genes/proteins, and B) methods that identify functionally related modules, inferring biological features from them. Among the available tools that perform network-based enrichment analysis, EnrichNet [9] and PINA v2.0 [10] are two of the most cited methods, representative of the A and B categories, respectively.

We recently developed NET-GE, a network-based gene enrichment analysis tool [11, 12]. NET-GE falls within the class B and it is based on a pre-processing phase aimed at identifying interconnected and compact modules in a molecular interaction network. However, differently from all the other approaches in class B, the modules found by our method are function-specific by construction, since they are built starting from seed sets collecting all the proteins related to a specific biological annotation

One of the main features of NET-GE is the possibility to enrich terms that are not originally present in the annotation of the starting gene/protein set (and thus not detectable through a standard enrichment). When tested on benchmark sets retrieved from the Online Mendelian Inheritance in Man (OMIM) resource (https://www.omim.org), NET-GE was able to enrich sets of genes related to the same disease, also highlighting new terms (i.e. terms not included in the annotations of the input set) [11].

Here, we present two study cases, demonstrating how NET-GE can help the interpretation and prioritization of variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder (ADHD) and the Obsessive-compulsive disorder (OCD).

METHODS

NET-GE background

The network-based enrichment makes use of precomputed annotation terms, as previously described [11]. Briefly, the human molecular-interaction network was downloaded from STRING v.10 (http://string-db.org). A second version of STRING, named STRING0.9, was obtained by retaining only the links with the STRING combined score ≥ 0.9. The database for annotating features were: GENE ONTOLOGY (as retrieved from the UniProt-GOA human 145 web resource: http://www.ebi.ac.uk/GOA); KEGG PATHWAY v77 and REACTOME PATHWAY v53. For each annotating feature, proteins sharing the same annotation term were collected in a seed set and then extended into a compact and connected module of the molecular-interaction network. Thus, the module was determined by computing all the shortest paths among the seeds genes/proteins and then by reducing the resulting sub-network into the minimal connecting network that preserves the distances among seeds. The minimal connecting network adds to the seeds a set of connecting nodes that are more reliably related to the reference annotation. Details about annotations and module extraction can be found in [11] and [12], respectively.

Over-representation analysis is performed by mapping the input set on each module and determining, through a Fisher's exact test, whether there are significant overlaps between the input set and the modules (seed sets in the case of standard enrichment). Multiple testing correction is then applied using the Bonferroni or the Benjamini-Hochberg (False Discovery Rate, FDR) procedure [13].

When we consider the standard enrichment, the background set is totally disconnected. On the contrary, with the network-based procedure we rely on the human interactome to precompute the annotation modules. Enrichment is computed over a changed reference set that includes also all the nodes connecting seeds with the same annotation. This may change the p-value.

NET-GE web server

A web server, implementing both a standard and a network-based gene enrichment was implemented as described in [12]. Briefly, NET-GE Web interface takes as input a list of genes/proteins (allowed identifiers are: UniProtKB AC, Ensembl and HGNC gene names). The enrichment can be performed considering the annotation modules based on STRING or STRING0.9. The enriched terms can derive from the GENE ONOLOGY (all the sub-ontologies), or from the KEGG or the REACTOME PATHWAYS. The user can select between two kinds of multiple testing correction methods (Bonferroni or the Benjamini-Hochberg correction), and the significance threshold. As output NET-GE reports: 1) two enrichment tables (one for the standard enrichment and one for network-based one), 2) a graph visualizing how the enriched terms are linked, and 3) the complete set of annotations (for both the enrichment modes). Terms not included in the annotations of the input proteins are highlighted with a double star.

RESULTS

To test the performance of NET-GE we used sets of proteins involved in Mendelian diseases [11]. We tested 244 different genetic disorders, each one associated to two or more proteins. Our method was able to detect functional associations not detectable by the standard enrichment. Moreover, the newly enriched terms that were absent in the original annotations of the input genes are likely to provide new knowledge on the phenotype under examination [11].

Here, we present two cases of study demonstrating how NET-GE can help the interpretation and prioritization of variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder and the Obsessive-compulsive disorder.

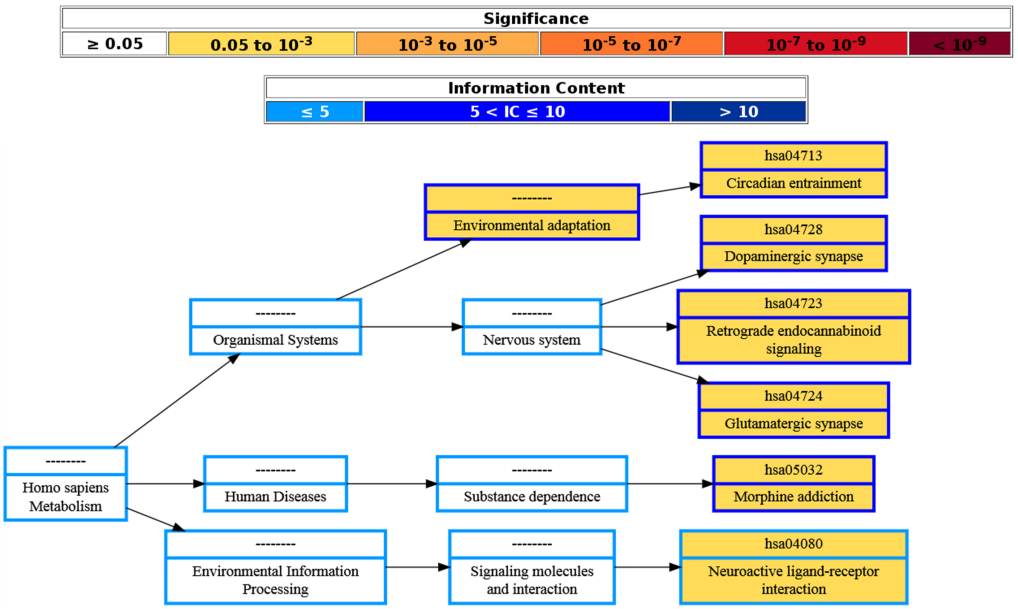
Attention Deficit Hyperactivity Disorder

In the following, we deal with a specific test set (http://net-ge.biocomp.unibo.it/enrich/tutorial) that includes two input proteins related to Attention Deficit Hyperactivity Disorder (ADHD; OMIM #143465), a neurodevelopmental disease of childhood affecting the cognitive and behavioral functions. The genetic disease is associated to variations in the dopamine receptors DRD4 (UniProtKB AC: P21917) and DRD5 (UniProtKB AC: P21918). Using as input the *DRD4* and *DRD5* genes, we carried out enrichment analyses by setting the significance threshold at 0.05 on the Bonferroni corrected p-values. Standard and network-based enrichments ran over the KEGG database. Terms enriched by NET-GE are shown in Figure 1. The standard enrichment on KEGG highlights neuroactive ligand-receptor interaction and dopaminergic synapse as the most significant pathways. The network-based procedure adds new terms, not associated to the input proteins, and involved in ADHD, considering the statistically significant subnetworks. The pathways sorted by significance are: circadian entrainment, morphine addiction, retrograde endocannabinoid signaling and glutamatergic synapse.

Interestingly enough, the enriched pathways had been previously described in literature as being diseases-related. Different experiments have described different pathways [14–17] and the network-based enrichment method retrieved them all from the inclusion of the connecting nodes in the annotation modules.

In Figure 1 the difference in annotation between the standard enrichment procedure and the network-based is shown. As explained in the Methods section, standard enrichment is computed over a totally disconnected reference set. The network-based procedure relies on the precomputed annotation modules and the reference set includes all the nodes that connect seeds with the same annotation. This may increase the p-value as in

Enriched Terms - Directed Acyclic Graph



Standard enrichment

[+/-] Show/Hide all results.

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description
S	hsa04728	2 [+] Show genes	135 Show genes Show protein info	3.722e-03	Dopaminergic synapse
S	hsa04080	2 [+] Show genes	291 Show genes Show protein info	1.736e-02	Neuroactive ligand-receptor interaction

Network-based enrichment

N** highlights enriched terms not included in the annotations of the input set.

[+/-] Show/Hide all results.

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description	
N**	hsa04713	2 [+] Show genes	192 Show genes Show protein info	1.230e-02	Circadian entrainment	graph visualization
N**	hsa05032	2 [+] Show genes	202 Show genes Show protein info	1.362e-02	Morphine addiction	graph visualization
N**	hsa04723	2 [+] Show genes	220 Show genes Show protein info	1.616e-02	Retrograde endocannabinoid signaling	graph visualization
N**	hsa04724	2 [+] Show genes	239 Show genes Show protein info	1.908e-02	Glutamatergic synapse	graph visualization
N**	--------	2 [+] Show genes	271 Show genes Show protein info	2.454e-02	Environmental adaptation	graph visualization
N	hsa04728	2 [+] Show genes	296 Show genes Show protein info	2.928e-02	Dopaminergic synapse	graph visualization

Enriched terms and link to the database entry

Mapped input genes/proteins

Annotated genes

Possibility to visualize the term-specific network

Legend:
→ New terms added by NET-GE and not directly associated to the input proteins/genes

Figure 1: Output of NET-GE for the enrichment of KEGG pathways in the ADHD study case. Enrichment analysis was carried out using as input the *DRD4* and *DRD5* genes. The upper panel shows the graph of the enriched terms and their relations. Box filling color represents the corrected p-value associated to the enriched term, while contour color represents its information content (see [11] and [12] for details). The lower panel presents the enriched terms in a tabular format. Terms highlighted with a double star are new annotations, not associated to the input proteins and enriched with the network-based procedure. p-values are corrected with the Bonferroni procedure.

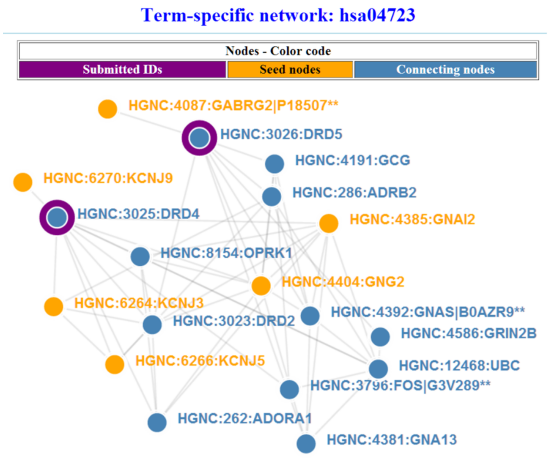


Figure 2: Graph of the first protein neighbours in the ADHD study case. The graph of the KEGG term hsa04723 (Retrograde endocannabinoid signaling) shows the two input proteins (in purple) and the first protein neighbours highlighted as seeds (in yellow) and new connecting genes (in blue). The connecting genes are added to the graph with the network-based enrichment procedure.

the case of the neuroactive ligand receptor interaction that is no longer listed among the terms obtained with the network-based procedure.

For comparison, we also tried PINA and EnrichNET. Considering as significant p-values<0.05 Benjamini-Hochberg corrected, PINA (tool "Identify enriched Interactome modules") did not retrieve any significantly over-represented module. EnrichNET authors recommend to analyse sets with at least 10 genes/proteins for reasons of statistical reliability. As a consequence, EnrichNET did not retrieve any significant term.

As evaluate the robustness of the method for small input sets composed of two to ten proteins, we computed the effect on the final stability of the enrichment when doubling (with random additions) the sizes of the input sets. We obtain that under these extreme conditions of noise, the stability of the enrichment ranges from 37 to 52%, depending on the annotation term and the network type (see Figure S1).

In Figure 2, the two input proteins are shown in the graph (purple circles) of the first protein neighbors, after network-based enrichment, detailing protein seeds of the Retrograde endocannabinoid signaling KEGG path (hsa04723, *Homo sapiens*) in yellow and the connecting nodes in blues (proteins that are retained after NET-GE based enrichment). The whole annotation network is downloadable (all seeds, nodes and arcs) and it is available for display.

Obsessive-compulsive Disorder

Obsessive-compulsive disorder (OCD) is a severe neuropsychiatric disorder characterized by the presence of obsessions and compulsions [18]. This disorder has been recently investigated in [18] by using whole-exome sequencing (WES).

Twenty OCD cases and their unaffected parents (parent–child trios) were screened for *de novo* missense mutations (i.e. mutation present only in the affected individual), identifying 27 OCD-related genes. Based on Ingenuity software (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis), three signaling pathways were identified as disease-related [18] sharing only one patient/one gene. In fact, among the 27 genes, only SMAD4 (gene mutated in only one patient) was present in the three enriched pathways.

With NET-GE we highlight four biological processes as the most significant ones (Figure 3, panel A), all related to the purine metabolism that has been proven to be associated with several neurological disorders [19–21]. However, and interestingly enough, 10 of the 27 initial genes have common annotations. Testing Molecular Functions, the standard enrichment procedure highlighted ATPase activity and the network-based procedure enriched thyroxine 5'-deiodinase activity (Figure 3, panel B), a new term not associated to the input proteins and involved in OCD [22].

Our results highlight the involvement of processes common to the gene panel and corroborates the notion that network-based enrichment consistently derives information from the connected annotation modules, including genes corresponding to 9 of the 14 patients analyzed in [18].

CONCLUSION

In this article, we presented the NET-GE web server [12], developed for tackling the problem of the human biological complexity. Specifically, NET-GE is a tool for associating biological processes and pathways with sets of human genes/proteins involved in the same phenotype. It performs standard and network-based enrichment analysis. The network-based procedure extracts from the STRING human interactome sub-networks of connecting proteins that share the same annotation [11]. We benchmarked NET-GE on two specific test cases, with a phenotype and its biological functions already described in literature. On this benchmark, the network-based procedure, considering genes/proteins in the context of their functional interaction network, enriched functional annotations that are experimentally validated. This version of NET-GE is preliminary to the inclusion of some additional features that can eventually add to the relevance of detecting emerging functional characteristics from a set of genes, such as the inclusion of ranking scores (e.g. fold of differentially expressed genes) or the usage of tissue-specific interactomes.

ACKNOWLEDGEMENTS

RC thanks COST Action BM1405 (European Union RTD Framework Program) and FARB UNIBO.

A) Biological Process

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description
N	GO:0009203	10 [+] Show genes	1506 Show genes Show protein info	1.505e-02	ribonucleoside triphosphate catabolic process
N	GO:0009207	10 [+] Show genes	1506 Show genes Show protein info	1.505e-02	purine ribonucleoside triphosphate catabolic process
N	GO:0009146	10 [+] Show genes	1510 Show genes Show protein info	1.541e-02	purine nucleoside triphosphate catabolic process
N	GO:0009143	10 [+] Show genes	1518 Show genes Show protein info	1.616e-02	nucleoside triphosphate catabolic process

B) Molecular Function

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description
N**	GO:0004800	2 [+] Show genes	8 Show genes Show protein info	8.505e-03	thyroxine 5'-deiodinase activity
N	GO:0017111	9 [+] Show genes	1786 Show genes Show protein info	1.146e-02	nucleoside-triphosphatase activity
N	GO:0016462	9 [+] Show genes	1919 Show genes Show protein info	2.033e-02	pyrophosphatase activity
N	GO:0016818	9 [+] Show genes	1924 Show genes Show protein info	2.075e-02	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
N	GO:0016817	9 [+] Show genes	1931 Show genes Show protein info	2.136e-02	hydrolase activity, acting on acid anhydrides

Figure 3: Output of NET-GE for Obsessive Compulsive Disorder for Biological Processes (panel A) and Molecular Function (panel B). Genes are derived from [18]. Terms highlighted with a double star are new annotations, not associated to the input proteins and enriched with the network-based procedure. p-values are corrected with the Bonferroni procedure.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution figure files, together with supplementary items listed below, are available at Genomics and Computational Biology online.

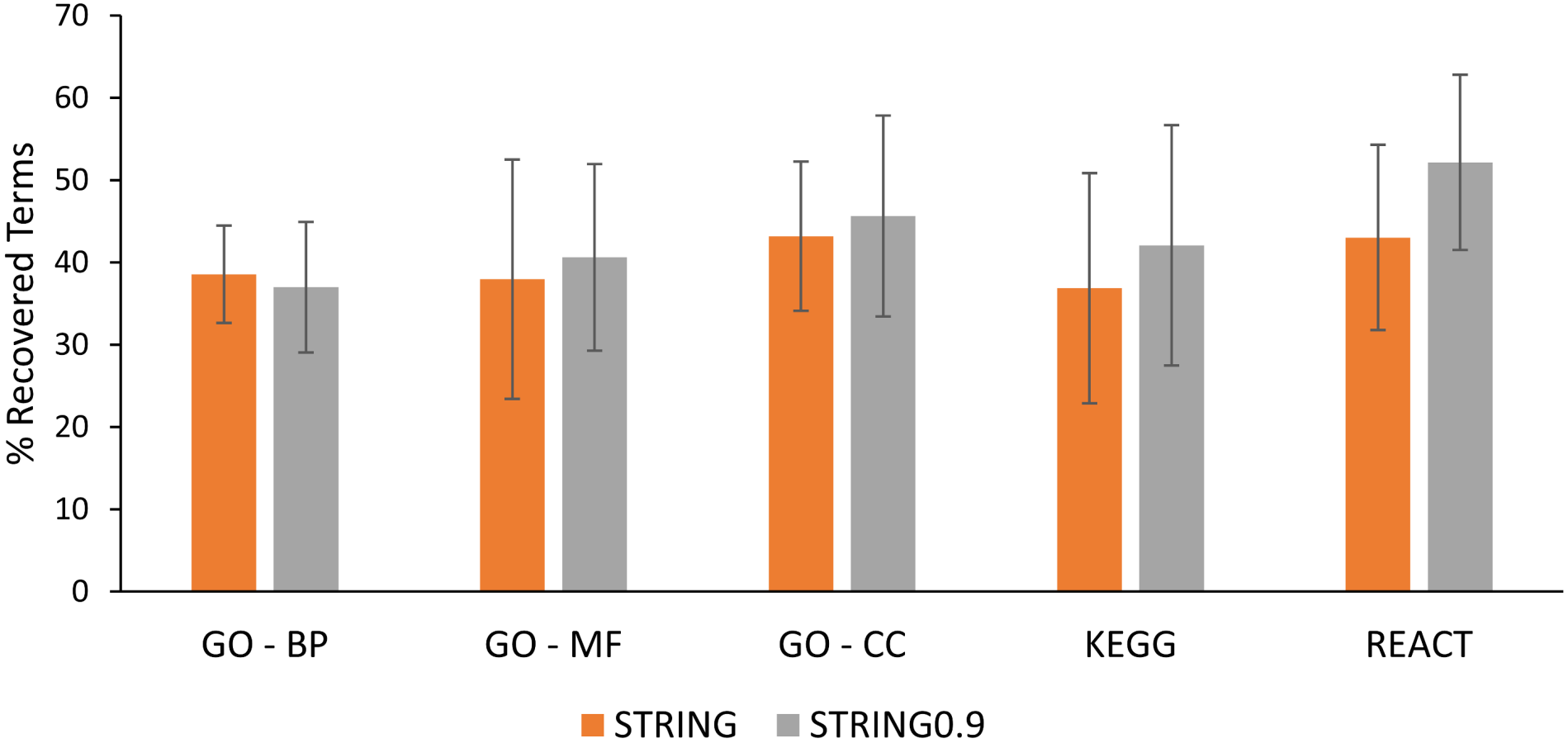
Supplementary Figure S1. **Testing the robustness of the network-based enrichment methods**. For small input sets comprising from two to ten proteins (derived from OMIM), we computed the effect of doubling (with random additions) the size on the final stability of the enrichment. This was done for all the annotation terms and the two different version of STRING (see Methods). Errors bars indicated standard deviations over a reference of 123 gene sets.

REFERENCES

1. Sander C. **Genomic Medicine and the Future of Health Care**. Science. 2000;287(5460):1977–1978. doi:10.1126/science.287.5460.1977.
2. The Gene Ontology Consortium. **Gene Ontology Consortium: going forward**. Nucleic Acids Research. 2015;43(D1):D1049–D1056. doi:10.1093/nar/gku1179.
3. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. **KEGG as a reference resource for gene and protein annotation.** Nucleic Acids Research. 2016;44(D1):D457–462. doi:10.1093/nar/gkv1070.
4. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. **The Reactome pathway Knowledgebase.** Nucleic Acids Research. 2016;44(D1):D481–487. doi:10.1093/nar/gkv1351.
5. Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. **Impact of outdated gene annotations on pathway enrichment analysis.** Nature Methods. 2016;13(9):705–706. doi:10.1038/nmeth.3963.
6. Laukens K, Naulaerts S, Berghe WV. **Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis.** Proteomics. 2015;15(5-6):981–996. doi:10.1002/pmic.201400296.
7. Mooney MA, Wilmot B. **Gene set analysis: A step-by-step guide.** American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics. 2015;168(7):517–527. doi:10.1002/ajmg.b.32328.
8. Huang DW, Sherman BT, Lempicki RA. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** Nucleic Acids Research. 2009;37(1):1–13. doi:10.1093/nar/gkn923.
9. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. **EnrichNet: network-based gene set enrichment analysis.** Bioinformatics. 2012;28(18):i451–i457. doi:10.1093/bioinformatics/bts389.
10. Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, et al. **PINA v2.0: mining interactome modules.** Nucleic Acids Research. 2012;40(D1):D862–865. doi:10.1093/nar/gkr967.
11. Di Lena P, Martelli PL, Fariselli P, Casadio R. **NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases.** BMC Genomics. 2015;16(Suppl 8):S6. doi:10.1186/1471-2164-16-S8-S6.
12. Bovo S, Di Lena P, Martelli PL, Fariselli P, Casadio R. **NET-GE: a web-server for NETwork-based human gene enrichment.** Bioinformatics. 2016;32(22):3489–3491. doi:10.1093/bioinformatics/btw508.
13. Noble WS. **How does multiple testing correction work?** Nature Biotechnology. 2009;27(12):1135–1137. doi:10.1038/nbt1209-1135.
14. Maltezos S, Horder J, Coghlan S, Skirrow C, O'Gorman R, Lavender TJ, et al. **Glutamate/glutamine and neuronal integrity in adults with ADHD: a proton MRS study.** Translational Psychiatry. 2014;4:e373. doi:10.1038/tp.2014.11.
15. Centonze D, Bari M, Di Michele B, Rossi S, Gasperi V, Pasini A, et al. **Altered anandamide degradation in attention-deficit/hyperactivity disorder.** Neurology. 2009;72(17):1526–1527. doi:10.1212/WNL.0b013e3181a2e8f6.
16. Gamble KL, May RS, Besing RC, Tankersly AP, Fargason RE. **Delayed sleep timing and symptoms in adults with attention-deficit/hyperactivity disorder: a controlled actigraphy study.** Chronobiology International. 2013;30(4):598–606. doi:10.3109/07420528.2012.754454.
17. Zhu J, Reith M. **Role of the Dopamine Transporter in the Action of Psychostimulants, Nicotine, and Other Drugs of Abuse**. CNS & Neurological Disorders - Drug Targets. 2008;7(5):393–409. doi:10.2174/187152708786927877.
18. Cappi C, Brentani H, Lima L, Sanders SJ, Zai G, Diniz BJ, et al. **Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways.** Translational Psychiatry. 2016;6:e764. doi:10.1038/tp.2016.30.
19. Micheli V, Camici M, Tozzi MG, Ipata PL, Sestini S, Bertelli M, et al. **Neurological Disorders of Purine and Pyrimidine Metabolism**. Current Topics in Medicinal Chemistry. 2011;11(8):923–947. doi:10.2174/156802611795347645.
20. Moretti A, Gorini A, Villa RF. **Affective disorders, antidepressant drugs and brain metabolism.** Molecular Psychiatry. 2003;8(9):773–785. doi:10.1038/sj.mp.4001353.
21. Hines DJ, Haydon PG. **Astrocytic adenosine: from synapses to psychiatric disorders.** Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1654):20130594. doi:10.1098/rstb.2013.0594.
22. Mermi O, Atmaca M. **Thyroid gland functions are affected in obsessive-compulsive disorder**. Anatolian Journal of Psychiatry. 2016;17(2):99–103. doi:10.5455/apd.178087.

Supplementary file:

45_Bovo_FigureS1.tif


70
60
50
40
30
20
10
0
% Recovered Terms
GO - BP
GO - MF
GO - CC
KEGG
REACT
STRING
STRING0.9