Zhang et al. BMC Bioinformatics (2020) 21:106
https://doi.org/10.1186/s12859-020-3408-y

BMC Bioinformatics

SOFTWARE Open Access

AllEnricher: a comprehensive gene set function enrichment tool for both model and non-model species

Du Zhang[1,2], Qi Hu[2], Xinxing Liu[1], Kai Zou[1], Emmanuel Konadu Sarkodie[1], Xueduan Liu[1*] and Fei Gao[2,3*]

Abstract

Background: Function genomic studies will generally result in lists of genes that may provide clues for exploring biological questions and discovering unanticipated functions, based on differential gene expression analysis, differential epigenomic analysis or co-expression network analysis. While tools have been developed to identify biological functions that are enriched in the genes sets, there remains a need for comprehensive tools that identify functional enrichment of genes for both model and non-model species from a different function classification perspective.

Results: We developed AllEnricher, a tool that calculates gene set function enrichment, with user-defined updatable libraries backing up for both model and non-model species as well as providing comprehensive functional interpretation from multiple dimensions, including GO, KEGG, Reactome, DO and DisGeNET.

Conclusions: AllEnricher incorporates up to date information from different public resources and provides a comprehensive resolution for biologists to make sense out of specific gene sets, making it an advanced open-source tool for gene set function analysis.

Keywords: Enrichment analysis, Function analysis, Pathway analysis, GO, KEGG, DO, Non-model species

Background

Functional genomics and large-scale genetic studies continuously generate a large number of gene sets (e.g. differentially expressed gene sets, co-expressed gene sets, or differential epigenomic modification gene sets, etc.). These gene sets are pivotal for elucidating molecular mechanisms in a biological system [1]. Investigating the relationship among these genes in the context of different function classification system provides clues for exploring biological questions and discovering unanticipated functions. Therefore, it is critical to characterizing gene-function relationships and mining gene-function associations of the gene sets.

Various kinds of databases have been developed for gene function classification. The most commonly used gene function database is the Gene Ontology (GO) [2]. Pathway-based database, like Kyoto Encyclopedia of Genes and Genomes (KEGG) [3] and Reactome [4], provide gene function interpretation through the perspective of biological reactions. Other databases like disease-based databases, such as Disease Ontology (DO) [5], DISEASE [6] and DisGeNET [7] were designed for molecular studies in disease. All these databases together provide comprehensive gene-function interpretations for the biologists.

Nonetheless, several analytic approaches based on different gene-function databases have been developed to decipher the biological significance of specific gene sets. Although proposed as the first generation of methods, Over-Representation Analysis (ORA) approaches still remain a commonly used method in exploring the functions implication of gene sets [1]. Based on this algorithm, many enrichment tools have been published, including GO-TermFinder [8], GOstat [9], WEGO [10], FunSet [11] for GO enrichment, KOBAS [12], clusterProfilers [13] for GO and KEGG enrichment, and DOSE

* Correspondence: xueduanliu@csu.edu.com; flys828@gmail.com
[1]School of Minerals Processing and Bioengineering, Central South University, Changsha 410083, China
[2]NEOMICS Institute, Shenzhen 518122, China
Full list of author information is available at the end of the article

[14] for disease enrichment. Though these tools can automatically calculate and visualize the significantly enriched function categories, various gene function analysis based on different tools and platforms make it complicated and tedious for biologists to choose and use. Therefore, collaborative tools like GO-Elite [15], MSigDB [16] and Enrichr [17] were developed to resolve these limitations. However, these tools either merely provide analysis for finite model species or the library they relied on are vulnerable to be out of date since their update depends on the timely maintenance of the author.

In this study, we developed a user-defined updatable application, which could be easily integrated into pipelines of functional genomic studies (RNA-seq, ATAC, BS-seq, etc.) and also can be incorporated into the five optimized public gene-function annotation collections (GO, KEGG, Reactome, DISEASE, and DisGeNET). Users of the application can update their local library to the latest version anytime they wish and decipher specific gene sets of both model and non-model species from appropriate gene function perspectives by enrichment analysis in just one single command.

Implementation

Public resources selection

The design framework of this tool is shown in Fig. 1. To establish local libraries as back up for AllEnricher, we firstly selected a series of public resources. The optimized public database must be timely updated and should incorporate gene-function annotations for both model and non-model species. We finally integrated five public resources into the local library, including GO, KEGG, Reactome, DISEASE, and DisGeNET (Fig. 2a).

Local gene ontology (GO) library construction

To obtain the GO annotation information of multiple species, we downloaded the public resource from NCBI and Gene Ontology Annotation (GOA). NCBI FTP (ftp://ftp.ncbi.nlm.nih.gov/) supplied the up to date GO annotation file, Gene Ontology (http://geneontology.org/) provided the obo file. The gaf file from GOA (https://www.ebi.ac.uk/GOA) provided all the GO annotations to proteins in the UniProt Knowledgebase (UniProtKB). The comprehensive gene information file supplied the corresponding relations between the NCBI official gene symbol and its gene ID (the unique identifier for a gene). All these files together make up the local GO library support for AllEnricher. We supplied a shell script to download and update this local database (*update_GOdb*) as per the user's requirement. Libraries for specified species can be built based on the local established GO database by *makeDB.go.sh*.

Local disease ontology (DO) library construction

The difficulty of DO analysis is to obtain the disease-gene annotation information. Though it has been developed as a standardized ontology for human disease, the Disease Ontology (http://disease-ontology.org/) falls short of up to date disease-gene annotation for the users. In 2013, The publication of Disease Gene Annotation (DGA, http://dga.nubic.northwestern.edu) [18] database provided an integrated environment to facilitate the analysis of disease–gene associations and explore potential gene interactions shared among multiple diseases.

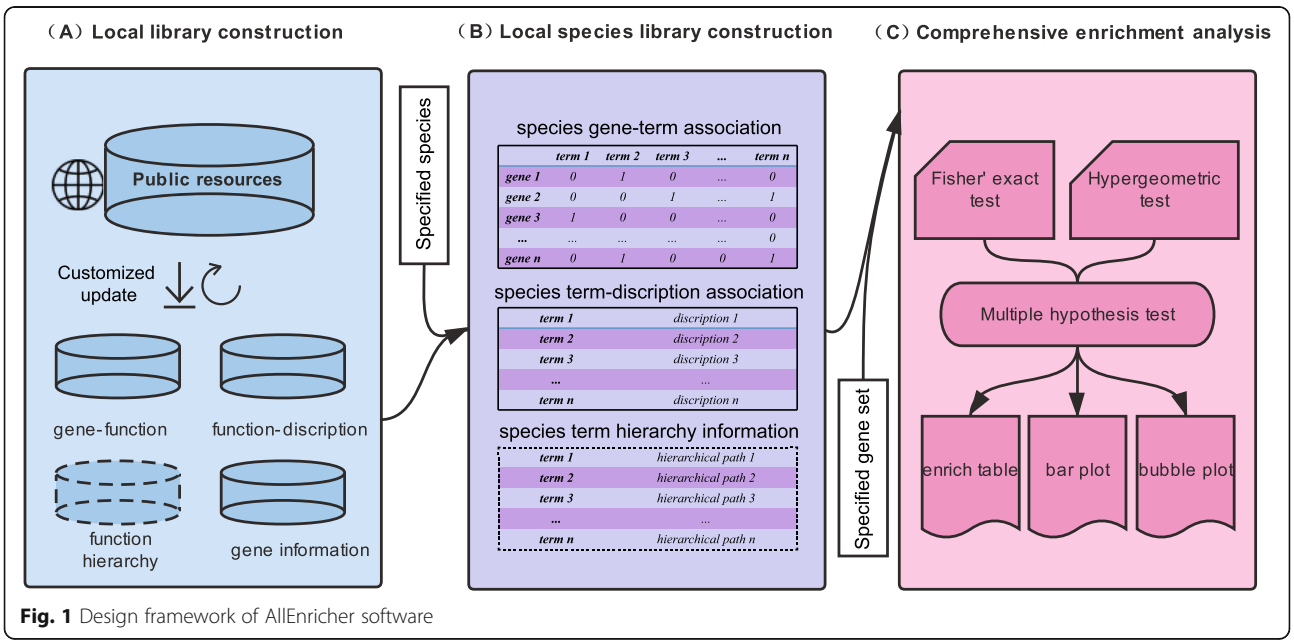

Fig. 1 Design framework of AllEnricher software

(a)

Local gene set library	Category	Support species	Source	Source update frequency
GO	Ontology	15464 species	ftp://ftp.ncbi.nlm.nih.gov/gene/ and ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/	Daily
KEGG	Pathway	498 Eukaryotes; 5179 Bacteria and 296 Archaea	https://www.kegg.jp/kegg/	Weekly
Reactome	Pathway	16 model species	https://reactome.org/	Quarterly
DISEASE	Disease; Ontology	human	https://diseases.jensenlab.org/	Weekly
DisGeNET	Disease	human	http://disgenet.org/	Annually

(b)

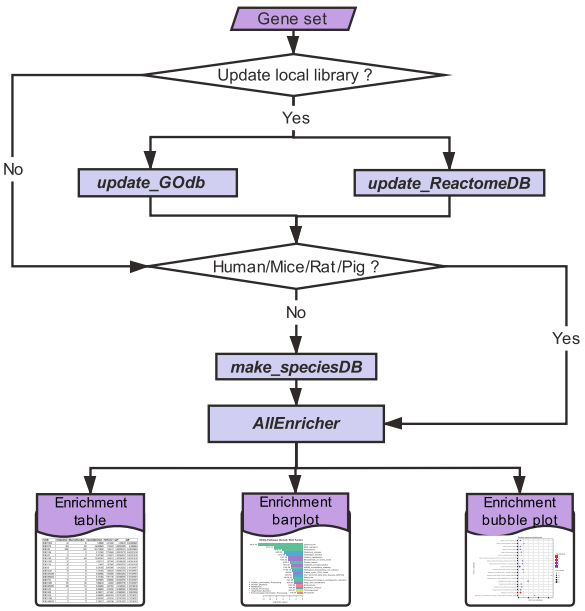

Fig. 2 Overview of the AllEnricher software current version. **a** Details of the local gene set libraries **b** General workflow of *AllEnricher*

However, currently, it has been out of service. DISEASES database (https://diseases.jensenlab.org/) [6] is a weekly updated web resource that integrates evidence on human disease-gene associations from automatic text mining, manually curated literature, cancer mutation data, and genome-wide association studies. We, therefore, utilized the non-redundant gene-disease annotation from the text mining channel, knowledge channel and experiment channel and integrated it as the local DO resource. The update of the local DO database was integrated into the building of AllEnricher comprehensive database for human (*makeDB.do.v1.0.sh*).

Local DisGeNET disease library construction

DisGeNET (http://disgenet.org/home/), which is another public gene-disease association database, is a discovery platform containing one of the largest publicly available collections of genes and variants associated with human diseases [7]. DisGeNET integrates data from expert curated repositories, GWAS catalogs, animal models and the scientific literature. We acquired all the gene-disease associations in DisGeNET and constructed the local DisGeNET disease library for AllEnricher. Local DISEASE and DisGeNET library only provide gene-disease annotations for humans, hence we merged and updated the progress into local library construction for human beings.

Local KEGG pathway library construction

KEGG PATHWAY Database (https://www.kegg.jp/kegg/pathway.html) is a collection of manually drawn pathway maps representing current knowledge on the molecular interaction. It is a commercial subscription-based database that does not offer a pathway-gene annotation file for specified species, and for this reason, we wrote several R scripts to download pathway-gene associations for specified species. The functional hierarchies for specified species could also be obtained from KEGG BRITE Database. The establishment of the local KEGG library for specified species was integrated into the shell script *makeDB.kegg.v1.0.sh*.

Local Reactome library construction

Reactome (https://reactome.org/) is a free, open-source and open-data pathway database which provides

comprehensive pathway knowledge to the biologist. We provide a script *update_ReactomeDB* to build and update the local Reactome database. First, we downloaded the pathway-gene annotation file of all the 16 kinds of species and then we filtered genes out of gene information files from NCBI. The local Reactome database for specified species was constructed based on these local resources by *makeDB.reactome.v1.0.sh*.

All the five gene-function libraries support for a specific species was built by script *make_speciesDB*. The accepted format for specifying genes is an official gene symbol from NCBI.

Gene set enrichment analysis and visualization

Fisher's exact test or hypergeometric test was employed to calculate the enrichment of the customized genes in the input gene set. The default genomic background gene sets for enrichment analysis were obtained from NCBI gene information. The False Discovery Rate (FDR) was controlled by multiple hypothesis testing with an alternative method of BH and FDR. The visualization of enrichment was accomplished via bar plots and bubble plots by R scripts. All the enrichment analysis and visualization steps had been integrated into the main program *AllEnricher*. The general workflow of AllEnricher is described in Fig. 2b.

Results

Command line application based on Unix

Shell, perl, and R were utilized to develop an easy to use command line application based on Unix environment to calculate gene enrichment of a user-provided gene set. Actually, the primary version of AllEnricher had been applied to several in house researches engaged in enrichment analysis of epigenetically modified gene sets and differentially expressed gene (DEG) sets of model species [19–21]. Here we used another two gene sets from previous studies to test AllEnricher.

Case study 1: function enrichment analysis in investigations of human disease

The first example is a DEG set based on RNA-Seq generated from ten matched pairs of cancer and non-cancerous tissues from Hepatocellular Carcinoma (HCC) patients [22]. AllEnricher (Fisher's Exact test, q-value < 0.05 by BH method) run on the 1378 DEGs identified in our study proved enrichment for cancer and tumorigenesis genes in both GO function and KEGG pathway analyses (Additional files 1 and 2), which are consistent with the main findings in this study. Moreover, DisGeNET disease enrichment analyses also reveal significant enrichment of genes associated with several kinds of carcinomas, with liver carcinoma emerging as the most enriched disease. These results validate the reliabiliy of AllEnricher and highlight specifically expressed genes in HCC that may confer important clues for understanding the molecular mechanisms of HCC pathogenesis.

Case study 2: function enrichment analysis in studies refers to non-model species

To illustrate how AllEnricher performs in gene set enrichment analysis on non-model species, we analyzed the RNA-seq data of golden snub-nosed monkeys (*Rhinopithecus roxellana*) living in the wild during both the winter and summer season [23]. We identified 2967 genes differentially expressed in different seasons using a well-established protocol [24]. To decipher the functional implications of this DEG list, *make_sepciesDB was used* to construct the local KEGG library for this species. As a result, 8327 genes annotated to 329 pathways constituted the unique local KEGG pathway library for the golden snub-nosed monkeys. KEGG pathway enrichment analysis by AllEnricher was applied to the seasonal DEGs (Hypergeometric test, q-value < 0.05 by BH method). Intriguingly, a wide range of associated physiological and metabolic pathways was enriched, including thermogenesis, oxidative phosphorylation, and pentose phosphate pathway (Additional files 3 and 4). The results obtained for seasonal stress and corresponding physiological response of golden snub-nosed monkeys in winter, provides a reasonable mechanism for the explaination of the adaption to the cold winter when food was scarce in the wild.

Comparison to other similar state-of-the-art tools

In order to investigate the similarities that exist between the AllEnricher and other similar tools, AllEnricher was compared to Enrichr, GO-Elite, clusterProfilers, and FunSet, which are the four leading tools that provide gene set enrichment analysis from comprehensive perspectives. Below is a summary of some important features of all the four tools, which refer to support species and libraries, library update, availability, ie. pipeline embeddable or customized background list (Table 1).

(1) Comprehensive function interpretation support: The same gene set provided by users could be interpreted from multiple aspects according to their purpose, which including Gene Ontology, KEGG pathway, Reactome pathway, Disease Ontology, and DisGeNET disease. Although the coverage of library collections is less than Enrichr and GO-Elite, gene-function annotations based on various kinds of database are planned to integrate as local libraries of AllEnricher to satisfy requirements of researches in a different field in the future, based on current program framework.

(2) Model species and non-model species support: The 1.0 version of AllEnricher already provides established

Table 1 Comparison to other similar state-of-the-art tools

Name	Support species	Support Libraries	Library update	Availability	Pipeline embeddable	Customized background list
AllEnricher	Model species and non-model species	GO, KEGG, Reactome, DISEASE, and DisGeNET	Customized	Standalone (https://github.com/zd105/AllEnricher)	Yes	Yes
Enrichr	Human, mouse and rat	GO, KEGG, GEO, InterPro, WikiPathways, MGI, Chromosome Location, Genome Browser PWMs, TargetScan, Reactome, BioCarta, TRANSFAC and JASPAR PWMs, Epigenomics Roadmap, ENCODE, ChEA, PPI databases, NURSA, CORUM, LINCS L1000, DEPOD, HumanCyc, NCI-Nature, Panther, KEA, HPO, GeneSigDB, CMAP, OMIM, VirusMINT, Achilles, dbGaP, Human Gene Atlas, Mouse Gene Atlas, ESCAPE, GTEx, HMDB and HomoloGene	Developer dependent	Web (http://amp.pharm.mssm.edu/Enrichr/)	No	No
GO-Elite	Over 60 species	GO, KEGG, GEO, InterPro, WikiPathways, MGI, Disease Ontology, GOSlim, Amadeus Metazoan compendium, PAZAR and AltAnalyze	Developer dependent	Standalone, Web (http://www.genmapp.org/go_elite/)	Yes	No
clusterProfilers	19 species	GO and KEGG	Developer dependent	Standalone (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html)	Yes	No
FunSet	11 species	GO	Developer dependent	Standalone, Web (http://funset.uno/)	Yes	Yes

local libraries for four kinds of most commonly studied model species, including humans (*Homo sapiens*), mouse (*Mus musculus*), Rat (*Rattus norvegicus*) and Pig (*Sus scrofa*). The number of species that supports the application is largely dependent on the public resources which the local library of AllEnricher is based on. Users could extend the supports for specific species by constructing the corresponding local species libraries. AllEnricher current version supports 15,464 species for GO enrichment, 5973 species for KEGG enrichment and merely human for disease enrichment (Fig. 1a), far more species than the other four tools. Supporting the analysis of various non-model species is the typical feature of AllEnricher.

(3) Customized library updates: The local library of AllEnricher was built based on frequently updated public resources (Fig. 2a). Compared to the other four similar tools, which library updates depend on the developers, several simple commands for customized library updates were designed. Therefore, users could obtain the latest data as they need.

(4) Customized background gene list: Enrichment analysis requires a background gene list. In general, researchers would take all the genes from the genome of specific species as the background gene list. However, the background gene list should be merely part of the genes from the whole genome in some cases. For example, when a DEG set is generated from samples of a specified tissue, where some parts of genes never expressed due to the high differentiation of cells, they should be excluded from the background gene list of enrichment analysis. AllEnricher provides flexible solutions to satisfy the application scenarios of the user-defined background gene list.

Conclusions

This study has demonstrated that a command line application based on the general Unix environment provides a robust way to carry out gene function enrichment, with support for multiple species and comprehensive functional perspectives. AllEnricher incorporates up to date information from different public resources and provides a comprehensive tool for biologists to make sense of specific gene lists. In summary, the wide application scenarios of AllEnricher makes it an advanced tool for gene set function enrichment analysis.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3408-y.

Additional file 1. Result tables enrichment analysis in case study 1.

Additional file 2. Result figures enrichment analysis in case study 1.

Additional file 3. Result figures enrichment analysis in case study 2.

Additional file 4. Result figures enrichment analysis in case study 2.

Abbreviations

DEG: Differentially expressed gene; DO: Disease Ontology; FDR: False Discovery Rate; GO: Gene Ontology; HCC: Hepatocellular Carcinoma; KEGG: Kyoto Encyclopedia of Genes and Genomes; ORA: Over-Representation Analysis

Acknowledgements

Not applicable.

Availability and requirements

Project name: AllEnricher.
Project home page: https://github.com/zd105/AllEnricher
Operating system(s): Unix.
Programming language: R, perl, and shell.
Other requirements: R >= 3.2, Perl version >= 5.10.1.
License: MIT.
Any restrictions to use by non-academics: non-academic use of KEGG pathway analysis generally requires a commercial license (refers to https://www.kegg.jp/kegg/legal.html).

Authors' contributions

DZ designed and developed the tool. QH contributed to the improvement of the original design. DZ wrote the manuscript and tested the software. EKS revised the manuscript. XXL, KZ, XDL, and FG supervised the project. The author(s) read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

AllEnricher source code is freely available in the GitHub repository (https://github.com/zd105/AllEnricher). The computing code is licensed under MIT. The online readme is the most extensive and continually updated source of documentation for AllEnricher, covering installation and user guide.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Minerals Processing and Bioengineering, Central South University, Changsha 410083, China. [2]NEOMICS Institute, Shenzhen 518122, China. [3]Comparative Pediatrics and Nutrition, Department of Veterinary and Animal Sciences, University of Copenhagen, DK-1870 Frederiksberg C, Denmark.

Received: 1 August 2019 Accepted: 11 February 2020
Published online: 17 March 2020

References

1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
2. Consortium TGO. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–8.
3. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109–14.
4. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55.
5. Bello SM, Shimoyama M, Mitraka E, Laulederkind SJF, Smith CL, Eppig JT, Schriml LM. Disease Ontology: improving and unifying disease annotations across species. Dis Model Mech. 2018;11(3):dmm032839.
6. Pletscher-Frankild S, Palleja A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2015;74:83–9.
7. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, Garcia-Garcia J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833–9.
8. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO:: TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710–5.
9. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23(2):257–8.
10. Ye J, Zhang Y, Cui H, Liu J, Wu Y, Cheng Y, Xu H, Huang X, Li S, Zhou A, et al. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. Nucleic Acids Res. 2018;46(W1):W71–5.
11. Hale ML, Thapa I, Ghersi D. FunSet: an open-source software and web server for performing and displaying gene ontology enrichment analysis. BMC Bioinformatics. 2019;20(1):359.
12. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39(Web Server issue):W316–22.
13. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
14. Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31(4):608–9.
15. 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.
16. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
17. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7.
18. Peng K, Xu W, Zheng J, Huang K, Wang H, Tong J, Lin Z, Liu J, Cheng W, Fu D, et al. The disease and gene annotations (DGA): an annotation resource for human disease. Nucleic Acids Res. 2013;41(Database issue):D553–60.
19. Chen YL, Zhang Y, Wang J, Chen N, Fang W, Zhong J, Liu Y, Qin R, Yu X, Sun Z, et al. A 17 gene panel for non-small-cell lung cancer prognosis identified through integrative epigenomic-transcriptomic analyses of hypoxia-induced epithelial-mesenchymal transition. Mol Oncol. 2019.
20. Gao F, Niu Y, Sun YE, Lu H, Chen Y, Li S, Kang Y, Luo Y, Si C, Yu J, et al. De novo DNA methylation during monkey pre-implantation embryogenesis. Cell Res. 2017;27(4):526–39.
21. Hu Y, Hu L, Gong D, Lu H, Xuan Y, Wang R, Wu CD, Zhang K, Gao F, et al. Genome-wide DNA methylation analysis in jejunum of Sus scrofa with intrauterine growth restriction. Mol Gen Genomics. 2018;293(4):807–18.
22. Huang Q, Lin B, Liu H, Ma X, Mo F, Yu W, Li L, Li H, Tian T, Wu D, et al. RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. PLoS One. 2011;6(10):e26168.
23. Zhang D, Hu Q, Hu Y, Zhang Y, Zhang Y, Cui P, Zhou Y, Liu X, Jiang J, Yang L, et al. Epigenetic and transcriptional signatures of ex situ conserved golden snub-nosed monkeys (Rhinopithecus roxellana). Biol Conserv. 2019;237:175–84.
24. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc. 2012;7(3):562–78.

Publisher's Note