https://doi.org/10.1093/nar/gkac1066

GPSAdb: a comprehensive web resource for interactive exploration of genetic perturbation RNA-seq datasets

Shipeng Guo[1,2,*], Zhougeng Xu[3], Xiangjun Dong[1], Dongjie Hu[1], Yanshuang Jiang[1], Qunxian Wang[1], Jie Zhang[1], Qian Zhou[1], Shengchun Liu[2,*] and Weihong Song[1,4,5,*]

[1]Chongqing Key Laboratory of Translational Medical Research in Cognitive Development and Learning and Memory Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, National Clinical Research Center for Child Health and Disorders, China International Science and Technology Cooperation Base of Child Development and Critical Disorders, Children's Hospital of Chongqing Medical University, Chongqing, China, [2]Department of Breast and Thyroid Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [3]National Key Laboratory of Plant Molecular Genetics (NKLPMG), CAS Center for Excellence in Molecular Plant Sciences (CEMPS), Institute of Plant Physiology and Ecology (SIPPE), Chinese Academy of Sciences (CAS), Shanghai, China, [4]Institute of Aging, Key Laboratory of Alzheimer's Disease of Zhejiang Province, Zhejiang Provincial Clinical Research Center for Mental Disorders, School of Mental Health and Kangning Hospital, The Second Affiliated Hospital and Yuying Children's Hospital, Wenzhou Medical University, Wenzhou, Zhejiang, China and [5]Oujiang Laboratory (Zhejiang Lab for Regenerative Medicine, Vision and Brain Health), Wenzhou, Zhejiang, China

Received August 26, 2022; Revised October 21, 2022; Editorial Decision October 24, 2022; Accepted October 26, 2022

ABSTRACT

Gene knock-out/down methods are commonly used to explore the functions of genes of interest, but a database that systematically collects perturbed data is not available currently. Manual curation of all the available human cell line perturbed RNA-seq datasets enabled us to develop a comprehensive human perturbation database (GPSAdb, https://www.gpsadb.com/). The current version of GPSAdb collected 3048 RNA-seq datasets associated with 1458 genes, which were knocked out/down by siRNA, shRNA, CRISPR/Cas9, or CRISPRi. The database provides full exploration of these datasets and generated 6096 new perturbed gene sets (up and down separately). GPSAdb integrated the gene sets and developed an online tool, genetic perturbation similarity analysis (GPSA), to identify candidate causal perturbations from differential gene expression data. In summary, GPSAdb is a powerful platform that aims to assist life science researchers to easily access and analyze public perturbed data and explore differential gene expression data in depth.

INTRODUCTION

Gene knock-out/down is commonly used to explore the causality and mechanisms underlying cell behavior. Genetic perturbation experiments provide a powerful approach where cells are modulated, and the downstream consequences are monitored. RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression (DGE) and is frequently used to monitor downstream consequences where genes were perturbed. Genetic perturbed data sets are accumulating rapidly, but they are complicated for wet-lab research communities to use due to the lack of bioinformatical skills.

Several web tools exist that provide information regarding perturbed data, such as GPA (1), KnockTF (2), Enrichr (3) and CMAP (4). GPA mainly collects perturbed data from microarrays and is currently decommissioned. The KnockTF database collects RNA-seq and microarray datasets related to transcription factors (TFs), but non-TF genes are not included. The Enrichr database offers analysis based on a wide range of gene sets, including some perturbation gene sets from GEO (5). Enrichr only provides analysis with over-representation analysis method (ORA) and gene set enrichment analysis (GSEA) method is not integrated. However, the number of perturbation gene sets from Enrichr is limited and the origin of the gene sets are not clearly described. Finally, CMAP contains transcriptome

[*]To whom correspondence should be addressed. Tel: +86 23 63637851; Email: guoshipeng2008@126.com
Correspondence may also be addressed to Shengchun Liu. Email: liushengchun1968@163.com
Correspondence may also be addressed to Weihong Song. Email: weihong@wmu.edu.cn

expression profiles derived from many perturbagens, most of which are small molecules (4). To date, databases systematically collected genetic perturbed data sets, but easy access and exploration of the datasets are not available. Here, we constructed the GPSAdb genetic perturbation database, with the aim to provide intensive resources and easy exploration of perturbed data to the research community.

To achieve this goal, we collected 3048 RNA-seq datasets associated with 1458 genes which was knocked out/down using siRNA, shRNA, CRISPR/Cas9, or CRISPRi by comprehensive manual curation of all the available human cell line perturbed RNA-seq datasets from GEO, SRA, and ENCODE. Through DGE analysis of GPSAdb datasets, 6096 new gene sets (up and down) were generated. GPSAdb integrated the 6096 new gene sets and developed a genetic perturbation similarity analysis (GPSA) tool to identify candidate causal perturbations from differential gene expression data. With the GPSA tool, life science researchers can gain a better understanding of DGE data.

We constructed a couple of other derivative tools as well, for instance, Query (finds a gene when perturbation affects the expression of the gene of interest) and Mediator (finds a gene that theoretically mediates).The GPSAdb is now available at https://www.gpsadb.com/ or http://guotosky.vip:13838/GPSA/.

MATERIALS AND METHODS

Data sources and processing

To obtain candidate genetic perturbed RNA-seq datasets, we searched NCBI GEO with the following query syntax: ((((((((((knockdown) OR knock-down) OR knock down) OR knockout) OR knoc-kout) OR knock out) OR shRNA) OR siRNA)) OR ((((CRISPRi) OR CRISPR interference)) OR dCas9))) AND ((Homo sapiens[Organism]) AND 4:1000[Number of Samples]), and a total of 9297 datasets were collected. We then manually curated those datasets to keep only genetic perturbed RNA-seq data, which should contain at least two samples in each treatment and control group. Finally, we obtained 2337 datasets from SRA. Similar searching and curating methods were applied to the ENCODE database and 711 datasets were obtained. In total, GPSAdb contains 3048 genetic perturbed RNA-seq datasets. The curated RNA-seq data was processed by a snakemake-based pipeline (https://github.com/xuzhougeng/auto_sra_rnaseq_pipeline). Briefly, the SRA data downloaded from GEO were firstly converted to fastq and the low-quality reads were filtered by fastp (v0.23.1) (6). The high-quality reads were then aligned to human genome (GRCh38) by STAR (v2.6.1d) (7). STAR was also used to count number reads per gene while mapping using GTF annotation downloaded from ENSEMBL (release-95). Finally, all gene-count files were merged into a single file for downstream DGE analysis (Figure 1).

DGE analysis and perturbed gene set production

Batch DGE analysis was conducted among the 3048 datasets using R package Deseq2 (8), and 6096 new perturbed gene sets were generated (1000 up genes and 1000 down genes for each gene set). The 3048 results of those differential analysis were organized for gene set enrichment analysis (GSEA) with different gene sets (9).

GSEA

GSEA analysis was performed among the 3048 datasets using R package clusterProfiler (10). The input data for GSEA was a pre-ranked gene list, wherein all genes were ranked by their log fold change value. The gene sets chosen to perform GSEA were Hallmarks (11), KEGG (12), Wikipathway (13) and Reactome (14). The results of all DGE analysis and GSEA analysis were organized and integrated into the Query and Mediator tool.

GPSA

GPSA compared the DGE of the 3048 perturbed datasets from GPSAdb to find which genes shares similar or reverse downstream effects with the user's input data with perturbation. GPSAdb generated 6096 new gene sets from 3048 gene perturbation RNA-seq datasets. The gene sets were split into two groups: up gene sets and down gene sets. GSEA analysis was performed with the input data and those 6096 gene sets. We computed the similarities between input data and GPSAdb datasets using a method similar to weighted connectivity score (WTCS) (4).

RESULTS AND DISCUSSION

Database use and access

User-friendly interface for browsing and exploration of 3048 perturbed datasets. GPSAdb contains five functional modules to access and explore perturbed data, including browse, query, mediator, enrichGPSA, and GPSA (Figure 2A). The browse tool in GPSAdb provides easy exploration of all 3048 perturbed datasets (Figure 2B). Users can search and click rows in gpsaMetadata panel to pick a dataset of interest. Three different panels (diffTable, GSEA, GPSA) were developed to facilitate data investigation. The diffTable panel displays differential results in MAplot and volcano plot. Users can easily locate the downstream targets of the perturbed gene. GPSAdb also incorporates the TwoGenePlot function module from GTBAdb (https://www.gtbadb.com/, a database from our group) to support tissue level correlation analysis between perturbed genes and their targets. Users can go to GSEA panel to perform GSEA and ORA with preferred gene sets. If users click RUN GPSA button in GPSA tab panel, the chosen dataset will be prepared for GPSA analysis and GPSAdb web pages will automatically switch to GPSA tool panel. With the browse tool, researchers can explore all 3048 datasets in an easy and intuitive way.

Querying candidate causal perturbations of your gene/gene set of interest. Locating downstream targets of a gene with the browse tool is simple. However, if users want to find which gene perturbation regulates the gene of interest, for example, finding genes upstream of TP53, few tools provide functionality. Since we have collected 3048 gene knock down/out RNA-seq datasets and differential analysis of all those datasets was performed, integrating all results and

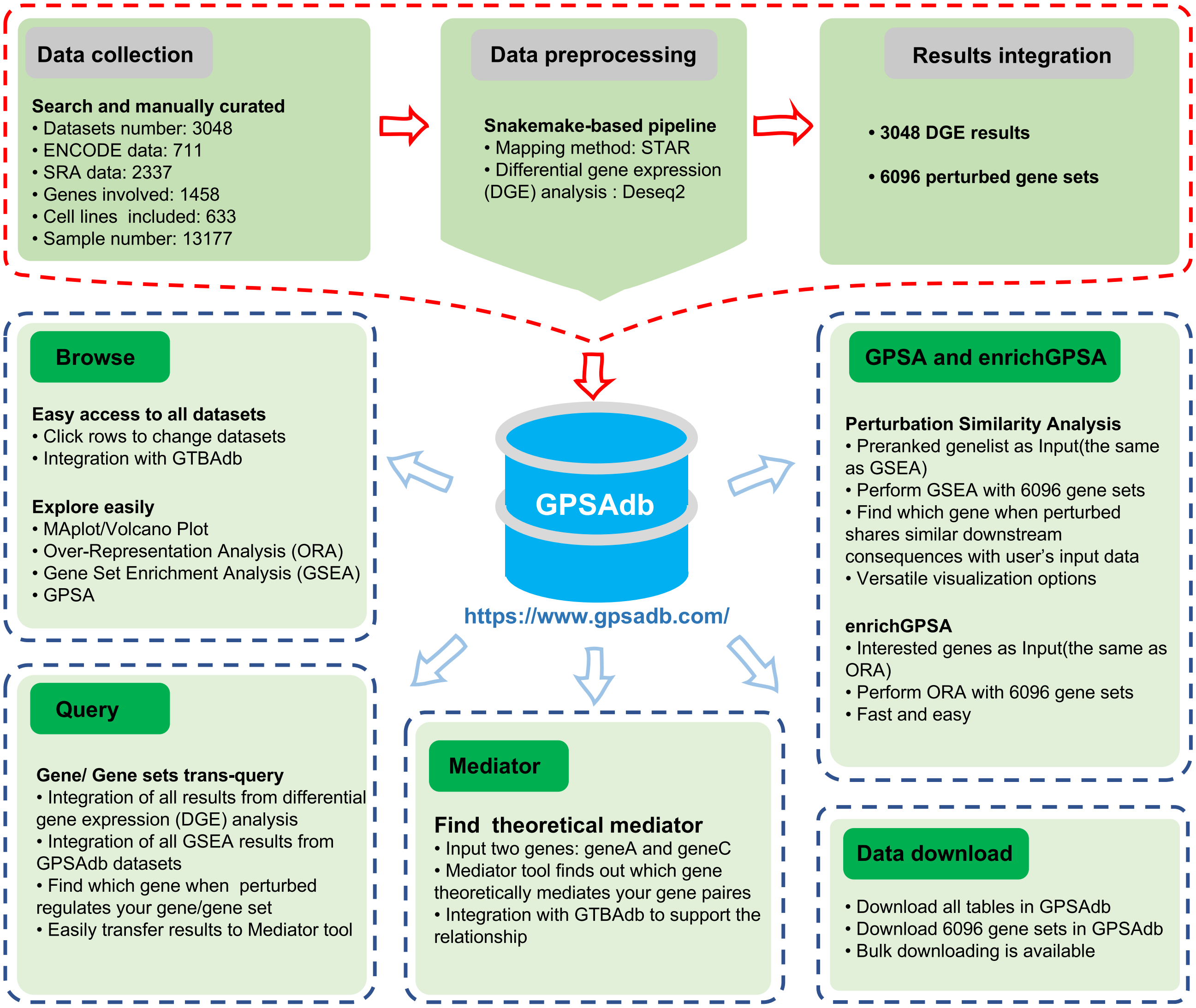
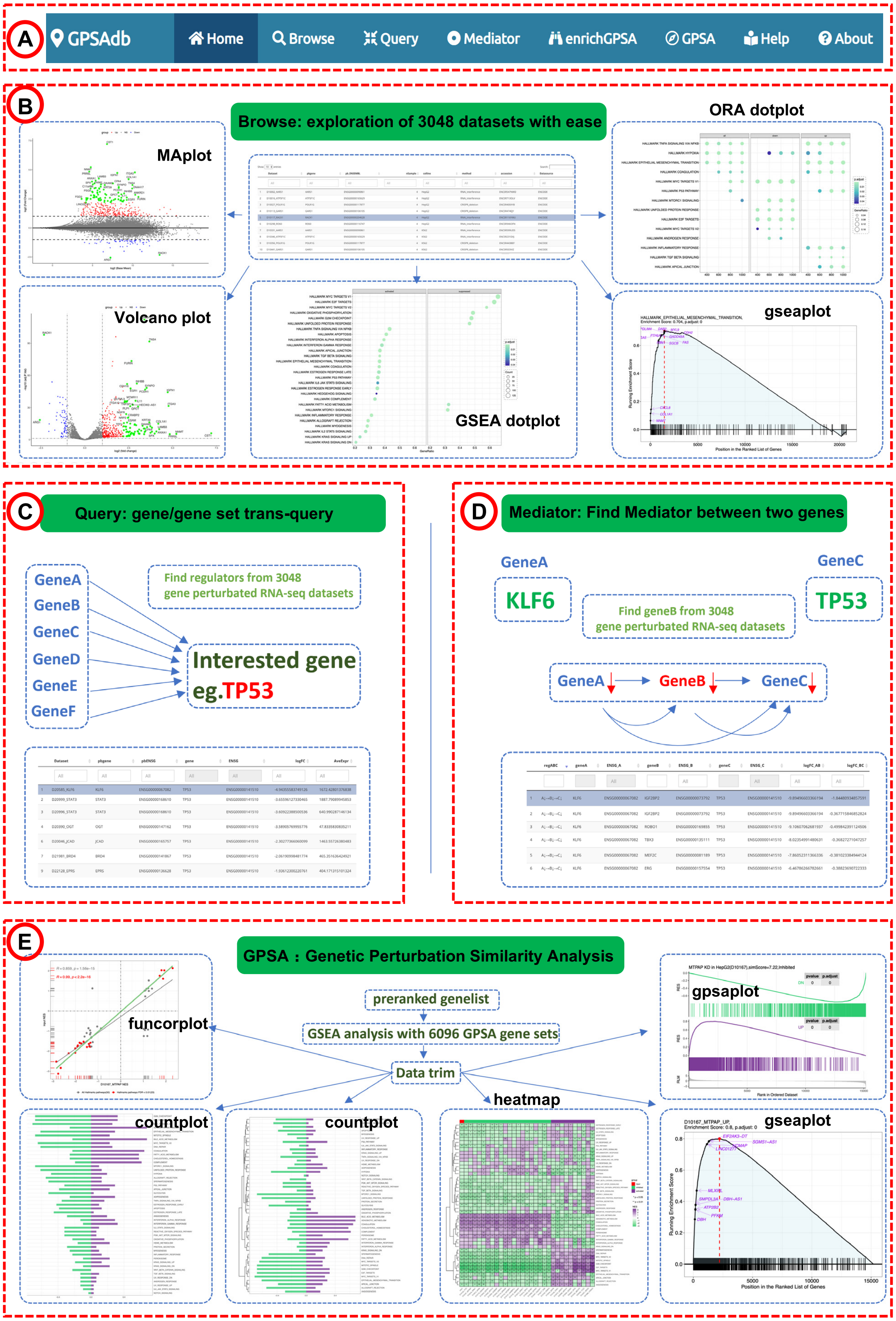

Data collection
Search and manually curated
• Datasets number: 3048
• ENCODE data: 711
• SRA data: 2337
• Genes involved: 1458
• Cell lines included: 633
• Sample number: 13177
Data preprocessing
Snakemake-based pipeline
• Mapping method: STAR
• Differential gene expression (DGE) analysis : Deseq2
Results integration
• 3048 DGE results
• 6096 perturbed gene sets
GPSAdb
https://www.gpsadb.com/
Browse
Easy access to all datasets
• Click rows to change datasets
• Integration with GTBAdb
Explore easily
• MAplot/Volcano Plot
• Over-Representation Analysis (ORA)
• Gene Set Enrichment Analysis (GSEA)
• GPSA
GPSA and enrichGPSA
Perturbation Similarity Analysis
• Preranked genelist as Input(the same as GSEA)
• Perform GSEA with 6096 gene sets
• Find which gene when perturbed shares similar downstream consequences with user's input data
• Versatile visualization options
enrichGPSA
• Interested genes as Input(the same as ORA)
• Perform ORA with 6096 gene sets
• Fast and easy
Query
Gene/ Gene sets trans-query
• Integration of all results from differential gene expression (DGE) analysis
• Integration of all GSEA results from GPSAdb datasets
• Find which gene when perturbed regulates your gene/gene set
• Easily transfer results to Mediator tool
Mediator
Find theoretical mediator
• Input two genes: geneA and geneC
• Mediator tool finds out which gene theoretically mediates your gene paires
• Integration with GTBAdb to support the relationship
Data download
• Download all tables in GPSAdb
• Download 6096 gene sets in GPSAdb
• Bulk downloading is available

Figure 1. Construction pipeline of the GPSAdb database. Genetic perturbed RNA-seq datasets (3048) were included in GPSAdb. All data underwent a uniform data processing pipeline and 3048 differential gene expression (DGE) results and 6096 new gene sets were generated. GPSAdb provides a user-friendly interface to browse and explore these data, along with several online tools such as Query, Mediator, enrichGPSA and GPSA.

providing users with a user-friendly query interface is possible. The query tool in GPSAdb locates the gene that regulates the gene of interest from 3048 genetic perturbed RNA-seq datasets (Figure 2C). All 3048 datasets in GPSAdb were subjected to GSEA with four sources of gene sets, and query of which gene regulates a specific gene set is also supported. We also developed a mediator tool to determine which gene theoretically mediates the gene pair (Figure 2D). However, it should be emphasized that the Mediator tool does not make predictions, and instead, it simply connects the existing regulation evidence from 3048 DGE results in GPSAdb.

Online tool for GPSA. Genetic perturbation methods, such as those that use siRNA, shRNA and CRISPR/Cas9, are important in scientific research. When referred to the mechanism of a specific gene or drug on cells, RNA-seq was generally performed after the cells were treated with a drug or when the gene was knocked down/out. The current downstream analysis of RNA-seq data is mainly dependent on pre-defined gene sets, such as KEGG, GO and MsigDB, among others. Newly defined high quality gene sets can broaden researchers' understanding of RNA-seq data (3). The GPSA tool uses 6096 newly defined gene sets from GPSAdb and aims to helps users identify which gene shares similar or inverse transcriptomic changes to user's input data (Figure 2E). The input data for GPSA is a pre-ranked gene list, which is the same as that required in GSEA. However, the GPSA input requires a table that contains two columns: gene symbols written in a form similar to 'TP53' and a variable that is used to rank those genes, such as logFC. GPSA automatically transformed the two-column table into a preferred pre-ranked gene list. Visualization of GPSA results was possible via a versatile set of tools, including funcorplot, countplot, heatmap, gpsaplot, and gseaplot (Figure 2E). With GPSA application, researchers can infer the molecular causes of the observed differences in their DGE data. We also developed enrichGPSA to perform ORA with the 6096 newly defined gene sets in case


A
GPSAdb
Home
Browse
Query
Mediator
enrichGPSA
GPSA
Help
About
B
Browse: exploration of 3048 datasets with ease
MAplot
Volcano plot
ORA dotplot
gseaplot
GSEA dotplot
C
Query: gene/gene set trans-query
GeneA
GeneB
GeneC
GeneD
GeneE
GeneF
Find regulators from 3048 gene perturbated RNA-seq datasets
Interested gene eg.TP53
D
Mediator: Find Mediator between two genes
GeneA
KLF6
GeneC
TP53
Find geneB from 3048 gene perturbated RNA-seq datasets
GeneA↓ → GeneB↓ → GeneC↓
E
GPSA：Genetic Perturbation Similarity Analysis
preranked genelist
GSEA analysis with 6096 GPSA gene sets
Data trim
funcorplot
gpsaplot
countplot
countplot
heatmap
gseaplot

	Dataset	pbgene	pbENSG	gene	ENSG	logFC	AveExpr
	All	All	All	All	All	All	All
1	D20585_KLF6	KLF6	ENSG00000067082	TP53	ENSG00000141510	-4.94355583749126	1672.42801376838
2	D20999_STAT3	STAT3	ENSG00000168610	TP53	ENSG00000141510	-3.65596127330465	1887.79089945853
3	D20996_STAT3	STAT3	ENSG00000168610	TP53	ENSG00000141510	-3.60922388500536	640.990287146134
4	D20390_OGT	OGT	ENSG00000147162	TP53	ENSG00000141510	-3.58905769955776	47.8335830835211
6	D20046_JCAD	JCAD	ENSG00000165757	TP53	ENSG00000141510	-2.30277366060099	1463.55726380483
7	D21981_BRD4	BRD4	ENSG00000141867	TP53	ENSG00000141510	-2.06190998481774	465.351636424921
9	D22128_EPRS	EPRS	ENSG00000136628	TP53	ENSG00000141510	-1.93612300220761	404.171315101324

	regABC	geneA	ENSG_A	geneB	ENSG_B	geneC	ENSG_C	logFC_AB	logFC_BC
			All		All		All	All	All
1	A↓→B↓→C↓	KLF6	ENSG00000067082	IGF2BP2	ENSG00000073792	TP53	ENSG00000141510	-9.89496603366194	-1.84480934857591
2	A↓→B↓→C↓	KLF6	ENSG00000067082	IGF2BP2	ENSG00000073792	TP53	ENSG00000141510	-9.89496603366194	-0.367715846852824
3	A↓→B↓→C↓	KLF6	ENSG00000067082	ROBO1	ENSG00000169855	TP53	ENSG00000141510	-9.10607062681937	-0.499842391124506
4	A↓→B↓→C↓	KLF6	ENSG00000067082	TBX3	ENSG00000135111	TP53	ENSG00000141510	-8.02354991480631	-0.36827271047257
5	A↓→B↓→C↓	KLF6	ENSG00000067082	MEF2C	ENSG00000081189	TP53	ENSG00000141510	-7.86052311366336	-0.381023384944124
6	A↓→B↓→C↓	KLF6	ENSG00000067082	ERG	ENSG00000157554	TP53	ENSG00000141510	-6.46786266782661	-0.38823690722333

Figure 2. Overview of the GPSAdb database workflow. (**A**) Five functional modules are provided in GPSAdb. (**B**) Data sets (3048) are browsed and explored with ease in multiple ways. (**C**) The perturbed gene that regulates the gene of interest or gene set of interest was queried. (**D**) A mediator was selected for the gene regulation pair of interest. (**E**) Gene perturbation similarity analysis (GPSA) module to select a gene when perturbances share similar downstream consequences with the input data.

users are interested with only a group of genes. The enrichGPSA analysis is easy and fast and the results indicates that in which up or down perturbation gene sets the interested genes are enriched.

System design and implementation. The GPSAdb runs on Ubuntu Linux (20.04 × 64) with 48GB memory and 24-core computed processing units (CPUs). GPSAdb was written in R (version 4.1.3, https://www.r-project.org/) based on the web framework R Shiny Server (version 1.5.17.973, https://rstudio.com/products/shiny/download-server/). The database utilizes several third-party tools, including Tidyverse (15), ggplot2 (Wickham, 2016), shinyWidgets (Perrier, 2022), shinydashboard (Chang, 2021), DT (Xie, 2021), aplot (Yu, 2021), STRINGdb (16), Deseq2 (17) and clusterProfiler (10). GPSAdb adopted a WYSIWYG (what you see is what you get) method to display and download plots. The plot is free to modify using the options we offered, following which the plot is downloaded in a WYSIWYG way. The plots are downloadable as three file types, namely, PDF, PNG, and SVG.

SUMMARY AND FUTURE DIRECTIONS

We have constructed an interactive, user-friendly database—GPSAdb—for users to browse, explore, visualize, and intuitively analyze 3048 genetic perturbed RNA-seq datasets. We also introduced a GPSA online tool to assist researchers exploring their own DGE data. With the application of GPSA, researchers can better infer the molecular causes of the observed difference in gene expression. Several derivative useful tools are also implemented in GPSAdb, including Query (looks up upstream regulators) and Mediator (finds out which gene theoretically mediates your gene pair). As perturbed datasets are growing rapidly, we will annually survey newly released perturbed data resources, update GPSAdb accordingly, and maintain it as a useful resource for the research community, bridging the gap between biology and bioinformatics.

DATA AVAILABILITY

The GPSAdb database is publicly available (https://www.gpsadb.com/ or http://guotosky.vip:13838/GPSA/). The RNA-seq data was processed by a snakemake-based pipeline (https://github.com/xuzhougeng/auto_sra_rnaseq_pipeline). All the plots and tables in GPSAdb are free to download. Download data in bulk is also supported in GPSAdb.

ACKNOWLEDGEMENTS

We gratefully acknowledge Dr Guangchuang Yu for exporting the functions of R package clusterProfiler and Prof. Daniel J. Rigden for the suggestions on database improvement.

FUNDING

Chongqing Science and Technology Commission [CSTC2021jscx-gksb-N0027]; Natural Science Foundation of Chongqing, China [2022NSCQ-BHX0551]; National Natural Science Foundation of China [81772979]; Key Laboratory of Alzheimer's Disease of Zhejiang Province and Zhejiang Provincial Clinical Research Center for Mental Disorders (to W.S., in part). Funding for open access charge: Chongqing Science and Technology Commission [CSTC2021jscx-gksb-N0027].

Conflict of interest statement. None declared.

REFERENCES

1. Xiao,Y., Gong,Y., Lv,Y., Lan,Y., Hu,J., Li,F., Xu,J., Bai,J., Deng,Y., Liu,L. *et al.* (2015) Gene perturbation atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Sci. Rep.*, **5**, 10889.
2. Feng,C., Song,C., Liu,Y., Qian,F., Gao,Y., Ning,Z., Wang,Q., Jiang,Y., Li,Y., Li,M. *et al.* (2019) KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Res.*, **48**, D93–D100.
3. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
4. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
5. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2010) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
6. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
7. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
8. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
9. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
10. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L. *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*, **2**, 100141.
11. Liberzon,A., Birger,C., Thorvaldsdóttir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
12. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
13. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,DeniseN., Hanspers,K., A. Miller,R., Digles,D., Lopes,ElissonN., Ehrhart,F *et al.* (2020) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
14. Gillespie,M., Jassal,B., Stephan,R., Milacic,M., Rothfels,K., Senff-Ribeiro,A., Griss,J., Sevilla,C., Matthews,L., Gong,C. *et al.* (2021) The reactome pathway knowledgebase 2022. *Nucleic Acids Res.*, **50**, D687–D692.
15. Wickham,H., Averick,M., Bryan,J., Chang,W., McGowan,L.D.A., François,R., Grolemund,G., Hayes,A., Henry,L., Hester,J. *et al.* (2019) Welcome to the tidyverse. *J. Open Source Software*, **4**, 1686.
16. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2020) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
17. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.