D1094–D1101 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 16 October 2022*
https://doi.org/10.1093/nar/gkac872

PertOrg 1.0: a comprehensive resource of multilevel alterations induced in model organisms by *in vivo* genetic perturbation

Zhaoyu Zhai†, Xuelu Zhang†, Lu Zhou†, Zhewei Lin, Ni Kuang, Qiang Li, Qinfeng Ma, Haodong Tao, Jieya Gao, Shiyong Ma and Jianbo Pan*

Center for Novel Target and Therapeutic Intervention, Institute of Life Sciences, Chongqing Medical University, Chongqing 400016, China

Received July 21, 2022; Revised September 04, 2022; Editorial Decision September 25, 2022; Accepted September 28, 2022

ABSTRACT

Genetically modified organisms (GMOs) can be generated to model human genetic disease or plant disease resistance, and they have contributed to the exploration and understanding of gene function, physiology, disease onset and drug target discovery. Here, PertOrg (http://www.inbirg.com/pertorg/) was introduced to provide multilevel alterations in GMOs. Raw data of 58 707 transcriptome profiles and associated information, such as phenotypic alterations, were collected and curated from studies involving *in vivo* genetic perturbation (e.g. knockdown, knockout and overexpression) in eight model organisms, including mouse, rat and zebrafish. The transcriptome profiles from before and after perturbation were organized into 10 116 comparison datasets, including 122 single-cell RNA-seq datasets. The raw data were checked and analysed using widely accepted and standardized pipelines to identify differentially expressed genes (DEGs) in perturbed organisms. As a result, 8 644 148 DEGs were identified and deposited as signatures of gene perturbations. Downstream functional enrichment analysis, cell type analysis and phenotypic alterations were also provided when available. Multiple search methods and analytical tools were created and implemented. Furthermore, case studies were presented to demonstrate how users can utilize the database. PertOrg 1.0 will be a valuable resource aiding in the exploration of gene functions, biological processes and disease models.

INTRODUCTION

Experimentation on humans for biomedical research is frequently unfeasible and/or unethical; therefore, model organisms are often studied using technologies that perturb normal function, such as mutagenesis, RNA interference and drug treatment (1). A genetically modified organism (GMO) is any organism whose genotype has been altered using genetic engineering techniques. GMOs can be generated to model human genetic disease or plant disease resistance and have been improving the exploration and understanding of gene function, physiology, disease onset and drug target discovery. GMOs can also be applied directly for food use (e.g. GM soybeans) or to make food additives (e.g. aspartame, an artificial sweetener) (2). Characterization of GMOs is needed for a better understanding and exploration of GMOs. To characterize GMOs, phenotypes can be qualitatively described, and gene expression profiles can be quantitatively measured. High-throughput sequencing technologies, e.g. microarray, RNA-seq and single-cell RNA-seq (scRNA-seq), allow a quantitative evaluation of gene expression and cell type composition to enable the comparison of model organisms before and after modification.

An increasing number of studies have reported high-throughput data of genetically modified model organisms, and the raw data generated are deposited in databases such as GEO and ArrayExpress (3,4). Some databases, such as KnockTF, LINCS and GPA (no longer accessible), were developed to deposit data on differentially expressed genes (DEGs) in human cell lines following genetic perturbations (5–7). However, 2D *in vitro* cell cultures do not fully recapitulate living tissue (e.g. immune cells are not present); therefore, they cannot mimic the natural structures of tissues *in vivo*. They are also unable to determine the organism's phenotype. A centralized data portal focusing on genetic perturbations *in vivo* for multiple model organisms could help better and more accurately understand the multilevel alter-

*To whom correspondence should be addressed. Tel: +86 23 684 80209; Fax: +86 23 684 80209; Email: panjianbo@cqmu.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

ations induced by genetic perturbations and further make full use of GMOs.

In this study, we introduced a user-friendly database, named Perturbing the Organisms (PertOrg) (http://www.inbirg.com/pertorg/), which provides a comprehensive resource of transcriptomic datasets from studies involving *in vivo* genetic perturbation (e.g. knockdown, knockout and overexpression) in eight model organisms, including mouse, rat and zebrafish. A total of 10 116 genetic perturbation datasets, among which 122 are scRNA-seq datasets, were manually curated. The raw data were checked and analysed using widely accepted and standardized pipelines, and 8 644 148 DEGs associated with gene perturbations were identified. Downstream changed functions/pathways/cell types as well as altered phenotypes are also provided when available. It is expected that this service will contribute to the understanding of gene functions, complex diseases and GMOs and is aimed at becoming a valuable resource for the research community.

MATERIALS AND METHODS

Workflow of PertOrg

The workflow of PertOrg 1.0 is presented in Figure 1. The methods employed to construct the database and the instructions for using the web server are described below.

Data collection and processing

We searched the GEO (3) and ArrayExpress (4) databases to retrieve genetic perturbation studies in model organisms using the keywords 'knockout, knockdown, overexpression, knockin, transgenic, shRNA, siRNA, RNAi, CRISPRi, CRISPRko and CRISPRedit'. The search results were further manually checked to determine whether they met the criteria, i.e. contained data from both before and after genetic perturbation. Relevant information was also collected, including phenotypic alterations when available. In a study, the samples, which could be grouped into 'control' (before perturbation) and 'case' (after perturbation), were organized into one perturbation dataset. This process was double-checked by at least two researchers. We further downloaded most of the metadata in raw format, such as expression profiles annotated by probe IDs and raw RNA sequencing data. For microarray data, the expression profiles from GEO were downloaded using the R package GEOquery (8), and the processed profiles in ArrayExpress were downloaded using ftp links. For RNA sequencing data, compressed FASTQ files were mainly downloaded from the European Nucleotide Archive (9) and DNA Data Bank of Japan (10). Customized workflows for different library strategies of experimental studies were adopted and are described below:

(i) *Microarray data*: For the convenience of downstream analysis, the probe identifiers of each transcriptome profile were converted to Entrez Gene IDs. Multiple probes matched to the same Entrez Gene ID were merged using the average expression values. Then, $\log_2$ transformation was performed on those gene expression profiles that were not transformed before. Differential expression analysis was performed using limma (v3.52.1) (11).

(ii) *RNA-seq data*: Reference genomes and annotations of eight model organisms were obtained from the RefSeq database (12). Then, after the md5 check and quality control by fastp (v0.23.1) (13), alignments were performed by HISAT2 (v2.2.1) (14). Furthermore, featureCounts (v2.0.1) (15) was used to estimate gene expression levels. DESeq2 (v1.36.0) (16) or edgeR (v3.38.1) (17) was used for DEG analysis of samples with or without replicates, respectively.

(iii) *scRNA-seq data*: scRNA-seq data from two different protocols, 10X Genomics and Smart-Seq2, were analysed using different pipelines (18,19). For 10X Genomics, we used CellRanger (v6.0.2) (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger) and its built-in reference genome to identify the barcode of each cell as well as the unique molecular identifier (UMI) to construct a matrix of UMI counts for each sample. For Smart-Seq2, the bulk RNA-seq analysis pipeline described above was applied. We used SCTransform (v0.3.3) (20) to normalize the count matrix, and Harmony (v0.1.0) (21) was applied to integrate the profiles in the dataset. A series of analyses were further performed using Seurat (v4.1.1) (22), including visualization, clustering and DEG analysis. Annotation of the cell types was performed by scType (23). All analyses mentioned above were based on default parameters to build a relatively standard analysis pipeline.

For each PertOrg dataset, we performed differential expression analysis and identified DEGs using a threshold *P*-value <0.05. If >1000 DEGs were found, we selected only the top 1000 DEGs ranked by absolute $\log_2$ fold change for downstream analysis.

Enrichment analyses

We performed enrichment analyses including upstream regulatory information and downstream functional analysis using the R package ClusterProfiler 4.0 (24). The upstream regulatory transcription factor-target gene information was extracted from TRRUST (version 2) (25). For downstream functional analysis, Gene Ontology (GO) (26) and KEGG pathway (27) enrichment analysis of DEGs and gene set enrichment analysis (GSEA) of each PertOrg dataset based on fold changes (28) were performed. Moreover, we investigated differential cell type markers derived from the CellMarker (29) and PanglaoDB (30) databases in mouse datasets. For the overlapping analysis of DEGs from PertOrg datasets, or with user-submitted gene sets, the significance was determined using a hypergeometric test.

Database implementation

The web server was built using technologies such as Nginx and uwsgi. Django and MySql were used for back-end data exchange, and Bootstrap 4 and Jquery were used for front-end visualization. Statistical analyses were performed

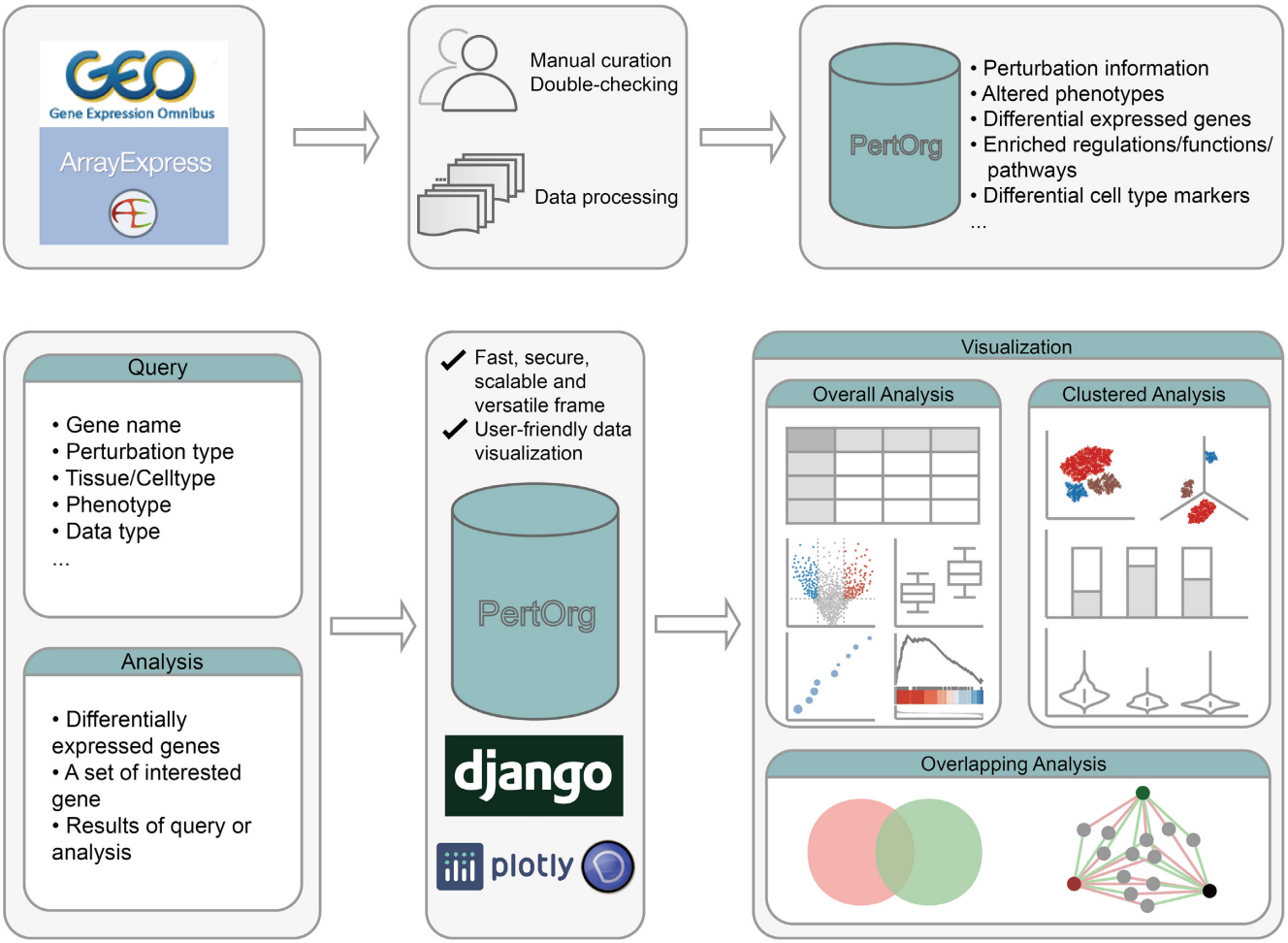

Figure 1. Workflow of PertOrg: First, data from studies involving *in vivo* genetic perturbation were manually collected and curated from GEO and ArrayExpress. The data were then organized as genetic perturbation datasets and analysed for DEGs and functional enrichment. The results were deposited in the user-friendly database, and tables and plots were provided for users.

using Python packages such as Pandas (v1.3.5), NumPy (v1.21.2), SciPy (v1.7.3) and GSEApy (v0.10.8). Data visualizations were carried out via several approaches, such as a volcano plot, heatmap, bubble chart, bar chart, UMAP and t-SNE.

RESULTS

Database usage

Three data retrieval methods, including quick search, advanced search and browse, were developed for accessing PertOrg 1.0 (Figure 2). For the rapid retrieval of genetic perturbation datasets, PertOrg provides a quick search method on the home page. Via the quick search form, the user can enter a keyword of gene perturbation, e.g. gene name (Slc39a8), perturbation type (knockout), tissue (kidney), phenotype (anaemia) or data type (scRNA-seq). The datasets with information containing the keyword are presented on the 'Result' page. In addition, PertOrg offers an advanced search on the search page for users to acquire perturbation datasets from different aspects. Users can input one or a list of genes (IDs or names) in the form and simultaneously select other feature (i.e. 'organism', 'tissue', 'perturbation type', 'data type' and 'phenotype') options to perform a more precise search. Perturbation datasets meeting the criteria are listed on the 'Result' page. Moreover, the list of perturbation datasets can be viewed in an interactive table on the 'Browse' page. Users can customize filters using 'perturbation type', 'data type', 'tissue type' and 'organism'.

After clicking the PertOrg ID on the 'Result' or 'Browse' page, the user will be redirected to the detailed information page for this dataset (Figure 3). The detailed information page contains the available perturbation information, altered phenotypes, differential genes, enriched upstream transcription factors, enriched downstream functions/pathways and differential cell type markers by genetic perturbations. The perturbation information includes gene information, perturbation type, tissue type, data source and external links to some well-known databases, such as the NCBI Entrez Gene database (31), Ensembl (32) and KnockTF (5), when available. In the 'differential genes' section, DEGs induced by genetic perturbation are listed in the table; volcano plots, heatmaps and boxplots

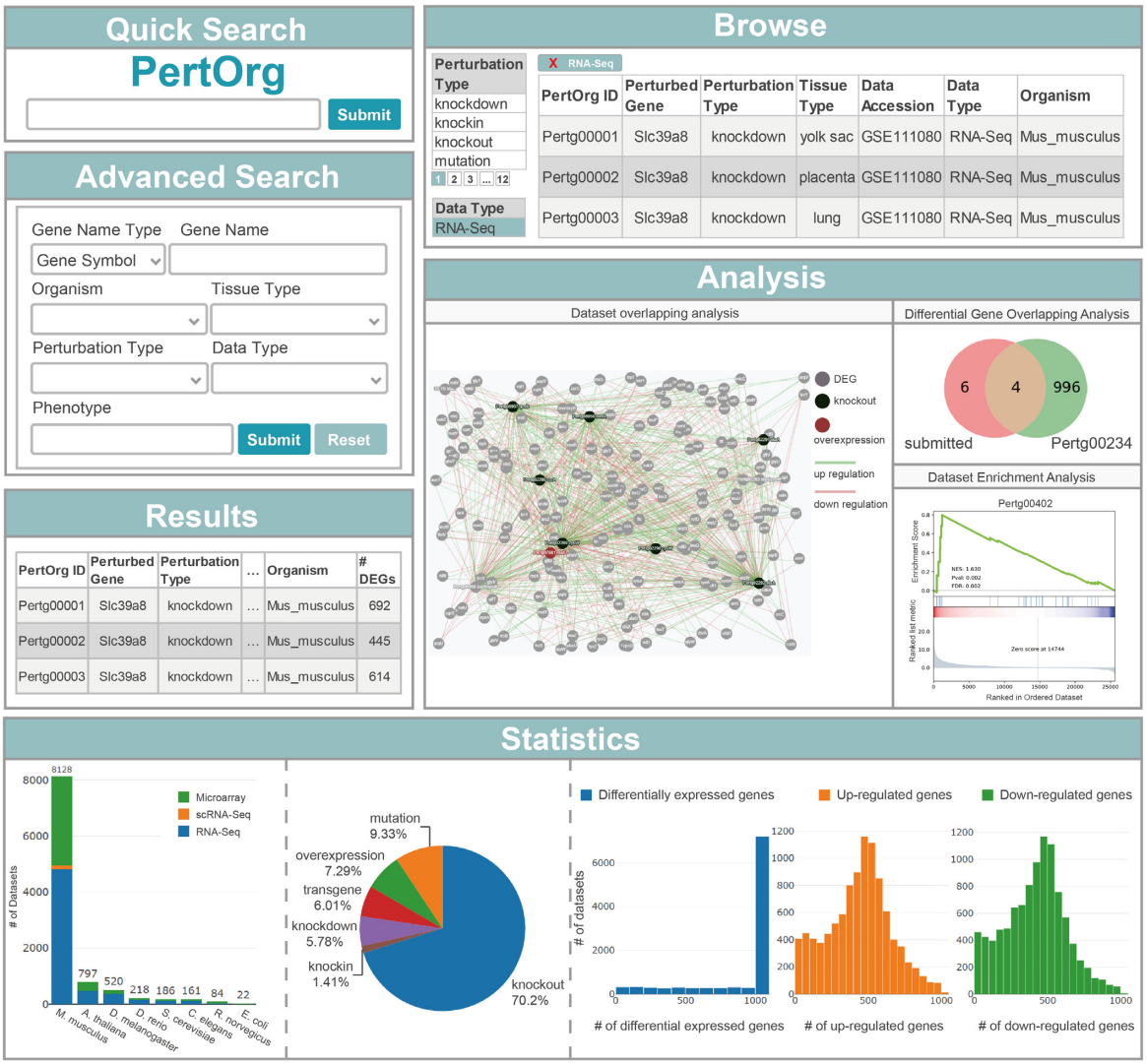

Figure 2. Contents of PertOrg: On the home page, users can conduct quick searches via the keyword of genetic perturbation. On the search page, users can perform more advanced searches by filtering one or more features of the perturbation dataset. On the 'Result' page, datasets meeting the criteria will be listed in a table. On the 'Browse' page, users can browse datasets filtering by perturbation type, data type, tissue type and organism. Three analytical tools were implemented: *differential gene overlapping analysis* helps to investigate the perturbation datasets that are significantly enriched in the user-given gene set by hypergeometric testing*; dataset enrichment analysis* helps to identify the perturbation datasets in which the user-given gene set is overrepresented; and *dataset overlapping analysis* shows the common functions and DEGs of the found perturbation datasets in the 'Result' page. Statistical graphics are also presented to visualize the distribution of datasets, perturbation type and DEGs on the statistics page.

are also provided for the users. For the mouse dataset, cell type markers that are also up/downregulated genes are selected in the table. Furthermore, for the scRNA-seq dataset, the single-cell clustering analysis will generate the cell clustering plot, cell ratio, cluster marker genes and differential genes in clusters. In a cell clustering plot, users can select different resolutions and different types (t-SNE or UMAP). For the mouse scRNA-seq dataset, plotting by inferred cell type is also provided when available. Users can search genes for expression in different clusters. The cell ratio plot shows the ratio of cells in each cluster for the case (after perturbation) and control (before perturbation) conditions. Moreover, cluster marker genes and DEGs in clusters are also listed in the table, and violin plots are provided for visualization.

Analytical tools

PertOrg provides three practical analytical tools, including differential gene overlapping analysis and dataset enrichment analysis on the 'Analysis' page and dataset overlapping analysis on the 'Result' page. These tools enable users to compare their gene signatures with PertOrg datasets or to compare retrieved PertOrg datasets (Figure 2).

Differential gene overlapping analysis helps to find perturbation datasets in which DEGs are significantly enriched in the user-given gene set via hypergeometric testing on the 'Analysis' page. It can also allow users to search DEGs in PertOrg. Users first submit a list of DEGs classified as up- or downregulated. Then, these genes are mapped to the relevant datasets filtered by organism, and hypergeometric tests are performed successively. Qualified datasets are

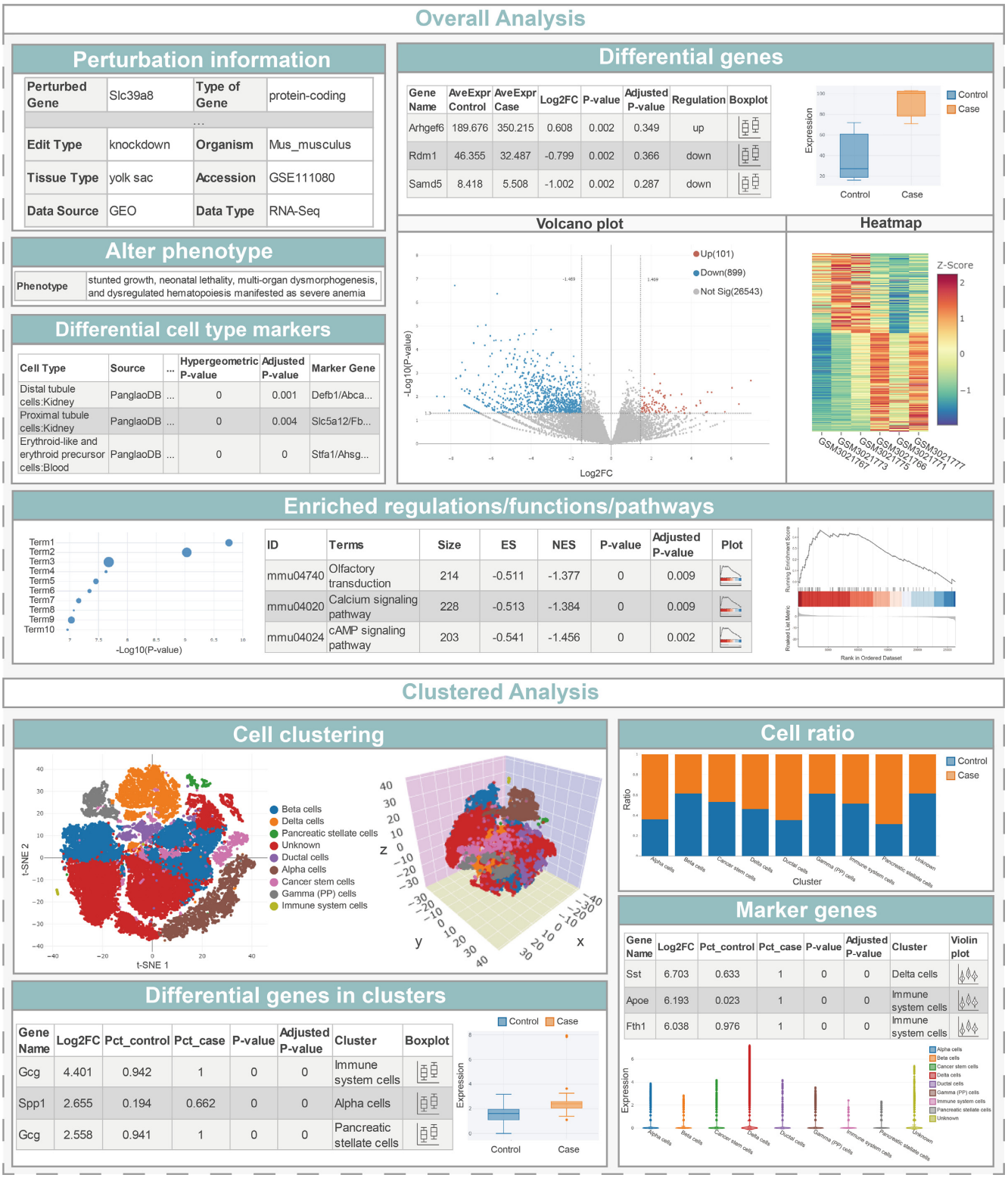

Figure 3. Detailed information page: Users can view details of the perturbation information, altered phenotypes, differential genes, enriched functions/pathways, differential cell type markers by genetic perturbation and single-cell RNA clustering analysis, including the cell ratio, cell markers and differential genes in clusters when available. A volcano plot, heatmap, bubble chart, bar chart, UMAP and t-SNE are provided for visualization.

sorted by *P*-value to obtain the most significant perturbation datasets. Furthermore, by clicking on the dataset ID in the result table, Venn diagram(s) and overlapping genes are shown on the detailed information page.

Dataset enrichment analysis helps to identify the perturbation datasets in which the user-given gene set is overrepresented on the 'Analysis' page. Users submit a set of genes, and the tool iterates the submitted gene set over all PertOrg datasets of that organism via the GSEA algorithm. Available datasets will be returned and sorted by enrichment score. When one of the resulting datasets is clicked, a section to display the GSEA results will be found on the detailed information page.

Dataset overlapping analysis will be provided on the 'Result' page if >1 perturbation dataset is searched. The overlapping downstream gene perturbation DEGs will be shown in a network that contains a maximum of 10 perturbations with more DEGs. Common GOs/pathways of related perturbed genes will also be listed.

Statistics and download

PertOrg 1.0 collected 58 707 bulk and single-cell transcriptome profiles, and 10 116 comparison datasets, including 122 scRNA-seq datasets, were organized and analysed. These datasets were focused on genetic perturbation *in vivo* in eight model organisms, including *Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Danio rerio*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Rattus norvegicus* and *Escherichia coli*. A total of 3958 different perturbed genes (protein-coding genes, microRNAs and long noncoding RNAs) were involved across 491 tissue types. A total of 8 644 148 DEGs associated with gene perturbations (e.g. knockdown, knockout and overexpression) were identified. For different organisms, mice have the most perturbation datasets, while for perturbation types, most datasets are of knocked out types (Figure 2). On the statistics page, interactive charts are provided. For example, the histogram of the number of samples shows that most perturbation datasets have four to eight samples.

Moreover, PertOrg provides downloads of summary information (e.g. perturbated gene, perturbation type, tissue type, data source, data type and organism) for all genetic perturbation datasets in 'csv' format. PertOrg also offers the 'gmt' file of DEGs from all datasets. In addition, differential gene expression analysis results (e.g. *P*-value and fold change) and gene expression profiles of each PertOrg dataset can also be downloaded. Finally, all results in tables can be exported from the interactive HTML table, and all interactive plots can be downloaded in PNG format.

Data submission

Although an increasing number of perturbation datasets are expected to be collected, many available datasets could still be missed. Therefore, users are encouraged to submit their datasets that meet PertOrg criteria for future integration. PertOrg allows users to submit perturbed genes, perturbation types, tissue types, data types and organisms, as well as a link to their data sources on the 'Submit' page. Once the submission is received, data will be carefully evaluated and further processed using our standard procedures in PertOrg, as described in the 'Materials and Methods' section. Finally, the dataset and the corresponding analysis results will be included in the future release of PertOrg.

Case studies

Two case studies investigating genetic perturbation-induced alterations are presented to demonstrate how users can utilize the database.

Case study 1: memory impairment-related genetic perturbation. In the quick search or advanced method, we can search 'memory impairment' as a phenotype. As a result, nine PertOrg datasets and three perturbed genes, including *Creb1*, *Adnp* and *Jade2*, were found (Supplementary Table S1). When the 'Overlapping analysis' button at the bottom of the 'Result' page was clicked, common features of those nine datasets were displayed. We found that these three perturbed genes were all involved in the process of 'neuron projection morphogenesis' (Supplementary Table S2). Deficiency of *Adnp* has been reported to cause brain disorders in both mice and humans, and *Adnp*-knockout mice have been used as disease models of brain disorders (33).

Case study 2: immune response induced by genetic perturbation. On the analysis page, to explore the immune response induced by genetic perturbation, we searched '*cd8a*' in the upregulated gene box and selected the organism 'mouse'. As a result, 307 PertOrg datasets were found (Supplementary Table S3). After clicking the scRNA-seq dataset 'Pertg06961' in the results table, the detailed information page showed that 68 DEGs induced by knockout of *Ptpn11* were involved in the biological process 'regulation of T cell activation' (Supplementary Table S4). In the 'differential cell type markers' section, we found that many immune cells were 'upregulated', as inferred from differentially expressed cell type markers (Supplementary Table S5). Moreover, the expression of *cd8a* was found to be upregulated by *Ptpn11* knockout in both overall cells and 'immune system cells' in single-cell cluster analysis (Supplementary Figure S1). It has been reported that *Ptpn11* inhibition triggers antitumor immunity by enhancing the function of CD8 cytotoxic T cells (34). Similar results were also observed in the bulk RNA dataset 'Pertg06178' by *Klf14* knockout. *Klf14* has been found to play a role in regulatory T-cell differentiation (35).

CONCLUSIONS AND FUTURE EXPANSIONS

GMOs have been widely used as models in biomedical research. Some GMOs, however, are produced for human consumption. Therefore, understanding the effect of GMOs by genetic perturbation is useful for studying gene functions and biological processes but also benefits human health and disease prevention. An increasing amount of transcriptomic data have been generated from studies involving genetic perturbations in model organisms. However, the data are unorganized and sparsely distributed, which poses a large barrier for knowledge mining. In PertOrg 1.0, these data were manually collected and processed by the unified methods so that they could be compared and analysed. As a

result, 10 116 datasets were organized in the current version. PertOrg provides multilevel (i.e. gene, pathway, cell type and phenotype) alterations induced by genetic perturbations. Compared with other related databases, such as KnockTF, GPA and LINCS, PertOrg 1.0 has three novel features: first, PertOrg has the largest number of datasets from multiple *in vivo* perturbed model organisms; second, the phenotype information is manually collected and shown in the detailed information page when available; and third, single-cell analyses on 122 scRNA-seq datasets are included. Further plans are in place to regularly update the database with newly published data every 3 months. Integrating *in vivo* and *in vitro* data may help to gain deep knowledge of alterations induced by genetic perturbation, e.g. from cellular interactions among different cell types or intrinsic properties of a cell line. Therefore, *in vitro* gene perturbation datasets, including human datasets, will also be integrated and compared with PertOrg datasets. Moreover, other omic data, such as genomics and proteomics, will be included. Other perturbation types, including drug treatment and external stimulation *in vivo*, are also scheduled for inclusion in the next version of PertOrg. Because the analysis of raw transcriptomic data is a time- and space-consuming process, PertOrg 1.0 does not currently support online analysis of user data. However, an in-house pipeline is scheduled to be developed to help users analyse their data using our standard procedures and to integrate the results with PertOrg datasets. Finally, additional analytical tools integrating gene–pathway–cell–phenotype–drug–disease will be developed and implemented.

In summary, we present PertOrg as a comprehensive resource of multilevel alterations induced in model organisms by *in vivo* genetic perturbation. PertOrg 1.0 allows users to link *in vivo* perturbed genes with DEGs, pathways, cell types, tissues and phenotypes to interrogate gene function, tissue development and phenogenesis. Moreover, PertOrg 1.0 is expected to help explore disease models and mechanisms and to assess therapeutic targets and potential gene therapy. As the most comprehensive GMO database available, we believe that PertOrg 1.0 will be a valuable resource for both bioscientists and bioinformaticians.

DATA AVAILABILITY

PertOrg 1.0 is freely available online at http://www.inbirg.com/pertorg/, and there is no login requirement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Part of the computing work in this paper was supported by the Supercomputing Center of Chongqing Medical University. The authors thank Dr Yongjun Dang and Dr Junchi Hu from Chongqing Medical University for their valuable discussions regarding this work.

FUNDING

Chongqing Medical University; National Natural Science Foundation of China [82104063]; University Innovation Research Group Project of Chongqing [CXQT21016]. Funding for open access charge: Chongqing Medical University.

Conflict of interest statement. None declared.

REFERENCES

1. Ma,D. and Liu,F. (2015) Genome editing and its applications in model organisms. *Genomics Proteomics Bioinformatics*, **13**, 336–344.
2. Buiatti,M., Christou,P. and Pastore,G. (2013) The application of GMOs in agriculture and in food production for a better nutrition: two different scientific points of view. *Genes Nutr.*, **8**, 255–270.
3. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
4. Sarkans,U., Füllgrabe,A., Ali,A., Athar,A., Behrangi,E., Diaz,N., Fexova,S., George,N., Iqbal,H., Kurri,S. *et al.* (2021) From ArrayExpress to BioStudies. *Nucleic Acids Res.*, **49**, D1502–D1506.
5. Feng,C., Song,C., Liu,Y., Qian,F., Gao,Y., Ning,Z., Wang,Q., Jiang,Y., Li,Y., Li,M. *et al.* (2020) KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Res.*, **48**, D93–D100.
6. Keenan,A.B., Jenkins,S.L., Jagodnik,K.M., Koplev,S., He,E., Torre,D., Wang,Z., Dohlman,A.B., Silverstein,M.C., Lachmann,A. *et al.* (2018) The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. *Cell Syst.*, **6**, 13–24.
7. 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.
8. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.
9. Harrison,P.W., Ahamed,A., Aslam,R., Alako,B.T.F., Burgin,J., Buso,N., Courtot,M., Fan,J., Gupta,D., Haseeb,M. *et al.* (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
10. Ogasawara,O., Kodama,Y., Mashima,J., Kosuge,T. and Fujisawa,T. (2019) DDBJ database updates and computational infrastructure enhancement. *Nucleic Acids Res.*, **48**, D45–D50.
11. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
12. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
13. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
14. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
15. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
16. 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.
17. Chen,Y., Lun,A.T.L. and Smyth,G.K. (2016) From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, **5**, 1438.
18. Zhang,Y., Zou,D., Zhu,T., Xu,T., Chen,M., Niu,G., Zong,W., Pan,R., Jing,W., Sang,J. *et al.* (2022) Gene Expression Nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels. *Nucleic Acids Res.*, **50**, D1016–D1024.
19. Zhang,Z., Cui,F., Lin,C., Zhao,L., Wang,C. and Zou,Q. (2021) Critical downstream analysis steps for single-cell RNA sequencing data. *Brief. Bioinform.*, **22**, bbab105.

20. Choudhary,S. and Satija,R. (2022) Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.*, **23**, 27.
21. Korsunsky,I., Millard,N., Fan,J., Slowikowski,K., Zhang,F., Wei,K., Baglaenko,Y., Brenner,M., Loh,P. and Raychaudhuri,S. (2019) Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**, 1289–1296.
22. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573.e29–3587.e29.
23. Ianevski,A., Giri,A.K. and Aittokallio,T. (2022) Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.*, **13**, 1246.
24. 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.
25. Han,H., Cho,J.-W., Lee,S., Yun,A., Kim,H., Bae,D., Yang,S., Kim,C.Y., Lee,M., Kim,E. *et al.* (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.*, **46**, D380–D386.
26. The Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
27. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
28. 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.
29. Zhang,X., Lan,Y., Xu,J., Quan,F., Zhao,E., Deng,C., Luo,T., Xu,L., Liao,G., Yan,M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
30. Franzén,O., Gan,L.-M. and Björkegren,J.L.M. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, **2019**, baz046.
31. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
32. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
33. Hacohen-Kleiman,G., Sragovich,S., Karmon,G., Gao,A.Y.L., Grigg,I., Pasmanik-Chor,M., Le,A., Korenková,V., McKinney,R.A. and Gozes,I. (2018) Activity-dependent neuroprotective protein deficiency models synaptic and developmental phenotypes of autism-like syndrome. *J. Clin. Invest.*, **128**, 4956–4969.
34. Zhao,M., Guo,W., Wu,Y., Yang,C., Zhong,L., Deng,G., Zhu,Y., Liu,W., Gu,Y., Lu,Y. *et al.* (2019) SHP2 inhibition triggers anti-tumor immunity and synergizes with PD-1 blockade. *Acta Pharm. Sin. B*, **9**, 304–315.
35. Sarmento,O.F., Svingen,P.A., Xiong,Y., Xavier,R.J., McGovern,D., Smyrk,T.C., Papadakis,K.A., Urrutia,R.A. and Faubion,W.A. (2015) A novel role for Kruppel-like factor 14 (KLF14) in T-regulatory cell differentiation. *Cell. Mol. Gastroenterol. Hepatol.*, **1**, 188–202.