Nucleic Acids Research, 2025, **53**, D1120–D1131
https://doi.org/10.1093/nar/gkae777
Advance access publication date: 11 September 2024
Database issue

PerturbDB for unraveling gene functions and regulatory networks

Bing Yang [1,†], Man Zhang [1,†], Yanmei Shi [1,†], Bing-Qi Zheng [1,†], Chuanping Shi [1], Daning Lu [1], Zhi-Zhi Yang [1], Yi-Ming Dong [1], Liwen Zhu [1], Xingyu Ma [1], Jingyuan Zhang [1], Jiehua He [1], Yin Zhang [1], Kaishun Hu [1], Haoming Lin [2,*], Jian-You Liao [1,3,*] and Dong Yin [1,*]

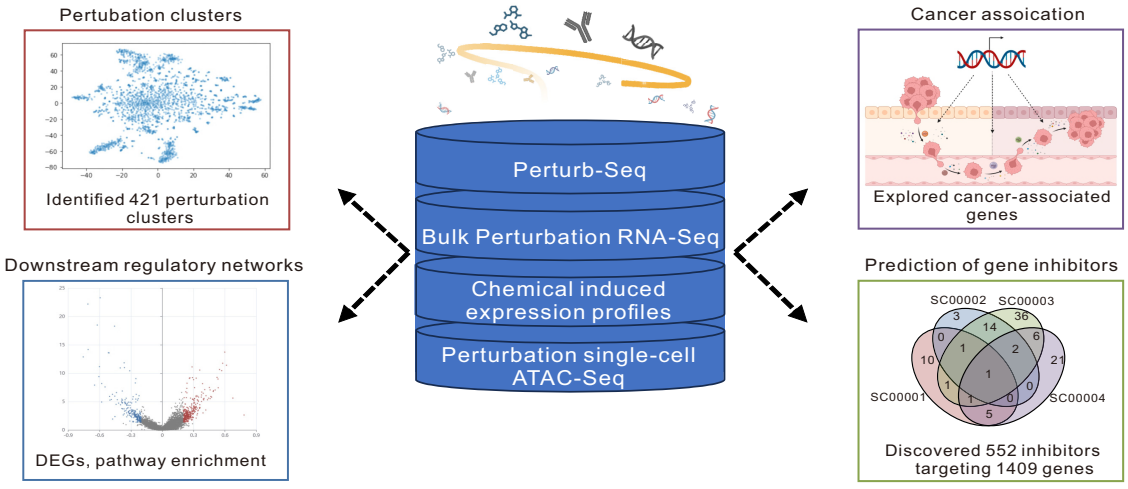
[1]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangdong–Hong Kong Joint Laboratory for RNA Medicine, Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, 107 Yan Jiang West Road, Guangzhou, Guangdong, 510120, China
[2]HBP Surgery Department, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, 107 Yan Jiang West Road, Guangzhou, Guangdong, 510120, China
[3]Center for Precision Medicine, Shenshan Central Hospital, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, 1 Heng Er Road, Dongyong Town, Shanwei, Guangdong, 516621, China

[*]To whom correspondence should be addressed. Tel: +86 18922182515; Email: yind3@mail.sysu.edu.cn
Correspondence may also be addressed to Jian-You Liao. Tel: +86 1358054805; Email: liaojy3@mail.sysu.edu.cn
Correspondence may also be addressed to Haoming Lin. Tel: +86 13902206967; Email: linhaom@mail.sysu.edu.cn
[†]The first four authors should be regarded as Joint First Authors.

Abstract

Perturb-Seq combines CRISPR (clustered regularly interspaced short palindromic repeats)-based genetic screens with single-cell RNA sequencing readouts for high-content phenotypic screens. Despite the rapid accumulation of Perturb-Seq datasets, there remains a lack of a user-friendly platform for their efficient reuse. Here, we developed PerturbDB (http://research.gzsys.org.cn/perturbdb), a platform to help users unveil gene functions using Perturb-Seq datasets. PerturbDB hosts 66 Perturb-Seq datasets, which encompass 4 518 521 single-cell transcriptomes derived from the knockdown of 10 194 genes across 19 different cell lines. All datasets were uniformly processed using the Mixscape algorithm. Genes were clustered by their perturbed transcriptomic phenotypes derived from Perturb-Seq data, resulting in 421 gene clusters, 157 of which were stable across different cellular contexts. Through integrating chemically perturbed transcriptomes with Perturb-Seq data, we identified 552 potential inhibitors targeting 1409 genes, including an mammalian target of rapamycin (mTOR) signaling inhibitor, retinol, which was experimentally verified. Moreover, we developed a 'Cancer' module to facilitate the understanding of the regulatory role of genes in cancer using Perturb-Seq data. An interactive web interface has also been developed, enabling users to visualize, analyze and download all the comprehensive datasets available in PerturbDB. PerturbDB will greatly drive gene functional studies and enhance our understanding of the regulatory roles of genes in diseases such as cancer.

Graphical abstract



Received: June 8, 2024. **Revised:** July 26, 2024. **Editorial Decision:** August 13, 2024. **Accepted:** September 5, 2024

Introduction

Perturb-Seq, an innovative and powerful technique, combines CRISPR (clustered regularly interspaced short palindromic repeats)-based screens with single-cell RNA sequencing (scRNA-Seq) readouts for high-content phenotypic screens to comprehensively map the transcriptional effects of genetic perturbations genome-wide (1–4). This technique is highly efficient for large-scale analysis of gene functions and regulatory networks. With the growing prevalence of Perturb-Seq technology, its datasets are rapidly accumulating, offering invaluable opportunities for a rapid understanding of gene functions. However, the large volume and diverse processing methodologies of these datasets pose significant challenges for researchers in repurposing Perturb-Seq datasets.

RNA-Seq of a heterogeneous population of cells subjected to knockdown/knockout of a specific gene is a widely used strategy to uncover gene functions and regulatory networks (5). To distinguish the traditional RNA-Seq strategy based on bulk cell populations from the recently emerged single-cell-based approach known as Perturb-Seq, we will hereafter refer to the former as bulk perturbation RNA-Seq. The bulk perturbation RNA-Seq technology has been frequently used in gene function studies for a long time with a huge amount of bulk perturbation RNA-Seq datasets accumulated. To facilitate researchers in reusing these bulk perturbation RNA-Seq datasets, several bulk perturbation RNA-Seq databases have been constructed, including KnockTF (6), GPSAdb (7) and Gene Perturbation Atlas (8).

The traditional bulk perturbation RNA-Seq is a low-throughput experiment with high costs and time-consuming procedures. Most bulk perturbation RNA-Seq techniques perturb only one or a few genes, resulting in a limited number of perturbations per experiment. Consequently, existing bulk perturbation RNA-Seq datasets are generated from a large number of experiments by numerous labs across various cell lines. Performing bulk perturbation RNA-Seq in various genetic backgrounds presents a challenge when it comes to directly comparing and integrating the transcriptional phenotype changes induced by different gene perturbations. However, this step is crucial for predicting gene functions and unraveling regulatory networks (1). The development of Perturb-Seq technology has enabled the rapid production of transcriptional phenotypes for tens of thousands of genetic perturbations in a single experiment and within one cell line, all at an acceptable cost (1,2). These transcriptional phenotypes are generated from cells with highly similar genetic backgrounds, empowering researchers to unveil gene functions through integrated analysis of transcriptional phenotypes resulting from different gene perturbations.

The field lacks an intuitive and interactive online platform that leverages the growing number of Perturb-Seq datasets for gene functional analysis. Currently, there are three platforms available for Perturb-Seq data analysis. scPerturb (9), designed for researchers with programming skills, contains a compendium of standardized datasets to aid the development and benchmarking of computational approaches. GeneSetR (10), a well-designed web server, offers gene set analysis using only three published Perturb-Seq datasets. The limited number of datasets in GeneSetR restrict its scope and application. PerturBase (11) is a database for interactively exploring a broad spectrum of Perturb-Seq datasets. However, it falls short in providing comprehensive functionalities for in-depth dataset exploration. This limits its ability to support thorough analyses and generate deeper insights from the data collected. Creating a user-friendly platform that enables in-depth and integrated analysis of Perturb-Seq datasets with other datasets, such as cancer genomic datasets, will enhance the investigation of gene functions and their regulatory role in diseases.

In this study, we manually collected 66 Perturb-Seq datasets and analyzed them using the uniform algorithm, Mixscape (12), to generate perturbed transcriptomic phenotypes for 3395 genes. We clustered these genes based on their perturbed transcriptomic phenotypes, resulting in 421 gene clusters. To enhance our analysis of Perturb-Seq datasets, we integrated them with The Cancer Genome Atlas (TCGA) program dataset (13–15), the Library of Integrated Network-Based Cellular Signatures (LINCS) program dataset (16) and bulk perturbation RNA-Seq datasets (17). These integrations help us to understand gene regulatory roles in cancers, identify novel gene inhibitors and predict gene functions. We also designed an intuitive and interactive web interface, PerturbDB (http://research.gzsys.org.cn/perturbdb), to easily assess to extensive Perturb-Seq datasets and the insights gained from these integrated analyses. The comprehensive features of PerturbDB will make it an invaluable and indispensable resource for both cancer and gene functional research.

Materials and methods

Data collection

To collect Perturb-Seq datasets, we performed a search on the National Center for Biotechnology Information (18) and Single Cell Portal (https://singlecell.broadinstitute.org/single_cell) using the keyword 'perturb seq', 'single cell crispr' or 'scRNA perturb'. After acquiring the accession numbers, expression matrices for all datasets were retrieved from databases, including the Gene Expression Omnibus (19), Zenodo (https://zenodo.org) and FigShare (https://figshare.com), for further analysis. Additional Perturb-Seq datasets were obtained from published articles and preprints, such as scPerturb (9) and PerturBase (11), to investigate gene function, gene expression in tumors and combined drug prediction. So far, our dataset comprises transcriptomic perturbation data from human species only. We have actively compiled data through March 2024. In total, 27 papers yielding 66 Perturb-Seq datasets were identified (Supplementary Table S1). Bulk perturbation RNA-Seq datasets were obtained from GPSAdb (7) under a Creative Commons Attribution 4.0 International License. Following filtering, we retained 2999 datasets with single gene perturbations.

The RNA-Seq transcriptomes and clinical data were retrieved from the TCGA database using the R/Bioconductor package TCGAbiolinks (v2.8.4) (20). The transcripts per million (TPM) data for gene expression levels in cancer cell lines were sourced from the DepMap portal's CCLE (Cancer Cell Line Encyclopedia) project (https://depmap.org/portal/data_page/?tab=allData). Similarly, the TPM data in normal human tissues were acquired from the GTEx (Genotype-Tissue Expression) project (https://www.gtexportal.org/home/downloads/adult-gtex/overview). The implementation of box plots for visualizing these messenger RNA (mRNA) expression data was accomplished through the utilization of Highcharts.js. Differential analysis was conducted using one-way analysis of variance (ANOVA), with disease state (tumor or normal) as the variable to calculate differential expression in the TCGA datasets. Kaplan–Meier survival analysis for the

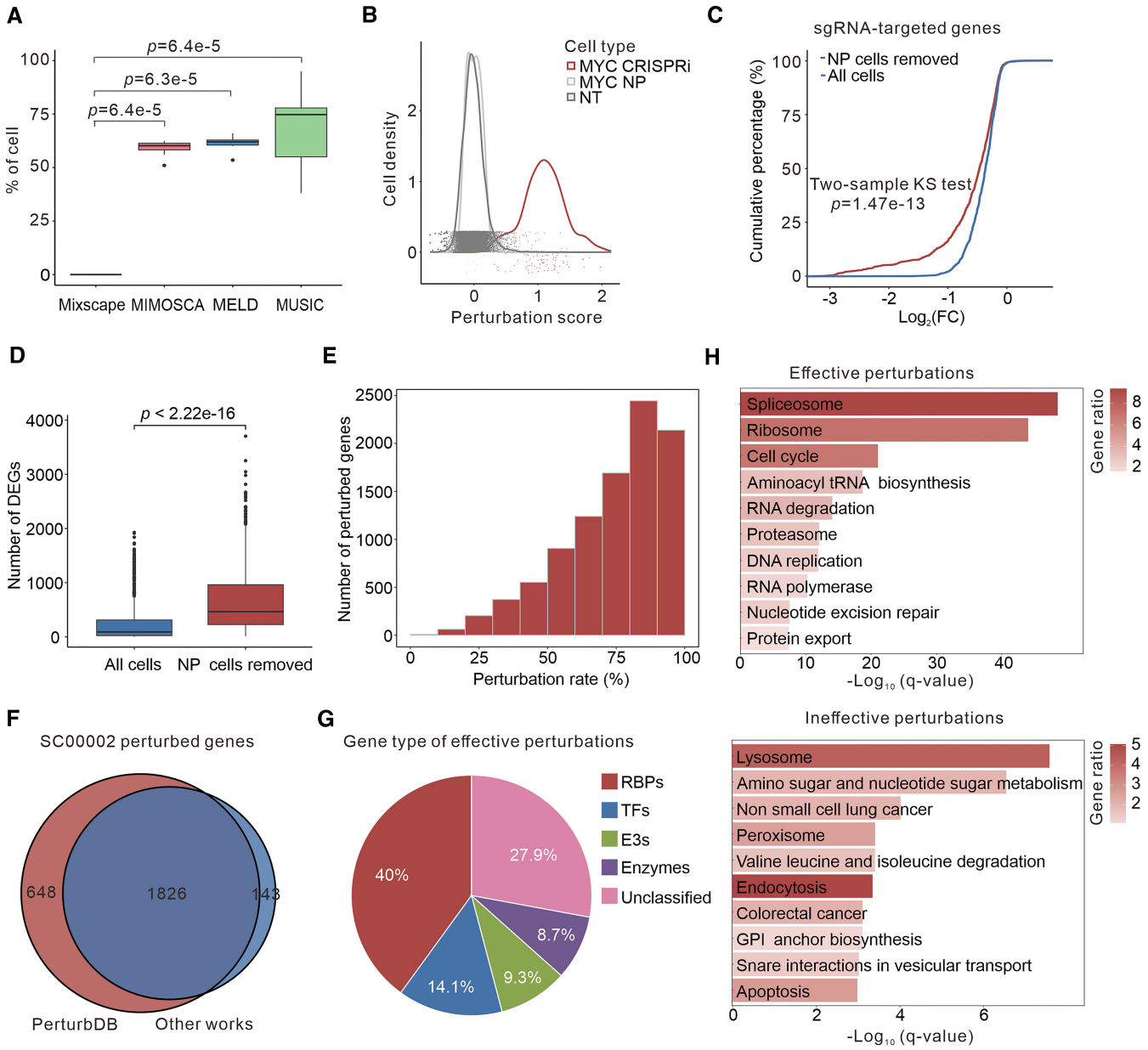


Figure 1. An overview of genome-scale Perturb-Seq datasets collected by PerturbDB. (**A**) Box plot showing the proportion of NT cells erroneously identified as perturbed cells by four algorithms: Mixscape, MIMOSCA, MELD and MUSIC ($n = 10$). Statistical significance was calculated by the Wilcoxon test. (**B**) For cells expressing the MYC proto-oncogene gRNA, density plot showing the distribution of perturbation scores and posterior probabilities for NT cells (gray) and CRISPR-edited cells (red) identified by the Mixscape algorithm. The horizontal axis indicates the distribution of cell perturbation scores, with cells having scores $\geq$0.5 identified as perturbed cells. The vertical axis represents cell density. NT, non-targeting; NP, non-perturbed. (**C**) Cumulative density curves showing expression changes of sgRNA-targeted genes in perturbed cells with <50% perturbation rate before and after the removal of NP cells. Statistical significance was calculated by a two-sample Kolmogorov–Smirnov test. (**D**) Box plot showing the number of DEGs in perturbed cells with <50% perturbation and NT cells before and after the removal of NP cells. Statistical significance was calculated by a paired Student's *t*-test. (**E**) Frequency histogram showing the distribution of the number of perturbed genes with varying perturbation efficiency as processed by the Mixscape algorithm. (**F**) Venn diagram comparing the total number of perturbed genes in PerturbDB with other published works. (**G**) Pie chart showing the proportion of efficiently perturbed genes coding for different categories of functional proteins. (**H**) KEGG pathway enrichment analysis of biological processes for effective (up) and non-effective (down) perturbed genes. The horizontal axis represents the adjusted *P*-value, calculated as $-\log_{10}$(*q*-value), while the vertical axis represents enriched biological processes.

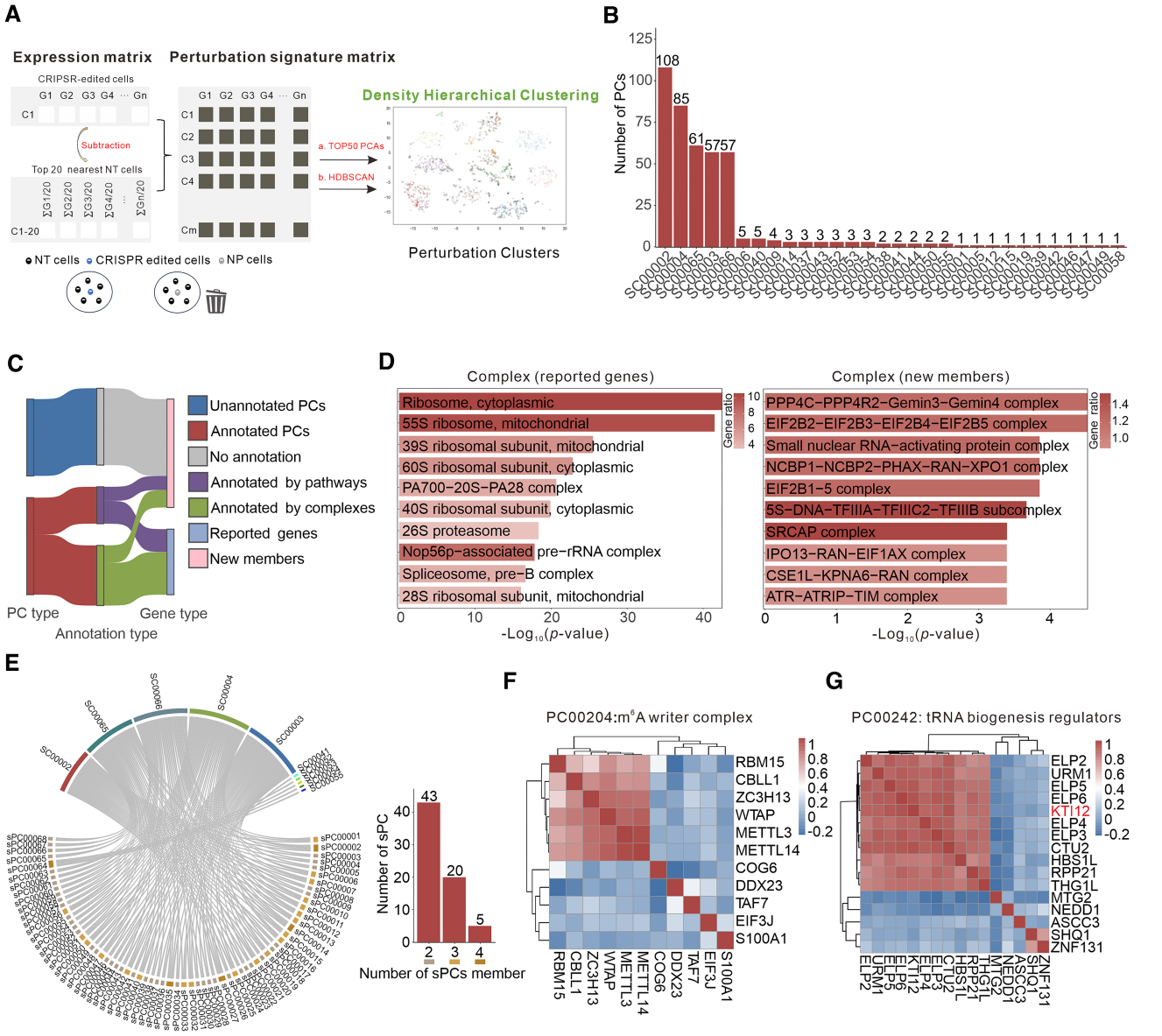


Figure 2. Construction and annotation of perturbation clusters. (**A**) Schematic of perturbation cluster identification. (**B**) Bar plot showing the distribution of perturbation cluster number identified in Perturb-Seq datasets collected by PerturbDB. (**C**) Sankey diagram showing the distribution of annotated and unannotated perturbation clusters labeled as genes with new functions (new members) and genes with known functions (reported genes) and their annotation types of pathways or complexes. (**D**) Enrichment analysis of reported genes (left) and new members (right) of annotated perturbation clusters by complex gene sets. The horizontal axis represents the adjusted *P*-value, calculated as $-\log_{10}$(*P*-value), while the vertical axis indicates enriched complexes. (**E**) Chord plot showing the distribution of stable perturbation clusters in different datasets, and bar plot (right) showing the number of members that form stable perturbation clusters. Stable perturbation cluster is the perturbation cluster that can be clustered in at least two datasets, with >60% of its members shared across different perturbation clusters. (**F**) Pearson correlation heatmap representing genes from PC00204 and five random gene selections, with all genes within this cluster confirmed as participants in N^6-methyladenosine (m^6A) modification. (**G**) Pearson correlation heatmap representing genes from PC00242 and five random gene selections, with all genes within this cluster confirmed as participants in transfer RNA (tRNA) biogenesis.

TCGA datasets was performed using the R package survival (v2.43-3).

Identification of effectively perturbed cells

The Mixscape pipeline of the Seurat package (v4.2.1) (21) was used for the identification of perturbed genes. All cells containing a single-guide RNA (sgRNA) and control cells were applied similarity analysis and perturbation scoring to identify the perturbed cells [with the function RunMixscape (object = eccite, assay = 'PRTB', slot = 'scale.data', labels = 'gene', nt.class.name = 'NT', min.de.genes = 5, iter.num = 10, de.assay = 'RNA', verbose = F)]. Only cells identified as a perturbed status (perturbation score >0.5) were included for further analysis.

Differential expression and functional enrichment analysis

Batch differentially expressed gene (DEG) analysis was conducted by Seurat 'FindMarkers' among the 3395 gene perturbations of Perturb-Seq. All DEG analysis results are available

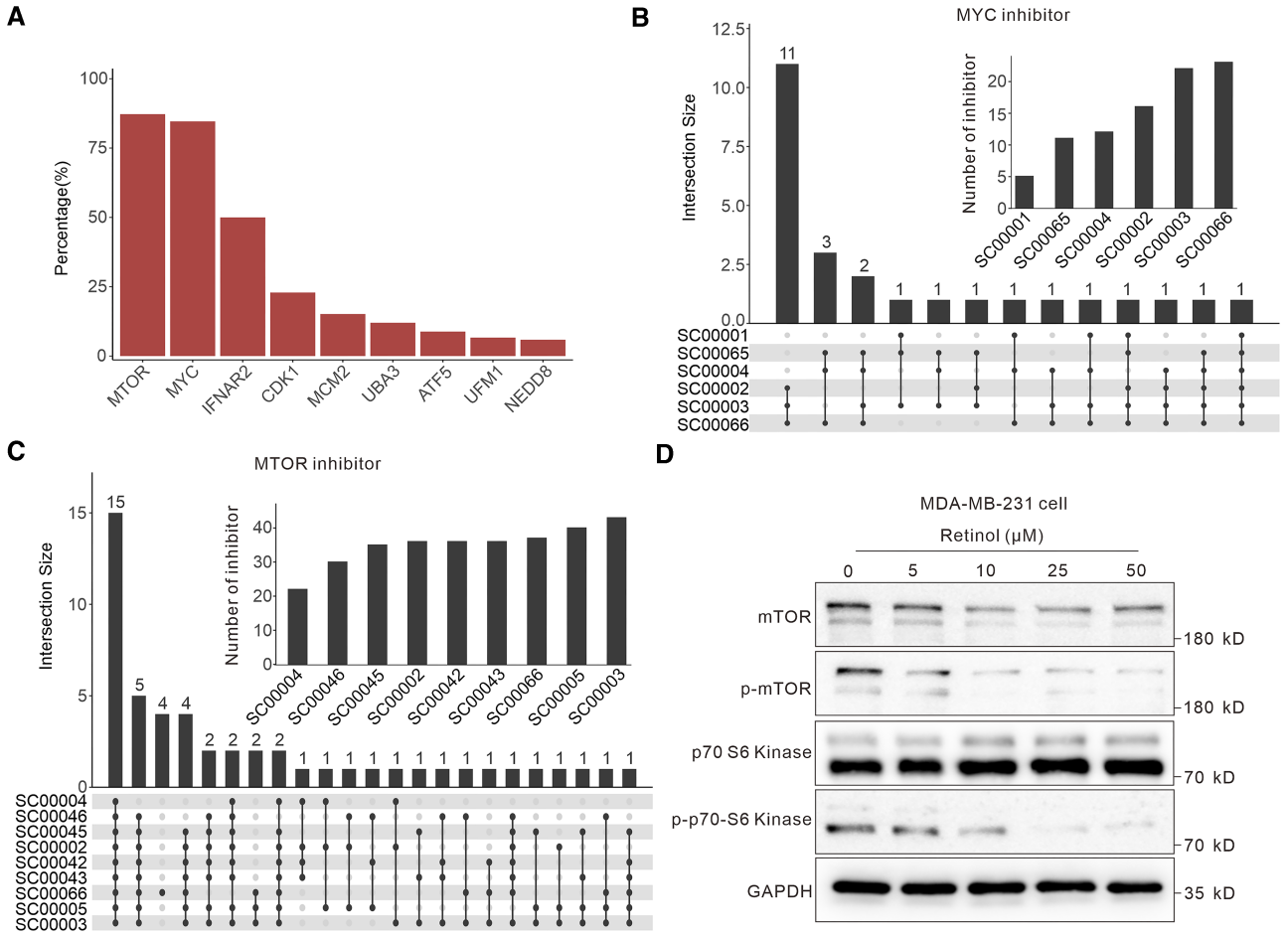


Figure 3. Identification of potential inhibitors for genes of interest. (**A**) Bar plot showing the proportion of reported inhibitors among those predicted by ≥3 datasets for nine genes identified by manual retrieval. (**B** and **C**) UpSet plots (bottom) displaying the reliable MYC inhibitors and MTOR inhibitors predicted by the integrated analysis of six Perturb-Seq datasets with L1000 datasets, where the horizontal axis represents the distribution of intersecting datasets (at least three datasets) and the vertical axis represents the number of inhibitors. Bar plots (top) showing the total number of reliable MYC inhibitors and MTOR inhibitors identified for all the datasets collected by PerturbDB. (**D**) Western blot showing key members' expression level changes upon treatment with retinol at different concentrations in MDA-MB-231 cells.

for downloading to meet users' needs for filtering DEGs at various thresholds. The expression matrices for both perturbed cells and control cells were extracted and subsequently analyzed. Functional enrichment analysis was performed using gene set enrichment analysis (GSEA) (22) within Kyoto Encyclopedia of Genes and Genomes (KEGG) (23), WikiPathways (24) and Hallmark (https://gsea-msigdb.org) databases.

Identification and annotation of perturbation clusters

To identify clusters of similar perturbations, CalcPerturbSig (object = eccite, assay = 'RNA', slot = 'data', gd.class = 'gene', nt.cell.class = 'NT', reduction = 'pca', ndims = 40, num.neighbors = 20, split.by = 'replicate', new.assay.name = 'PRTB') was used to call a perturbation signature matrix. Then, the perturbation signature matrix was applied principal component analysis (PRTB-PCAs). The HDBSCAN algorithm (metric = 'correlation', min_cluster_size = 4, min_samples = 1, cluster_selection_method = 'eom') was performed to identify clusters using PRTB-PCAs in each Perturb-Seq dataset. Perturbation cluster annotation was performed using the R package clusterProfiler (25). The background geometric mean titer files for the hypergeometric distribution test were downloaded from CPDB (26) and CORUM (27), including KEGG (28), WikiPathways (24), Reactome (29) and protein complex information. Only the annotated pathway covering over one-half of the perturbation cluster members was selected for the best complex or pathway annotation.

Analysis and visualization of perturbation single-cell ATAC-Seq data

The perturbation single-cell Assay of Transposase Accessible Chromatin sequencing (scATAC-Seq) datasets were downloaded from the Zenodo database (https://zenodo.org/). Clustering was applied the shared nearest neighbor method using gene score data. Differential significance analysis of gene activity scores following gene perturbation was conducted using the FindMarkers function.

Integrating gene and drug perturbations for predictive drug analysis

The basic information of 2837 chemicals was downloaded from the CLUE database (30,31). The prediction of gene inhibitors was performed by L1000 query tools. The input data of prediction were generated by the following steps: (i) DEGs

derived from the Perturb-Seq and bulk perturbation RNA-Seq datasets were separated into two categories: upregulated and downregulated; (ii) for both upregulated and downregulated genes, solely the top 150 genes, ranked based on $\log_2$(fold change), were employed for the inhibitor prediction exercise; and (iii) DEGs <15 were not included in prediction. The result files comprise outputs of the similarity between DEGs and transcriptional signatures after drug perturbation.

Tools of perturbation-induced GSEA and correlation heatmap visualization (DrawCorrHeatmap)

PerturbDB produced background datasets for perturbation-induced GSEA (PGSEA) based on specific principles. Each dataset of perturbed genes was divided into two groups: upregulated and downregulated gene sets. Gene sets containing <100 genes were excluded due to quality control. Consequently, 16 134 gene sets were retained in PerturbDB. The input data for PGSEA consisted of a list of DEGs [with or without $\log_2$(fold change)]. After clicking the 'Run' button, PGSEA conducts enrichment analysis with the input data and 16 134 gene sets. The PGSEA output identifies which of the perturbed gene sets in PerturbDB are enriched, assisting users in deducing the effects of the input DEGs.

DrawCorrHeatmap employs previously computed dimensionally reduced data from perturbation signature matrices as a foundation for calculating similarities using the Pearson correlation coefficient [cor(data, method = 'pearson')]. Heatmaps are subsequently generated with the pheatmap function's default settings. The DrawCorrHeatmap tool is designed for ease of use: users simply select their desired dataset and specific genes, click the 'Plot' button and instantly receive a heatmap that illustrates the similarity among cell populations following gene perturbation. This functionality allows users to examine the similarity across any combination of genes, thereby aiding in making informed conclusions about their functions.

Database construction

PerturbDB was built based on the Ant Design of Vue (https://1x.antdv.com/docs/vue/introduce-cn/) framework with Vue 2.0; all results mentioned above were stored in MySQL (version 8.0.13). We have tested it on various web browsers, including Google Chrome (preferred), Internet Explorer and Safari of macOS. The PerturbDB website is freely available online (http://research.gzsys.org.cn/perturbdb) without registration.

Cell culture and treatment

Human breast cancer cell line MDA-MB-231 cells were obtained from the ATCC (Manassas, VA, USA) and cultured in Dulbecco's modified Eagle's medium (Life Technologies, Grand Island, NY, USA) supplemented with 1% penicillin–streptomycin and 10% fetal bovine serum (Life Technologies) with 5% CO_2 at 37°C. Following a 24-h period post-seeding, the cells were exposed to a range of retinol concentrations (0, 5, 10, 25 and 50 μM). After a 24-h treatment with retinol, the cells were harvested for subsequent experimental procedures.

Western blot

Cells were lysed with RIPA lysis buffer supplemented with protease and phosphatase inhibitor cocktails (NCM Biotech, China) for 30 min at 4°C and then centrifuged at 13 000 × *g* for 30 min. Protein supernatants were separated by 10% (w/v) sodium dodecyl sulfate–polyacrylamide gel electrophoresis and transferred to polyvinylidene fluoride membranes (Millipore Corporation, USA). After blocking with 5% milk for 1 h at room temperature, the membranes were incubated with the primary antibodies at 4°C overnight. The following primary antibodies were used: anti-MTOR (2983, Cell Signaling Technology, USA), anti-phosphorylated MTOR (5536, Cell Signaling Technology, USA), anti-p70 S6 Kinase (34475, Cell Signaling Technology, USA), anti-phosphorylated p70 S6 Kinase (9234, Cell Signaling Technology, USA) and anti-GAPDH (60004-1-Ig, Proteintech, China). Unprocessed western blot images are provided in Supplementary Figure S4.

Results

An overview of Perturb-Seq datasets collected by PerturbDB

We manually collected a total of 66 Perturb-Seq datasets, consisting of 3 877 310 single-cell transcriptomes undergoing genetic perturbation and 641 211 single-cell transcriptomes of non-targeting (NT) cells. These single-cell datasets were generated from the 22 802 perturbations with the goal of targeting 10 194 genes across 19 different cell lines, including THP1, K562, RPE1, Calu-3, neurons, T cells, MCF10A, MOML-13, HEK293, HepG2, Jurkat and patient-derived melanoma cells. The number of single cells subjected to each perturbation varied, with an average of 170 single cells per perturbation. It is important to note that not all perturbation results in a transcriptional phenotype change at the single-cell level.

Previous research efforts employed diverse algorithms to detect transcriptional phenotype changes in these datasets, making integrated analysis of these datasets challenging. So far, numerous algorithms have been developed to assess sgRNA perturbation effect (32–35). Among them, Mixscape (12), MIMOSCA (2), MELD (36) and MUSIC (37) can distinguish cells that 'escape' perturbation from those significantly perturbed by the same sgRNA. These 'escaping' cells, potentially resulting from in-frame insertion and deletion mutations triggered by sgRNA (12,32,37), may confound the evaluation of perturbation effects. Utilizing cells subjected with control sgRNAs, namely NT cells, we observed that only Mixscape accurately identified all NT cells as non-perturbed (NP) cells (Figure 1A and B). Therefore, we selected Mixscape for analyzing all Perturb-Seq datasets. Following the exclusion of NP cells, we found that the expression changes in sgRNA-targeted genes were significantly higher in perturbed cells compared to all cells (Figure 1C). Moreover, the number of DEGs significantly increased after the removal of NP cells (Figure 1D). These results underscored the importance of excluding NP cells to accurately estimate the perturbation outcomes. Through Mixscape, we found that the genetic perturbation of 3395 genes resulted in significant changes in the transcriptional phenotypes of cells, utilizing the 66 Perturb-Seq datasets. These perturbations were referred to as effective perturbations. For each effective perturbation, Mixscape identified on average 74.1% of single cells showing significant alterations in their transcriptional phenotypes (Figure 1E).

Among the effective perturbations identified in SC00002, 73.8% (1826/2474) had been previously identified, while 26.2% (648/2474) were newly discovered (Figure 1F). The

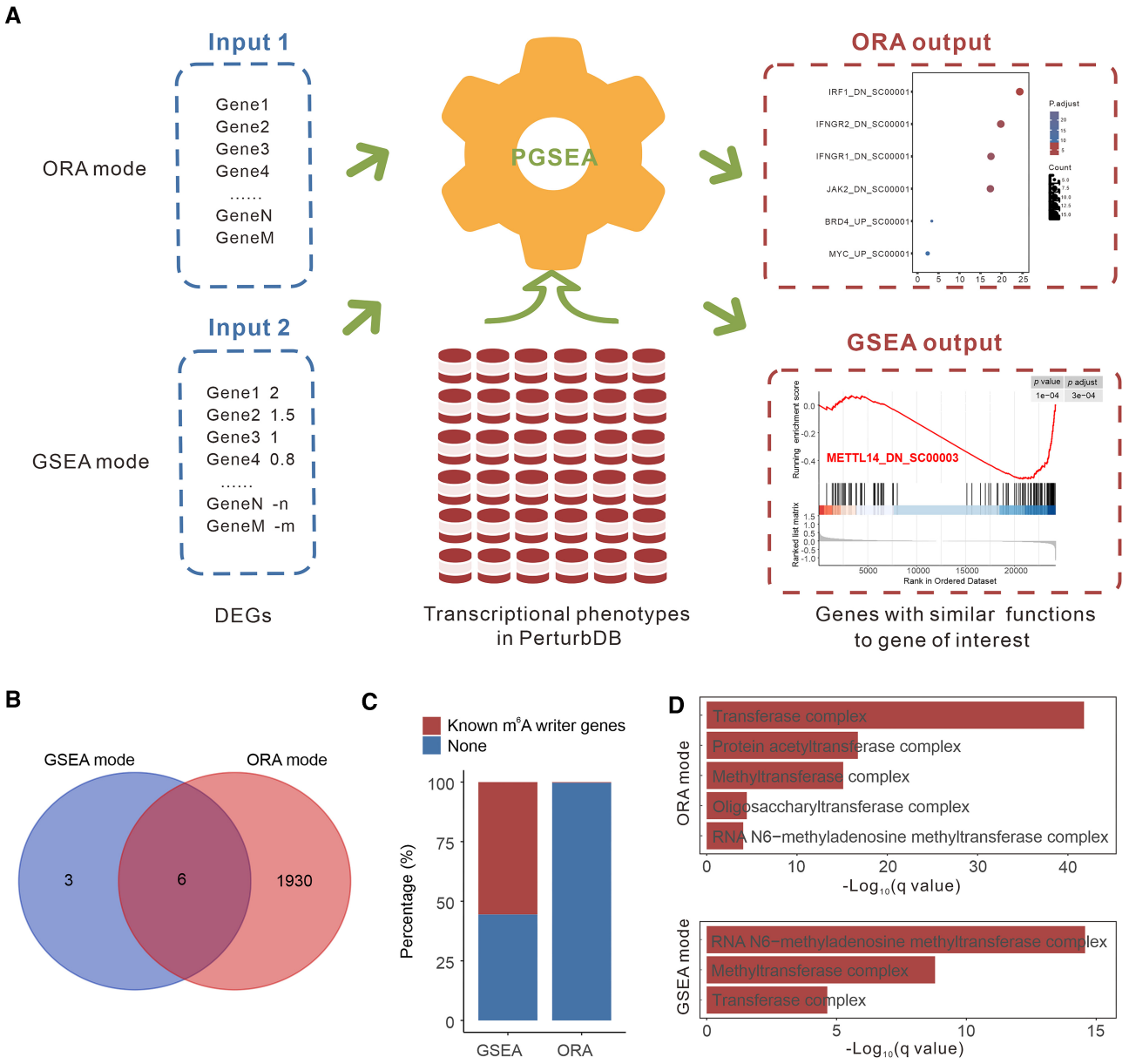


Figure 4. PGSEA as a valuable tool built into PerturbDB to reveal new genes with similar gene functions. (**A**) The flowchart of the PGSEA tool. (**B**) In the case of bulk perturbation RNA-Seq data from K562 cells subjected to METTL3 knockdown by CRISPR interference, Venn diagram showing the output number of genes from the two modes of the PGSEA tool. (**C**) Stacked chart showing the distribution of gene output annotated to m^{6}A writer in two modes of the PGSEA tool. (**D**) GO enrichment analysis of gene output in two modes of the PGSEA tool, highlighting the enrichment of pathways related to methyltransferase activities.

gene targets of effective perturbations exhibited the highest number of RNA binding proteins (RBPs) (Figure 1G). Interestingly, our findings revealed that 51% of gene targets of effective perturbations are essential genes, which are critical for cell survival (38). Consistently, KEGG pathway enrichment analysis showed that gene targets of effective perturbation were more likely to be enriched in fundamental cellular processes, such as ribosome biogenesis, DNA replication and RNA degradation (Figure 1H).

Annotation of gene functions based on transcriptional phenotypes

Gene function study is a challenging task for researchers. The transcriptome of cells undergoing genetic perturbations, referred to as perturbed transcriptome, offers insight into gene functions. Perturb-Seq is capable of generating perturbed transcriptomes for numerous genes within the same genetic background. This enables us to cluster genes with similar functions based on their perturbed transcriptome. To assess the similarity between the perturbed transcriptomes of various genes, we first created a perturbation signature matrix for each cell by subtracting the averaged mRNA expression of the top 20 nearest NT cells from that of the target perturbed cells (Figure 2A). Subsequently, we employed PCA to reduce the dimensionality of the perturbation signature space. Then, the perturbed genes were clustered using the top 50 most important PCA dimensions using the HDBSCAN algorithm (39). Ultimately, we identified 421 perturbation clusters from 30 Perturb-Seq datasets (Figure 2B and

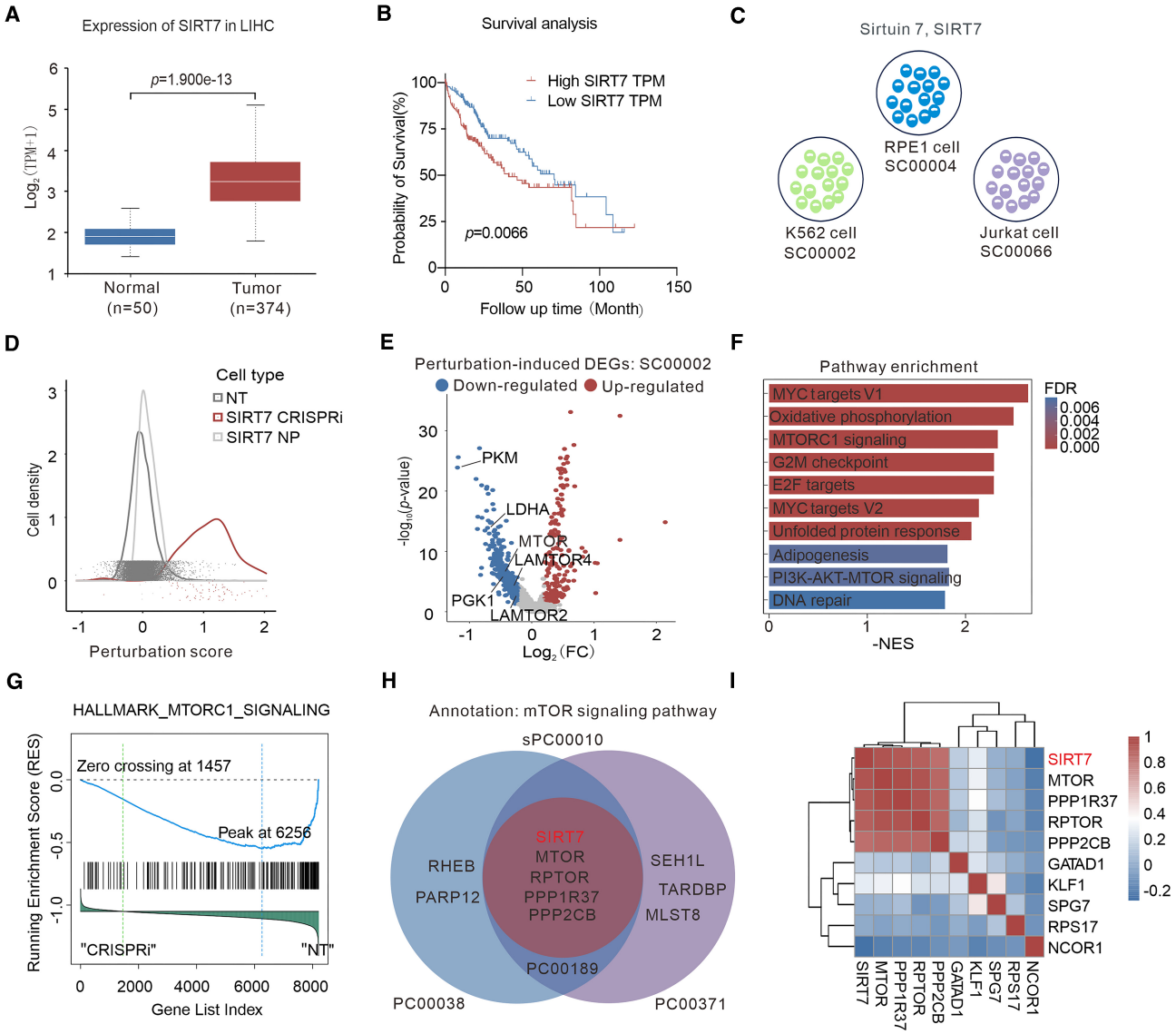


Figure 5. Uncovering the role of SIRT7 in LIHC progression utilizing the PerturbDB platform. (**A**) SIRT7 mRNA expression of 50 normal and 374 tumor samples from LIHC patients. Statistical significance was calculated by one-way ANOVA. (**B**) Kaplan–Meier curve showing the correlation between SIRT7 expression and survival prognosis of LIHC. Statistical significance was calculated by a log-rank (Mantel–Cox) test. (**C**) A protein encoding gene, Sirtuin 7 (SIRT7), was identified in three datasets, SC00002, SC00004 and SC00066, from three cell types, K562 cell, RPE1 cell and Jurkat cell, collected by PerturbDB. (**D**) Density plot showing the distribution of perturbation scores and posterior probabilities for NT cells (gray) and cells expressing SIRT7 gRNA (red) identified by the Mixscape algorithm in SC00002 (K562 cells). NT, non-targeting; NP, non-perturbed. (**E**) Volcano plot showing DEGs in K562 cells expressing SIRT7 gRNA compared to NT cells. Labeled genes are reported downstream genes of the mTOR signaling pathway. (**F**) Bar plot showing the top 10 pathways enriched from GSEA of DEGs in SIRT7 CRISPR-edited K562 cells and NT cells. DEGs, differentially expressed genes. (**G**) MTORC1 signaling pathway enrichment result for DEGs in SIRT7 CRISPR-edited K562 cells and NT cells using the Hallmark gene set. (**H**) Venn diagram showing the intersection of members of the three perturbation clusters annotating to the mTOR signaling pathway from SC00002, SC00004 and SC00066 datasets, mostly key genes of the mTOR signaling pathway and a novel gene, SIRT7. (**I**) Pearson correlation heatmap representing genes from the mTOR pathway and five random gene selections in SC00002, with most of the genes in the cluster being key members of the mTOR pathway and one new member, SIRT7.

Supplementary Table S2), with the median cluster size being 11 members (Figure 2B).

To determine the biological relevance of various perturbation clusters, we annotated them using curated biological relationship datasets, namely CORUM (27), Reactome (40), WikiPathways (24) and KEGG (41). These annotated perturbation clusters encompassed a total of 2680 genes. Of these, 1810 were already known to be associated with the identified complexes or pathways, whereas 870 represented novel members (Figure 2C). This analysis enabled the identification of potential novel roles for 2122 genes. Subsequent complex enrichment analysis showed that genes implicated in these novel roles were significantly enriched in remarkably different complexes compared to other genes within the same annotated perturbation clusters (Figure 2D). We next investigated the recurrence of perturbation clusters across datasets to facilitate the exploration of their potential conserved regulatory functions in various cells. Two perturbation clusters from separate Perturb-Seq datasets that shared over 60% of their members were designated as stable perturbation clusters

across these datasets. Our analysis identified 43 stable perturbation clusters in two datasets, 20 stable perturbation clusters in three datasets, and 5 stable perturbation clusters in four datasets (Figure 2E). With the growth of Perturb-Seq data in PerturbDB, an increased number of stable perturbation clusters will be identified in the future.

Annotation results showed that many perturbation clusters were derived from well-known complexes or pathways. For example, PC00204 contained critical elements of the m^{6}A writer complex, including RBM15, CBLL1, METTL14, METTL3, WTAP and ZC3H13 (Figure 2F). PC00001 consists of key components of the JAK2/STAT1 signaling pathway, including STAT1, JAK2, IFNGR1, IRF1 and IFNGR2 (42) (Supplementary Figure S1). PC00242 comprises a cluster of proteins, KTI12, CTU2, ELP2, ELP3, ELP4, ELP5, ELP6, RPP21, HBS1L, THG1L and URM1, primarily associated with tRNA biogenesis, except for KTI12 and HBS1L (Figure 2G). Recent investigations by Krutyhołowa *et al*. revealed that KTI12 is essential for tRNA biogenesis by regulating tRNA modifications (43), which further highlighted the powerful predictive capabilities of our perturbation clusters for identifying new gene functions. Moreover, our analysis identified numerous perturbation clusters with functions that are currently unknown, providing crucial insights for understanding the regulatory networks essential for cellular homeostasis.

Discovering inhibitors through transcriptional phenotype comparison

The LINCS L1000 project has provided extensive transcriptional phenotype datasets derived from the treatment of nine different cell lines with 2837 chemicals (16). Given that the transcriptional changes induced by genetic perturbations might mirror those triggered by chemicals targeting the same genes, we posited that it is plausible to identify gene inhibitors by comparing transcriptional responses to genetic and chemical perturbations. Utilizing the CLUE algorithm (44), we calculated similarity score for each comparison. Chemicals achieving a CLUE score of at least 95 in three or more Perturb-Seq/bulk perturbation RNA-Seq datasets were classified as potential gene inhibitors. Our approach led to the identification of 552 chemicals that may inhibit 1409 genes. After manual literature review of 263 inhibitors targeting nine classical genes involved in various biological processes, we discovered that, on average, 32.6% of these inhibitors were previously reported to suppress the activities or expression of their target genes (Figure 3A).

Notably, 22 out of 26 MYC inhibitors identified have been documented to inhibit MYC protein or mRNA expression (Figure 3B and Supplementary Table S3). Additionally, 41 out of 47 mammalian target of rapamycin (mTOR) inhibitors identified have been confirmed to inhibit mTOR signaling (Figure 3C and Supplementary Table S3). We also identified retinol, also known as vitamin A, as an mTOR signaling inhibitor. Experimental verification confirmed its efficacy in suppressing mTOR signaling (Figure 3D). This finding aligns with previous finding that retinoic acid, a metabolic product of retinol, can also regulate mTOR signaling (45). These results demonstrated the robustness of our integrative methodology in uncovering novel gene inhibitors. As the database of genetic and chemical perturbation-induced transcriptional phenotypes within PerturbDB, we anticipate the identification of an increasing number of inhibitors.

Analyzing gene functions utilizing PGSEA

To facilitate users to analyze gene functions, we have developed the PGSEA tool. This tool allows the identification of genes in PerturbDB that share a similar perturbed transcriptomic phenotype with a gene of interest (Figure 4A). PGSEA employs two distinct methods: the hypergeometric test-based over-representation analysis (ORA) method (46) and the GSEA method (22). Utilizing the METTL3 knockdown bulk perturbation RNA-Seq dataset (17), we demonstrated the efficacy of tool. In the GSEA mode, PGSEA identified nine genes sharing similar functions with METTL3, including five known components of the m^{6}A writer complex: WTAP, ZC3H13, METTL3, CBLL1 and METTL14 (Figure 4B and C). In contrast, the ORA mode identified 1936 genes, 6 of which belong to the m^{6}A writer complex: WTAP, METTL3, CBLL1, RBM15, VIRMA and METTL14 (Figure 4B and C). Gene ontology (GO) enrichment analysis showed that both modes significantly enriched for the 'RNA N^6-methyladenosine methyltransferase complex' term (Figure 4D), suggesting that both modes are effective in predicting gene functions, with the GSEA mode showing higher specificity than the ORA mode. As the Perturb-Seq/bulk perturbation RNA-Seq datasets within PerturbDB continue to expand, the capability of PGSEA to predict gene functions will correspondingly increase.

Exploring gene functions and regulatory networks with PerturbDB

To enable users to effectively utilize the vast amount of Perturb-Seq dataset and functional analysis results generated in this study, we have developed an intuitive web interface called PerturbDB (http://research.gzsys.org.cn/perturbdb), featuring browsing, searching, visualization and exploration functionalities. With PerturbDB, users can readily access information on downstream gene alterations and pathway enrichment alterations when a gene of interest is subjected to genetic perturbation in various cell lines. Additionally, the platform enables exploration of cancer-associated genes within the 'Cancer' module, potential gene inhibitors in the 'Inhibitor' module and gene functions using the 'Perturbation cluster' module or the PGSEA tool. This valuable resource greatly aids users in deciphering gene functions and their downstream regulatory networks. Finally, PerturbDB incorporated perturbation scATAC-Seq datasets to complement the multi-omics dataset. This addition facilitates the analysis of gene activity and RNA-Seq expression changes, aiding users in identifying potential upstream regulatory relationships.

Next, we illustrate the utility of PerturbDB with a specific example: within the 'Cancer' module, SIRT7 is significantly upregulated in several cancers, including liver hepatocellular carcinoma (LIHC), kidney renal clear cell carcinoma, uterine corpus endometrial carcinoma, kidney renal papillary cell carcinoma and lung adenocarcinoma (Figure 5A and B, and Supplementary Figure S2A–D). Moreover, this upregulation of SIRT7 is remarkably correlated with poorer survival outcomes in these cancers (Supplementary Figure S2E–H). Utilizing the search functionality of PerturbDB, we found that the inhibition of SIRT7 commonly suppresses mTOR signaling in K562, RPE1 and Jurkat cell lines (Figure 5C–G and Supplementary Figure S3A–J). In the 'Perturbation cluster' module, SIRT7 was identified as a member of the sPC00010 in three cell lines (Figure 5H). The cluster is annotated as

the 'mTOR signaling pathway', positioning SIRT7 as a regulator of this pathway (Figure 5I and Supplementary Figure S3I and J). This correlation is supported by the findings of Tsai *et al.* (47), further demonstrating the role of SIRT7 in the mTOR pathway. These results suggest that SIRT7 may promote oncogenic potential across various cancers by activating mTOR signaling. The potential regulatory role of SIRT7 in LIHC regulation is worth further experimental investigation. Most of the above results were directly extracted from the PerturbDB web interface, demonstrating its efficiency in elucidating gene functions and regulatory networks.

Discussion and conclusions

In this study, we have developed an interactive gene functional analysis platform called PerturbDB (http://research.gzsys.org.cn/perturbdb). This platform enables researchers to readily access downstream regulatory networks, potential functions and inhibitors of genes they are interested in. To facilitate functional analysis of genes that are not yet included in PerturbDB, we have also developed a tool called PGSEA. Using this tool, users can infer gene functions by leveraging the large volume of Perturb-Seq datasets included in PerturbDB. Furthermore, all data in PerturbDB can be freely and conveniently downloaded in bulk mode.

To facilitate users with easy access to all resources in PerturbDB, we developed a user-friendly and intuitive web interface, which includes browsing, searching, visualization and exploration functionalities. Within the web interface, users can easily browse data through five distinct modules: Gene, Dataset, Perturbation cluster, Inhibitor and Cancer. Moreover, users can simply enter a gene symbol or ID in the input box located on the homepage or the top left corner of any page. PerturbDB will then return all relevant information derived from these five modules for the specified gene on a single page. The user-friendly nature of PerturbDB significantly enhances the effectiveness of users in exploring gene functions and regulatory networks.

The identification of 143 unannotated perturbation clusters stands as a significant contribution to understanding complex cellular functions and cancer regulation. These perturbation clusters, potentially representing novel protein complexes and signaling pathways, present exciting opportunities for new discoveries despite their current lack of existing annotations. Detailed literature reviews could shed light on the roles of genes within these perturbation clusters. For instance, PC00007 includes genes such as KPNB1, ANKRD17, DLD and CSE1L. KPNB1 and CSE1L are known for their roles in nuclear transportation of proteins (48), whereas DLD is known as a dehydrogenase critical for energy metabolism (49). ANKRD17, a widely expressed RBP, is implicated in the Hippo signaling pathway regulation (50,51). The convergence of these four genes within a functionally similar cluster suggests a potential coordination between protein nuclear transportation, energy metabolism and Hippo pathway. Therefore, a thorough exploration of the unannotated perturbation clusters identified by us could provide valuable insights into the mechanisms governing various cellular physiological and pathological processes.

In response to the rapid accumulation of Perturb-Seq datasets, we are committed to continuously updating PerturbDB, ensuring that it remains a comprehensive and up-to-date repository. Anticipated future developments include incorporation of a deep learning tool, designed to streamline the discovery of gene functions. Given the existing uncertainties in the roles of numerous genes in cancer progression, PerturbDB will emerge as an indispensable tool for researchers, aiding significantly in the elucidation of complex gene regulatory networks.

Data availability

The PerturbDB database is available at http://research.gzsys.org.cn/perturbdb. All the data in PerturbDB are available for free download. All the analysis code used in this study is available at https://github.com/sysyangb/PerturbDB and https://zenodo.org/doi/10.5281/zenodo.13327403.

Supplementary data

Supplementary Data are available at NAR Online.

Funding

National Key Research and Development Program of China [2021YFA0909300]; National Natural Science Foundation of China [82373023, 82073067, 82273033 and 82072924]; the Science and Technology Planning Project of Guangdong Province [2019B020226003, 2023B1212060013, 2022B1515020100 and 2020B1212030004]; Guangzhou Bureau of Science and Information Technology [202201020575 and 2024A04J6554]. Funding for open access charge: Sun Yat-Sen Memorial Hospital.

Conflict of interest statement

None declared.

References

1. Replogle,J.M., Saunders,R.A., Pogson,A.N., Hussmann,J.A., Lenail,A., Guna,A., Mascibroda,L., Wagner,E.J., Adelman,K., Lithwick-Yanai,G., *et al.* (2022) Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell*, **185**, 2559–2575.
2. Dixit,A., Parnas,O., Li,B., Chen,J., Fulco,C.P., Jerby-Arnon,L., Marjanovic,N.D., Dionne,D., Burks,T., Raychowdhury,R., *et al.* (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, **167**, 1853–1866.
3. Schraivogel,D., Gschwind,A.R., Milbank,J.H., Leonce,D.R., Jakob,P., Mathur,L., Korbel,J.O., Merten,C.A., Velten,L. and Steinmetz,L.M. (2020) Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods*, **17**, 629–635.
4. Ursu,O., Neal,J.T., Shea,E., Thakore,P.I., Jerby-Arnon,L., Nguyen,L., Dionne,D., Diaz,C., Bauman,J., Mosaad,M.M., *et al.* (2022) Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat. Biotechnol.*, **40**, 896–905.
5. Shifrut,E., Carnevale,J., Tobin,V., Roth,T.L., Woo,J.M., Bui,C.T., Li,P.J., Diolaiti,M.E., Ashworth,A. and Marson,A. (2018) Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, **175**, 1958–1971.
6. 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.

7. Guo,S., Xu,Z., Dong,X., Hu,D., Jiang,Y., Wang,Q., Zhang,J., Zhou,Q., Liu,S. and Song,W. (2023) GPSAdb: a comprehensive web resource for interactive exploration of genetic perturbation RNA-seq datasets. *Nucleic Acids Res.*, **51**, D964–D968.
8. 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.
9. Peidli,S., Green,T.D., Shen,C., Gross,T., Min,J., Garda,S., Yuan,B., Schumacher,L.J., Taylor-King,J.P., Marks,D.S., *et al.* (2024) scPerturb: harmonized single-cell perturbation data. *Nat. Methods*, **21**, 531–540.
10. Omer,F. and Kuzu,F.S. (2023) GeneSetR: a web server for gene set analysis based on genome-wide Perturb-Seq data. bioRxiv doi: https://doi.org/10.1101/2023.09.18.558211, 18 September 2023, preprint: not peer reviewed.
11. Zhiting Wei,D.S., Duan,B., Gao,Y., Yu,Q., Guo,L. and Liu,Q.. (2024) PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization. bioRxiv doi: https://doi.org/10.1101/2024.02.03.578767, 05 February 2024, preprint: not peer reviewed.
12. Papalexi,E., Mimitou,E.P., Butler,A.W., Foster,S., Bracken,B., Mauck,W.M., Wessels,H.H., Hao,Y., Yeung,B.Z., Smibert,P., *et al.* (2021) Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat. Genet.*, **53**, 322–331.
13. The Cancer Genome Atlas Research Network (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**, 378–384.
14. Liu,J., Lichtenberg,T., Hoadley,K.A., Poisson,L.M., Lazar,A.J., Cherniack,A.D., Kovatich,A.J., Benz,C.C., Levine,D.A., Lee,A.V., *et al.* (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, **173**, 400–416.
15. The Cancer Genome Atlas Research Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
16. Duan,Q., Flynn,C., Niepel,M., Hafner,M., Muhlich,J.L., Fernandez,N.F., Rouillard,A.D., Tan,C.M., Chen,E.Y., Golub,T.R., *et al.* (2014) LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.*, **42**, W449–W460.
17. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
18. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W., *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.
19. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
20. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I., *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
21. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M. 3rd, 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**, 3573–3587.
22. 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.
23. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
24. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K., R,A.M., Digles,D., Lopes,E.N., Ehrhart,F., *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
25. Yu,G., Wang,L.G., Han,Y. and He,Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
26. Lo,W.C., Lee,C.C., Lee,C.Y. and Lyu,P.C. (2009) CPDB: a database of circular permutation in proteins. *Nucleic Acids Res.*, **37**, D328–D332.
27. Tsitsiridis,G., Steinkamp,R., Giurgiu,M., Brauner,B., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2023) CORUM: the comprehensive resource of mammalian protein complexes—2022. *Nucleic Acids Res.*, **51**, D539–D545.
28. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
29. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B., *et al.* (2018) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
30. 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.
31. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A., Ross,K.N., *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
32. Frangieh,C.J., Melms,J.C., Thakore,P.I., Geiger-Schuller,K.R., Ho,P., Luoma,A.M., Cleary,B., Jerby-Arnon,L., Malu,S., Cuoco,M.S., *et al.* (2021) Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat. Genet.*, **53**, 332–341.
33. Norman,T.M., Horlbeck,M.A., Replogle,J.M., Ge,A.Y., Xu,A., Jost,M., Gilbert,L.A. and Weissman,J.S. (2019) Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, **365**, 786–793.
34. Jost,M., Santos,D.A., Saunders,R.A., Horlbeck,M.A., Hawkins,J.S., Scaria,S.M., Norman,T.M., Hussmann,J.A., Liem,C.R., Gross,C.A., *et al.* (2020) Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.*, **38**, 355–364.
35. Replogle,J.M., Bonnar,J.L., Pogson,A.N., Liem,C.R., Maier,N.K., Ding,Y., Russell,B.J., Wang,X., Leng,K., Guna,A., *et al.* (2022) Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors. *eLife*, **11**, e81856.
36. Burkhardt,D.B., Stanley,J.S., Tong,A., Perdigoto,A.L., Gigante,S.A., Herold,K.C., Wolf,G., Giraldez,A.J., van Dijk,D. and Krishnaswamy,S. (2021) Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol.*, **39**, 619–629.
37. Duan,B., Zhou,C., Zhu,C., Yu,Y., Li,G., Zhang,S., Zhang,C., Ye,X., Ma,H., Qu,S., *et al.* (2019) Model-based understanding of single-cell CRISPR screening. *Nat. Commun.*, **10**, 2233.
38. Funk,L., Su,K.C., Ly,J., Feldman,D., Singh,A., Moodie,B., Blainey,P.C. and Cheeseman,I.M. (2022) The phenotypic landscape of essential human genes. *Cell*, **185**, 4634–4653.
39. Shao,J., Tanner,S.W., Thompson,N. and Cheatham,T.E. (2007) Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.*, **3**, 2312–2334.
40. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R., *et al.* (2020) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
41. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

42. Horiuchi,M., Hayashida,W., Akishita,M., Yamada,S., Lehtonen,J.Y., Tamura,K., Daviet,L., Chen,Y.E., Hamai,M., Cui,T.X., *et al.* (2000) Interferon-gamma induces AT_2 receptor expression in fibroblasts by Jak/STAT pathway and interferon regulatory factor-1. *Circ. Res.*, **86**, 233–240.
43. Krutyhołowa,R., Hammermeister,A., Zabel,R., Abdel-Fattah,W., Reinhardt-Tews,A., Helm,M., Stark,M.J.R., Breunig,K.D., Schaffrath,R. and Glatt,S. (2019) Kti12, a PSTK-like tRNA dependent ATPase essential for tRNA modification by Elongator. *Nucleic Acids Res.*, **47**, 4814–4830.
44. Ovchinnikov,A., Pérez Verona,I., Pogudin,G. and Tribastone,M. (2021) CLUE: exact maximal reduction of kinetic models by constrained lumping of differential equations. *Bioinformatics*, **37**, 3385.
45. Long,C., Zhou,Y., Shen,L., Yu,Y., Hu,D., Liu,X., Lin,T., He,D., Xu,T., Zhang,D., *et al.* (2022) Retinoic acid can improve autophagy through depression of the PI3K–Akt–mTOR signaling pathway via RARα to restore spermatogenesis in cryptorchid infertile rats. *Genes Dis.*, **9**, 1368–1377.
46. Pomyen,Y., Segura,M., Ebbels,T.M. and Keun,H.C. (2015) Over-representation of correlation analysis (ORCA): a method for identifying associations between variable sets. *Bioinformatics*, **31**, 102–108.
47. Tsai,Y.C., Greco,T.M. and Cristea,I.M. (2014) Sirtuin 7 plays a role in ribosome biogenesis and protein synthesis. *Mol. Cell. Proteomics*, **13**, 73–83.
48. Dong,Q., Li,X., Wang,C.Z., Xu,S., Yuan,G., Shao,W., Liu,B., Zheng,Y., Wang,H., Lei,X., *et al.* (2018) Roles of the CSE1L-mediated nuclear import pathway in epigenetic silencing. *Proc. Natl Acad. Sci. U.S.A.*, **115**, E4013–E4022.
49. Malty,R.H., Aoki,H., Kumar,A., Phanse,S., Amin,S., Zhang,Q., Minic,Z., Goebels,F., Musso,G., Wu,Z., *et al.* (2017) A map of human mitochondrial protein interactions linked to neurodegeneration reveals new mechanisms of redox homeostasis and NF-κB signaling. *Cell Syst.*, **5**, 564–577.
50. Sansores-Garcia,L., Atkins,M., Moya,I.M., Shahmoradgoli,M., Tao,C., Mills,G.B. and Halder,G. (2013) Mask is required for the activity of the Hippo pathway effector Yki/YAP. *Curr. Biol.*, **23**, 229–235.
51. Sidor,C.M., Brain,R. and Thompson,B.J. (2013) Mask proteins are cofactors of Yorkie/YAP in the Hippo pathway. *Curr. Biol.*, **23**, 223–228.

Received: June 8, 2024. **Revised:** July 26, 2024. **Editorial Decision:** August 13, 2024. **Accepted:** September 5, 2024